APA/C-DAC International Conference on
Digital Preservation and Development of Trusted Digital Repositories

5–6 February 2014
New Delhi, India

Sponsored as part of the Centre of Excellence for Digital Preservation by Department of Electronics and Information Technology (DeitY), Government of India

Edited by: Dinesh Katre and David Giaretta

Jointly Organized by

Excel India Publishers
New Delhi
Programme Schedule

APA/C-DAC International Conference on
Digital Preservation and Development
of Trusted Digital Repositories

5-6 February 2014

INTRODUCTION

As noted in the IDC country brief, the digital universe in India is expected to grow 23 fold between 2012–2020. The digital bits captured or created each year in the country is expected to grow from 127 exabytes to 2.9 zettabytes between 2012 and 2020. Asia’s giants India and China together would account for the 29% of the total digital universe in 2020 which is estimated to be 40 zettabytes. The IDC digital universe report also includes similar country briefs and growth projections about China, US and Western Europe. The growth projections pertaining to the production and capture of digital information or big data is propelled by internet, social networks, e-government and enterprise applications, migration from analogue to digital, cloud based applications, machine to machine and mobile communication, etc. The explosion of digital information has caught most of us unprepared to handle the threats posed by rapidly changing technologies and digital obsolescence of computer hardware, software, file formats and storage media. The possibilities of data corruption, physical damage and disasters continue to endanger digitally encoded information. The legal, financial, and administrative consequences of loss of digital information pose major concerns.

In this context, the need and relevance of digital preservation is already recognized by UNESCO which has decided to play active advocacy role to make digital preservation frameworks and practices as per the Vancouver Declaration, The Memory of the World in the Digital Age: Digitization and Preservation, 2012. It is challenging to ensure the accessibility, readability, usability, reliability and authenticity of such digital information over a long period. The ephemeral and intangible nature of digital information demands continuous and active management for its long term preservation. It requires the development of digital preservation best practices, methods, tools, systems, and infrastructural setup including trusted and sustainable digital repositories.

To frame a response to the challenge, the purpose of this international conference is to facilitate deliberations and sharing of domain specific case studies between digital preservation experts, technology developers, researchers, archivists, legal and law enforcement professionals, policy makers, and stakeholders.
### Conference Schedule

**Wednesday, February 5, 2014**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session/Activity Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 AM-09:00 AM</td>
<td>Registration</td>
</tr>
<tr>
<td>09:00 AM-10:30 AM</td>
<td><strong>Inaugural Session</strong>&lt;br&gt;  - Welcome Address by Dr. Hemant Darbari, Executive Director, C-DAC, Pune&lt;br&gt;  - Keynote by Dr. Juan Bicarregui, Chair-Alliance for Permanent Access (APA)&lt;br&gt;  - Address by Dr. G.V. Ramaraju, Group Coordinator &amp; HOD, R&amp;D in IT Division, Department of Electronics &amp; Information Technology (DeitY), Government of India&lt;br&gt;  - Address by Prof. Rajat Moona, Director General, C-DAC&lt;br&gt;  - Address by J. Satyanarayana, Secretary, Department of Electronics &amp; Information Technology (DeitY), Government of India&lt;br&gt;  - Opening Remarks by Dr. David Giaretta, Director Alliance for Permanent Access (APA)&lt;br&gt;  - Vote of thanks by Dr. Dinesh Katre, Associate Director &amp; HoD, C-DAC, Pune</td>
</tr>
<tr>
<td>10:30 AM-11:00 AM</td>
<td>Tea Break</td>
</tr>
</tbody>
</table>
| 11:00 AM-13:00 PM | **Session 1: Conference Theme**<br>  - Keynote: Who Audits Trusted Digital Repositories?, David Giaretta, Director, Alliance for Permanent Access (APA)<br>  - Digital Preservation and Development of Trusted Digital Repositories: An Indian Perspective, Dinesh Katre  
  - Session 2: Preserving and Managing Trust in Electronic Records  
| 13:00 PM-14:00 PM | Lunch Break                                                                                                                                                                                                           |
| 16:00 PM-16:15 PM | Tea Break                                                                                                                                                                                                           |
| 16:15 PM-16:45 PM | **Keynote:** Cloud Infrastructure for Trusted Digital Repositories, Mukul Sinha, Managing Director, ESCL, India  
  - Panel Discussion: Envisaging the National Digital Preservation Infrastructure for India  
    - Dr. Mukul Sinha, Managing Director, ESCL, India  
    - Dr. Gautam Bose, Consultant, NIC  
    - J.D. Shiers, Data Management Group, CERN  
    - Natasa Milic-Frayling-Principal Researcher-Microsoft Research  
    - Shirley Crompton, e-Science Department, STFC  
| 16:45 PM-17:45 PM | Panel Discussion: Envisaging the National Digital Preservation Infrastructure for India  
  - Dr. Mukul Sinha, Managing Director, ESCL, India  
  - Dr. Gautam Bose, Consultant, NIC  
  - J.D. Shiers, Data Management Group, CERN  
  - Natasa Milic-Frayling-Principal Researcher-Microsoft Research  
  - Shirley Crompton, e-Science Department, STFC  
<p>| 19.30 PM to 21.00 PM | Conference Dinner                                                                                                                                                                                                     |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 09:30 AM-11:00 PM| Session 4: Digital Preservation of Scientific Data  
  **Keynote**: Exa-scale Data Preservation in High Energy Physics, J.D. Shiers, Data Management Group, CERN  
  Digital Preservation of Remote Sensing Data, V.V.S. Nageswara Rao  
  The purpose behind repurposing [historic] data, Elizabeth Griffin  
  Towards a Virtual Centre of Excellence supporting Digital Preservation of the Records of Science, Matthias Hemmje |
| 11:00 AM-11:15 PM| Tea Break                                    |
| 11:45 AM-13:00 PM| Session 5: Digital Preservation of Audio, Video and Cultural Heritage  
  Securing Scotland’s 3D Digital Heritage: Finding a practicable middle ground by Emily Nimmo  
  Audiovisual Archives in India in the Digital Era and Initiative for setting up a ‘National Cultural Audiovisual Archives’, Protopanand Jha  
  Sahapedia: An Online Repository and Digital Platform on Indian Arts, Culture and Heritage, Sudha Gopalakrishnan  
  Digital Preservation of Cultural Digital Archives-Work in Progress, Neeraj Mittal, Milind Kapre, Sandeep Gaikwad, Dinesh Katre  
  Collaborative Services through Websites of Institutional Repositories in India, Vaishali B. Wadnerkar, Minakshi R. Sontakke |
| 13:00 PM-14:00 PM| Lunch Break                                  |
| 14:00 AM-15:20 PM| Session 6: Development of Digital Archives  
  Digital Preservation and Permanent Access to Print Media Resources at the TAKC, India: A Case Study by Venkata Kesavan  
  Diplomatics in the Context of National Archives of Zimbabwe: A New Frontier for Extending Records Management Services, D. Maboreke and L. Muchena  
  Development of Digital Preservation and Repository Infrastructure for National Archives of India, T. Hussain, Dinesh Katre, Shashank Puntamkar,  
  Digital Preservation of Information Resources in Academic Libraries in Nigeria, Attahiru Saminu |
| 15:20 PM-16:20 PM| Session 7: Legal Aspects of Digital Preservation  
  Copyright of Digital Content-Myths & Realities, Vakul Sharma  
  Digital Preservation of Court’s Disposed Case Records: A Case Study from Indian Judicial System’s perspective by Papul Abichandani, Rishi Prakash  
  Digital Preservation for legal compliances. Where does Swaziland stands by Thiyam Satyabati Devi, Ntombikayise Nomsa Mathabela  
  Role of Digital Forensics in Digital Preservation as per the Indian Legal Requirements by Yogendra Tank, Nikhil Padhiyar, Bhavesh Gabani, Dinesh Katre |
| 16:20 PM-16:30 PM| Tea Break                                   |
| 16:30 PM-17:00 PM| Session 8: Preservation and Value: Support based on a coherent view of preservation, David Giaretta  
  **Panel Discussion**: Long Term Sustenance of Digital Preservation and Trusted Digital Repositories  
  **Panellists**:  
  Dr. David Giaretta, Deputy Director, APA  
  Dr. G.V. Ramaraju, Group Coordinator and Head, R&D in IT, DeitY  
  Dr. Matthias Hemmje, Professor, Fern University in Hagen  
  V. Srinivas, Director General, National Archives of India  
  P.R. Goswami, Director, IGNCA  
  Dr. Luciana Duranti, Director, InterPARES Trust, UBC  
  Dr. Dinesh Katre, Associate Director & HoD, C-DAC |
Message
Foreword

Welcome to the 2014 APA/C-DAC International Conference on Digital Preservation & Development of Trusted Digital Repositories!

There have been many conferences all over the world on the topic of digital preservation but this one is perhaps the first one which focuses on development of Trusted Digital Repositories, which is of immense interest in India as well as other countries. We are delighted that you have decided to contribute to, or participate in, this conference.

To respond to the challenges posed by rapidly changing technologies and digital obsolescence, the purpose of this international conference is to facilitate deliberations and sharing of domain specific case studies between digital preservation experts, archivists, technology developers, researchers, legal and law enforcement professionals, policy makers, and stakeholders. We have therefore put together a conference programme that provides coverage on these wide ranging topics:

- Digital preservation in various domains such as e-government, scientific and research data, geospatial information, electronic health records, audio and video, media archives, government and cultural heritage archives, corporate data archives, etc.
- Domain specific adaptation of Open Archival Information System (OAIS) Reference Model.
- Digital disasters and recovery.
- Technological infrastructure development for trusted digital repositories.
- Trusted digital repositories in cloud.
- Digital preservation for legal compliances.
- Digital forensic for authenticity of information.
- The challenges and techniques for maintaining trust in electronic records over long period.
- Integrating e-record management with e-government.
- Approaches for long term sustainability of digital repositories.
- Data mining and repurposing of digital information stored in digital repositories.
- Preparing for audit and certification of digital repositories.
- Design of domain specific digital preservation policy framework.

These subjects will be covered through discussions in the pre-conference workshops, keynotes, paper presentations and panel sessions.
The organization of this international event has been possible through the support and contribution of many people and organizations. The Alliance for Permanent Access (APA), whose current membership is mostly European, has provided this conference with several international speakers and experts who can share their experience of establishing the digital repositories and domain specific case studies of digital preservation. This conference is organized by Human-Centred Design and Computing Group, C-DAC, Pune as part of the Centre of Excellence for Digital Preservation project sponsored by the Department of Electronics and Information Technology (DeitY), Government of India. We are thankful to all dignitaries for sending us encouraging messages and highlighting the significance of this event. We acknowledge and thank all the authors of the papers, invited speakers and panelists for their valuable contributions.

We hope that this conference will generate the ripples and momentum for establishing the digital preservation infrastructure and trusted digital repositories in India as well as the countries in Asia Pacific region.

As the organizers, we wish you an excellent Conference.

**Conference Co-Chairs**

Dr. Dinesh Katre  
Associate Director & HoD  
Human Centred Design & Computing Group, C-DAC, 4th Floor  
RMZ Westend Centre, III, Aundh, Pune – 411007, India  
dinesh@cdac.in, digitalpreservation@cdac.in

Dr. David Giaretta, MBE  
Director  
Alliance for Permanent Access (APA)  
2 High Street, Yetminster Dorset DT9 6LF, UK  
director@alliancepermanentaccess.org
Acknowledgement

We acknowledge and thank all the paper authors, presenters and speakers for their valuable contributions to this conference.

We acknowledge and thank the R&D in IT Division, Department of Electronics and Information Technology (DeitY), Government of India, as this conference became possible through the funding and mandate granted for the Centre of Excellence for Digital Preservation project. We are extremely thankful to Dr. G.V. Ramaraju, Group Coordinator and HoD of R&D in IT Division, DeitY for encouraging our activities in digital preservation. We also thank S.A. Kumar, Director, R&D in IT Division, DeitY for his guidance and support.

We are thankful to Alliance for Permanent Access (APA) for the encouragement and active participation in making this conference possible. Our special thanks to Krystina Giaretta, Office Manager, Alliance for Permanent Access, UK for proactively supporting us throughout the organization of this conference.

We are thankful to Prof. Rajat Moona, Director General, C-DAC and Dr. Hemant Darbari, Executive Director, C-DAC Pune for encouragement throughout the organization of this event. Our sincere thanks to the members of Human-Centred Design and Computing Group at C-DAC Pune and various departments within C-DAC that have directly or indirectly helped us in organizing this conference.
Committee

Conference Co-chairs

Dr. Dinesh Katre
Associate Director & HoD, C-DAC

Dr. David Giaretta, MBE
Director, Alliance for Permanent Access

Advisory Committee

Dr. Luciana Duranti
Director, Inter PARES
University of British Columbia, Canada

Prof. John Wood
Secretary General
Association of Commonwealth Universities, UK

Dr. Juan Bicarregui
Head of Data Services Division
Scientific Computing Department
STFC-Rutherford Appleton Laboratory, UK

Monica Marinucci
Director for Research
Oracle (Europe, Middle East and Africa)

S.A. Kumar
Director, R&D in IT Division
Department of Electronics & Information Technology (DeitY),
Government of India

Renu Budhiraja
Scientist G, NeGD, DeitY
New Delhi, India

Dr. Gautam Bose
Deputy Director General
NIC, New Delhi, India

Dr. Mukul Sinha
Managing Director
ESCL, Noida, India

Bill LeFurgy
Digital Initiatives Manager
National Digital Information Infrastructure and
Preservation Program at the Library of Congress, USA

Dr. Reagan W. Moore
School of Information and Library Science
University of North Carolina at Chapel Hill, USA

Prof. Matthias Hemmje
Fern University Hagen
Germany

Dr. Jussi Nuorteva
Director General
The National Archives of Finland
# Programme Committee

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution and Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlo Meghini</td>
<td>Consiglio Nazionale Delle Ricerche, Institute of Information Science and Technologies, Italy</td>
</tr>
<tr>
<td>Dr. Anne Thurston</td>
<td>OBE Director International Records Management Trust, UK</td>
</tr>
<tr>
<td>Dr. Peter Doorn</td>
<td>Data Archiving and Networking Services (DANS) Netherlands</td>
</tr>
<tr>
<td>Abhishek Jee</td>
<td>Director e-Governance and Green IT, NASSCOM, Bangalore</td>
</tr>
<tr>
<td>Vakul Sharma</td>
<td>Cyber Law Expert &amp; Advocate New Delhi, Supreme Court, India</td>
</tr>
<tr>
<td>Dr. Ramesh Gaur</td>
<td>Head-Librarian Jawaharlal Nehru University, New Delhi, India</td>
</tr>
<tr>
<td>Prof. Makhlouf Shabou Basma</td>
<td>University of Applied Sciences Western Switzerland</td>
</tr>
<tr>
<td>Dr. Usha Mujoo Munshi</td>
<td>Librarian IIPA, New Delhi, India</td>
</tr>
<tr>
<td>Richard Rinehart</td>
<td>Samek Art Gallery Bucknell University, USA</td>
</tr>
<tr>
<td>Dr. Sanjeev Kumar</td>
<td>Deputy Director NRC, NIFT, New Delhi, India</td>
</tr>
<tr>
<td>Pratapanand Jha</td>
<td>Director IGNCA, New Delhi, India</td>
</tr>
<tr>
<td>R. Venkata Kesavan</td>
<td>DGM &amp; Head TAKC The Times Group, India</td>
</tr>
<tr>
<td>Dr. Jeremy Leighton John</td>
<td>Curator of eMSS The British Library, London, UK</td>
</tr>
<tr>
<td>Barbara Sierman</td>
<td>Digital Preservation Manager National Library of Netherlands</td>
</tr>
<tr>
<td>Jason R. Baron</td>
<td>College of Information Studies University of Maryland, USA</td>
</tr>
<tr>
<td>Dr. Christopher Lee</td>
<td>School of Information and Library Science University of North Carolina, USA</td>
</tr>
<tr>
<td>T. Hussain</td>
<td>Assistant Director National Archives of India, New Delhi, India</td>
</tr>
</tbody>
</table>
Local Organizing Committee

Dr. Dinesh Katre  
Associate Director & HoD  
HCDC Group, C-DAC Pune

Shashank Puntamkar  
Joint Director  
HCDC Group, C-DAC Pune

Sudhir Patel  
Principal Technical Officer  
HCDC Group, C-DAC Pune

Rajendra Joglekar  
Principal Technical Officer  
HCDC Group, C-DAC Pune

Srinu Naik  
Principal Technical Officer  
HCDC Group, C-DAC Pune

Niranjan Patgaokar  
Technical Officer  
HCDC Group, C-DAC Pune

Digvijay Gohain  
HoD  
Media and Communications Group, C-DAC Pune

Nitin Karale  
Project Engineer  
HCDC Group, C-DAC Pune

Chandrakant Sontakke  
Secretary, HCDC Group  
C-DAC Pune

Rajeev Rudrakshi  
Principal Technical Officer  
Web Group, C-DAC Pune

Agnelo Rodrigo  
Senior Technical Officer  
Web Group, C-DAC Pune
# Contents

<table>
<thead>
<tr>
<th>Programme Schedule</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messages</td>
<td>ix</td>
</tr>
<tr>
<td>Foreword</td>
<td>xix</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>xxi</td>
</tr>
<tr>
<td>Committees</td>
<td>xxi</td>
</tr>
</tbody>
</table>

## CONFERENCE THEME

1. **Who Audits TDRs**
   
   *David Giaretta, Robert R. Downs, Simon Lambert, John S. Hughes, Mark Conrad, Bruce Anmbacher, Terry Longstreth, John G. Garrett, Helen R. Tibbo, Barbara Sierman and Krystina Giaretta*
   
   Page 3

2. **Digital Preservation and Development of Trusted Digital Repositories: An Indian Perspective**
   
   *Dinesh Katre*
   
   Page 11

## TRUST IN ELECTRONIC RECORDS

3. **Preservation in the Cloud: Towards an International Framework for a Balance of Trust and Trustworthiness**
   
   *Luciana Duranti*
   
   Page 23

4. **Authenticity in Records Systems: Emerging Research in Digital Preservation**
   
   *Adam Jansen*
   
   Page 39

5. **Electronic Records Capturing using eGOV-PID Standard**
   
   *Sourabh Koriya, Jayshree Pawar, Suman Behara, Srinu Naik and Dinesh Katre*
   
   Page 47

6. **Long-term Preservation, Management and Utility of EHR**
   
   *Gaur Sunder*
   
   Page 58

## DIGITAL PRESERVATION INFRASTRUCTURE

7. **Towards IT Support for Sustained Access and Use of Digital Content**
   
   *Natasa Milic-Frayling*
   
   Page 71

8. **Building a Federated Infrastructure for Preservation of and Access to Research Data in the Netherlands: The Front Office-Back Office Model**
   
   *Peter Doorn, Ingrid Dillo and Paula Witkamp*
   
   Page 72

9. **SCIDIP-ES: A Sustainable Data Preservation Infrastructure to Support OAIS Conformant Archives**
   
   *Shirley Crompton, David Giaretta, Brian Matthews, Holger Brocks, Felix Engel, Arif Sboa and Fulvio Marelli*
   
   Page 78
   Roman Graf, Ross King and Sven Schlarb  87

11. Scaling Up and Scaling Out: Leveraging Preservation Infrastructure and Experience to Benefit the Community  
   Kate Wittenberg, Amy Kirchhoff and Sheila Morrissey  98

12. Digital Repository on Cloud Infrastructure: Issues & Challenges  
   Mukul K. Sinha  107

### DIGITAL PRESERVATION OF SCIENTIFIC DATA

13. Exa-scale Data Preservation in High Energy Physics  
   J.D. Shiers  123

   V.V.S. Nageswara Rao, K.V. Rathna Kumar, Vinod Bathale, M.V. Ravi Kumar and S. Muralikrishnan  128

15. The Purpose Behind Repurposing [HISTORIC] Data: A Case Study in Ozone  
   Elizabeth Griffin  133

16. Towards a Virtual Centre of Excellence Supporting Digital Preservation of the Records of Science  
   Matthias Hemmje, Ruben Riestra, Simon Lambert and David Giaretta  142

### DIGITAL PRESERVATION OF CULTURAL HERITAGE

17. Securing Scotland’s 3D Digital Heritage: Finding a Practicable Middle Ground  
   Emily Nimmo  155

18. Audiovisual Archives in India in the Digital Era and Initiative for Setting up a ‘National Cultural Audiovisual Archives’  
   Pratapanand Jha  163

19. Digital Preservation of Cultural Digital Archives—Work in Progress  
   Neeraj Mittal, Milind Kapre, Sandeep Gaikwad, Dinesh Katre and O.N. Chaubey  169

20. Collaborative Services through Websites of Institutional Repositories in India: An Evaluation  
   Vaishali B. Wadnerkar and Minakshi R. Sontakke  176
## DEVELOPMENT OF DIGITAL ARCHIVES

21. Digital Preservation and Permanent Access to Print Media Resources at the TAKC, India: A Case Study  
   R. Venkata Kesavan  
   187

22. Diplomats in the Context of National Archives of Zimbabwe: A New Frontier for Extending Records Management Services  
   D. Maboreke and L. Muchefa  
   198

23. Ongoing Development of Digital Preservation and Repository Infrastructure for National Archives of India  
   Dinesh Katre, Shashank Puntamkar, T. Hussain and Ankit Sharma  
   202

   Attahiru Saminu  
   208

## LEGAL ASPECTS OF DIGITAL PRESERVATION

25. Copyright of Digital Content—Myths & Realities  
   Vakul Sharma  
   217

26. Digital Preservation of Court’s Disposed Case Records—A Case Study from Indian Judicial System’s Perspective  
   Payal Abichandani and Rishi Prakash  
   220

27. Digital Preservation for Legal Compliances: Where does Swaziland Stand?  
   Thiyam Satyabati Devi and Ntombikayise Nomsa Mathabela  
   228

28. Role of Digital Forensics in Digital Preservation as per the Indian Legal Requirements  
   Yogendra Tank, Bhavesh Gabani, Nikhil Padhiyar and Dinesh Katre  
   237

## VALUE FROM DIGITAL PRESERVATION

29. Preservation and Value: Support based on a Coherent View of Preservation  
   David Giaretta  
   251

AUTHOR INDEX  

261
Conference Theme
Abstract—There has been, for decades, a demand for a way to evaluate how well repositories are preserving their holdings. Progress has been made to put a full ISO audit and certification process into effect, one important component of which is the availability of auditors. This paper describes the work which has been carried out to specify the requirements for evaluating repositories as Trusted Digital Repositories (TDRs) and the competencies which the auditors will need to determine whether repositories can be certified as TDRs.

Keywords: Digital Preservation, Trustworthy Digital Repository, Audit, Certification

BACKGROUND


- A critical component of digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections.
A process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information.

The issue of certification, and how to evaluate trust into the future, as opposed to a relatively temporary trust which may be more simply tested, has been a recurring request, repeated in many subsequent studies and workshops. Organisations with preservation policies frequently mention the intention to become a certified repository. The Open Archival Information System (OAIS) Reference Model([2]), now adopted as the “de facto” standard for building digital archives([3]) and published as ISO 14721, contained a roadmap for follow-on standards; one of these was for certification of archives.

STANDARDS
The OAIS Reference Model introduced a number of fundamental concepts in digital preservation, including the Functional Model and, most importantly the Information Model. The conformance section of the standard specifies that:

A conforming OAIS Archive implementation shall support the model of information …. The OAIS Reference Model does not define or require any particular method of implementation of these concepts.

A conforming OAIS Archive shall fulfil the responsibilities listed in [following in the standard].

Despite this definition, it has been observed([4]) that there have been many claims of “OAIS compliance” and the route commonly, and very mistakenly, taken is by mapping their repository to the OAIS functional model. The mistake is of course that such claims ignore the mandatory responsibilities and the information model of the standard. However, as also noted, OAIS does not lend itself to self-auditing or certification—as it was intended to be a reference model, not a blueprint. It was therefore necessary to create the standard foreseen in OAIS, to address not just OAIS conformance but more generally the trustworthiness of a digital repository.

RLG and NARA led the first phase which produced the TRAC document([5]). The first draft document was published for review and comment and subjected to test audits. After revision it was published by CRL as a CRL-RLG-NARA document.

TRAC was, as had been planned, taken into the CCSDS organisation, in the Repository Audit and Certification Working Group (RAC-WG), as a foundation to be developed into the draft standard, Audit and Certification of Trustworthy Digital Repositories([6]). This was then passed on to ISO and has been published as ISO 16363([7]). Although there is significant overlap with TRAC, TRAC was somewhat skewed towards self-audits of digital libraries. In contrast ISO 16363 was designed from the start to form the basis for a full external audit process of all types of repositories: government or private, or managed by consortia or other institutions, concerned with preservation of information from cultural to science to commercial, and with international, trained, consistent, cohorts of auditors to supply whatever the scale of demand.
An important part of this strategy was the creation of a further standard, Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories (ISO 16919) [8, 9], which defines the way in which the external audit and certification must be undertaken and specifies the qualities that the auditing body must possess. There is a hierarchy of standards concerned with good auditing practice [10], [11] and [12]. ISO 16919 is positioned within this hierarchy in order to ensure that these good practices can be applied to the evaluation of the trustworthiness of digital repositories.

Another part of this strategy was to ensure that the external audit process would be reasonably achievable in practical terms. For this there is no substitute for experience. To do this a number of test audits [13] were performed in four countries and on five different types of archives before the two standards were finalised in order to ensure (1) that the metrics were understandable and applicable to a variety of repositories and (2) that the evaluation of the evidence could be done in a consistent manner.

It is worth noting that these standards were produced while the OAIS was being revised by the CCSDS Data Archive and Ingest Working Group (DAI-WG). There was a significant overlap between the membership of RAC-WG and DAI-WG which ensured the consistency of the standards, as well as several associated CCSDS standards such as the PAIS set dealing with interoperations between data producers and TDRs.

ISO 16363: AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

In setting up a standard for audit and certification, it is clear that one cannot cover all possible situations that might arise, nor can one prescribe exactly what each repository must do. This is the case with all types of audits. Instead one must leave a lot to the judgment of the auditors, but within useful guidelines.

To understand the way in which the metrics in ISO 16363 (referred to below as the “metrics document”) were written, it is helpful to think about the document in the following way, building it up in the same way as the authors of that document.

A very important thing to understand is that in judging a repository one could look at many types of issues related to that repository. For example, is the lighting adequate, are the facilities wheelchair accessible, does the repository respond to requests within 3 minutes, is it easy to find what one is looking for, and so on. However these are not the things against which the repository is to be judged here. Instead ISO 16363 focuses on how well a repository preserves the digitally encoded information with which it has been entrusted.

With this in mind, one could say that since the audit and certification depends on the judgement of the auditors, then there is a single ‘metric’ to be satisfied, namely “Make sure the repository does a good job in preserving its holdings”.

Of course this would not be adequate. We need to provide more guidance for the auditors. Therefore we start by saying “Well at least look at the organisation–make sure it cannot suddenly go out of business, and also make sure that they know how to preserve the digital objects.” With these considerations, one can say that there are two guidelines for auditors:

- Look at the organisation and its overall viability to ensure that it has sufficient commitment, and would have time, to hand over its holdings in an adequate way in the event that it could no longer preserve and provide access to them.
- Look at the way it takes care of the digital material.

In fact there is a third area, which one could argue is part of the second one, namely:

- Make sure that the digital holdings are secure and cannot be unknowingly altered.

The reason for this third category is to avoid the necessity for the repository to also undergo a security audit separately (ISO 27000). Therefore, it seemed sensible to provide a separate category which could possibly be replaced by ISO 27000 certification—but such additional certification is definitely not required.

Therefore we have three main headings:

- Organisational Infrastructure.
- Digital Object Management.
- Infrastructure and Security Risk Management.

Continuing this process we can specify the topics where the auditor really needs to be sure to look. The following breakdown results:

- Organisational Infrastructure:
  - Governance & organizational viability.
  - Organizational structure & staffing.
  - Procedural accountability & preservation policy framework.
  - Financial sustainability.
  - Contracts, licenses, & liabilities.

- Digital Object:
  - INGEST: Acquisition of Content.
  - INGEST: Creation of The AIP.
  - Preservation Planning.
  - AIP Preservation.
Who Audits TDRs

- Information Management.
- Access Management.
- Infrastructure and Security Risk Management:
  - Technical Infrastructure RISK Management.
  - Security Risk Management.

This organisation into topics is not unique and indeed several different sub-categories have been tried, but this one seemed to fit well.

Looking in even more detail, the guidance for the auditors is further split out into metrics. These metrics are well-defined criteria that the auditor should examine carefully and that should be satisfied, if they are applicable to the repository being audited. Some of these metrics are broken into sub-metrics indicating that the auditor needs to check the details more precisely; a few of these sub-metrics have some even more specific sub-sub-metrics specified. Even these sub-sub-metrics are not hugely specific—it is still a matter for the judgement of the auditor to interpret how the repository has addressed the metrics. Indeed the metrics themselves are a matter of judgement; in this case, of course, it is the judgement of the working group which produced the metrics.

Here are two examples from the working groups' test audits:

Metric 3.3.1 The Repository Shall Have Defined Its Designated Community and Associated Knowledge Base(S) and shall have these Definitions Appropriately Accessible.

The Test Audit Team interpretation of this metric during the <SITE> Test Audit was that it required evidence that there was a defined Designated Community for each AIP. The metric was not satisfied. There were no records or documentary evidence presented to describe any Designated Community, nor any AIP associations with any Designated Community.

Metric 3.3.2 The Repository SHALL have Preservation Policies in Place to Ensure its Preservation Strategic Plan will be Met.

The metric was not satisfied. Documentation of Preservation Policy, associated workflows and operating procedures was incomplete and out of date.

These two particular examples are taken because during the test audits these areas were underdeveloped in most of the tested repositories.

COMPETENCIES

Within ISO, CASCO ([14]) is responsible for overseeing conformance testing standards. In the past requirements were specified for training and/or qualifications for auditors. However this is changing. ISO CASCO now favours specifying the competencies that auditors must have, and is in the process of updating many of these standards. One advantage of specifying competencies is that the standard can be much more specific, and should help to ensure that audits conducted around the world are consistent.

This section describes the approach taken to describe the competencies in the version of 16919 submitted to ISO for final review.
Approach

Auditors must have generally applicable skills, but we expect that auditors of a candidate TDR would need additional competencies concerned with digital preservation and knowledge related to the metrics described in ISO 16363. A set of core competencies, based on the experience with the test audits, was created for ISO 16919.

A complete ISO audit process is performed in a number of stages, by a whole team, although only two people would normally perform a site visit. Therefore, each individual auditor would not necessarily have to possess all the specified competencies as long as the complete set of competencies is possessed by the team. Following the normal ISO pattern each competency could be marked as applicable to one or more of:

- Application review.
- Audit team selection.
- Audit planning activities.
- Auditing activities.
- Certification decision.
- Auditor evaluation.

Each of the competencies was stated along the lines of “Possesses the knowledge to…”, followed by a general description of the competency and then a number of bullet points of specific items to be included.

Requirements

The following table summarises the intent of the competencies under each of the main sub-headings.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organisational Infrastructure</strong></td>
<td></td>
</tr>
<tr>
<td>Governance &amp; Organizational Viability</td>
<td>Evaluate an organisation’s commitment to preservation by checking documents such as a mission statement, preservation plans, and collection policies.</td>
</tr>
<tr>
<td>Organisational Structure &amp; Staffing</td>
<td>Evaluate an organisation’s commitment to adequate staffing by checking organizational charts, job descriptions, training requirements, and professional development plans.</td>
</tr>
<tr>
<td>Procedural Accountability &amp; Preservation Policy Framework</td>
<td>Evaluate that an organisation is committed to transparency and accountability, that it has defined its designated community, and that the preservation plans are adequate for the community.</td>
</tr>
<tr>
<td>Financial Sustainability</td>
<td>Evaluate an organisation’s commitment to sustaining the repository over time by checking business plans, budgets, contingency plans, and risk management plans.</td>
</tr>
<tr>
<td>Contracts, Licenses, &amp; Liabilities</td>
<td>Evaluate an organisation’s legal preparedness by evaluating assess contracts, licenses, agreements, and permissions statements.</td>
</tr>
<tr>
<td><strong>Digital Object Management</strong></td>
<td></td>
</tr>
<tr>
<td>Ingest: Acquisition Of Content</td>
<td>Evaluate that the organisation has identified the Content Information and the Information Properties that the repository will preserve for the community.</td>
</tr>
<tr>
<td>Ingest: Creation Of The AIP</td>
<td>Evaluate that the organisation understands and has suitable definitions for the types of AIPs required by the community.</td>
</tr>
</tbody>
</table>

*Table 1 (Contd.)*
...Table 1 (Contd.)

<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preservation Planning</td>
<td>Evaluate an organisation’s understanding of digital preservation strategies, where and how they should be applied, and the changes that may endanger preservation.</td>
</tr>
<tr>
<td>AIP Preservation</td>
<td>Evaluate an organisation’s understanding of the parts of an AIP, their implementation, and risks to preservation due to change.</td>
</tr>
<tr>
<td>Information Management</td>
<td>Evaluate an organisation’s understanding of information technology for discovery, retrieval, and linkage of Content Information.</td>
</tr>
<tr>
<td>Access Management</td>
<td>Evaluate an organisation’s understanding of access policies and their implementation with regard to the designated community.</td>
</tr>
<tr>
<td>Technical Infrastructure Risk Management</td>
<td>Evaluate an organisation’s understanding of computing and storage technologies sufficient to identify impact due to change.</td>
</tr>
<tr>
<td>Security Risk Management</td>
<td>Evaluate an organisation’s commitment to repository security.</td>
</tr>
</tbody>
</table>

CONCLUSION

This paper describes the results of several years’ work to establish the standards on which an ISO audit and certification process for trustworthy digital repositories can be based. In particular, the set of competencies required of the auditors, described here, forms the bedrock of the value of the audits, and completes the series of three linked standards:

- OAIS (ISO 14721).
- Audit and Certification of Trustworthy Digital Repositories (ISO 16363).
- Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories (ISO 16919).

Organisations that intend to perform formal audits and to certify repositories will have to demonstrate to their national accreditation bodies that they have auditors who possess these competencies. Once ISO 16919 has been published, the next steps are to help such organisations achieve accreditation to conduct audits of candidate TDRs so that the full market can develop.

ACKNOWLEDGEMENT

The work reported here is result of the efforts of the whole of the CCSDS Repository Audit and Certification Working Group.

REFERENCES

Digital Preservation and Development of Trusted Digital Repositories: An Indian Perspective

Dinesh Katre
Centre for Development of Advanced Computing, Pune, India
E-mail: dinesh@cdac.in

Abstract—This paper presents an overview of the coordinated efforts between Department of Electronics and Information Technology and C-DAC Pune in envisaging the Indian National Digital Preservation Programme. It also presents the major milestones accomplished by the team of Centre of Excellence for Digital Preservation at C-DAC Pune, contributed towards completing the building blocks of trusted digital repositories as national digital preservation infrastructure.

INTRODUCTION
The R&D in IT Division, Department of Electronics and Information Technology, Government of India recognized the need of digital preservation way back in 2007 and decided to involve the R&D team at C-DAC Pune in envisaging the National Digital Preservation Programme. Since then several millstones have been accomplished which are briefly reported as under.

The Human-Centred Design & Computing Group at C-DAC Pune organized the Indo-US Workshop on International Trends in Digital Preservation in March 2009 in order to seek the recommendations from international experts for the Indian National Digital Preservation Programme. The workshop was organized as guided by the R&D in IT Division, Department of Electronics and Information Technology (DeitY), Government of India. The Indo-US Science and Technology Forum extended the required funding support to make this event possible, which was participated by several experts from India as well as the American National Digital Information Infrastructure and Preservation Programme (NDIIPP) and Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR), UK. This workshop lead the DeitY in entrusting the author of this paper to produce the National Study Report on Digital Preservation Requirements of India in year 2010 as a sponsored time-bound project.
National Study Report on Digital Preservation Requirements of India, 2010

During this study, it was observed that with the ever increasing proliferation of Information Technology, the volume of digital information generated by the government, private and social sectors is exponentially growing. This digital content, though ephemeral and non-tangible in nature, forms a significant part of human heritage for future generations which is constantly threatened due to digital obsolescence and frequent changes in the technologies. In this context, long term digital preservation has emerged as a new interdisciplinary area of research and development to ensure that the digital information preserved in its original, readable and trustworthy form in spite of obsolescence of everything: hardware, software, processes, format, people, etc.

The study revealed that India is extremely vulnerable to loss of digital information and bitter legal consequences in the absence of national digital preservation infrastructure and policy, as Government of India is hugely investing in the computerization and digitalization of its departments at national, state and district levels through the 30 mission mode projects of NeGP. The Information Technology Act, Public Records Act, Right To Information Act and several other laws categorically specify the statutory obligation to retain or preserve the e-records and the digital surrogates produced using computer or other devices by the government organizations. In case of failing to reproduce the digital information in its original, authentic, reliable form and within stipulated time-frame it can lead to penal consequences. Therefore, Government of India must equip itself in the defense of the digital information by creating long-term digital preservation policies, techniques, tools and infrastructures for the benefit of all organizations across diverse domains.

This report includes the overview of international digital presentation projects, study of legal imperatives, various technical standards, and the consolidation of recommendations given by the national expert group which included archivists, technologists and stakeholder representatives of 30 organizations from diverse domains such as e-governance, government records, audio, video and film archives, cultural heritage, health, science and education, insurance and banking, law, etc. It specifies the short term and long term action plans with specific R&D projects to be initiated under the National Digital Preservation Programme. The report provided a recommendation to establish the Centre of Excellence for Digital Preservation to develop the necessary technical competencies, pilot digital repositories and international collaborations.

CENTRE OF EXCELLENCE FOR DIGITAL PRESERVATION, 2011

As recommended in the national study report, as a flagship project under the Indian National Digital Preservation Programme, Department of Electronics and Information Technology (DeitY) approved the Centre of Excellence for Digital Preservation project to be executed by C-DAC Pune with the following objectives.
Objectives

- Conduct research and development in digital preservation to produce the tools, technologies, solutions and infrastructures to ensure that the digital information remains discoverable, accessible, readable, usable, reliable, authentic and trustworthy on long term basis.
- Develop the digital preservation repositories and provide help in nurturing the network of Trustworthy Digital Repositories (national digital preservation infrastructure) as a long-term goal
- Define the digital preservation standards by involving the experts from stakeholder organizations, consolidate and disseminate the best practices generated through various digital preservation projects implemented across India.
- Build the technical competencies necessary for the audit and certification of trustworthy digital repositories, curriculum design and training for data managers, e-records keepers and archivists.
- Spread awareness about the potential threats and risks due to digital obsolescence, the digital preservation best practices and maintain technology watch on continuing basis.

Fig. 1: Overview of Project Activities

Collaborations

In this project, C-DAC Pune has established collaborations with National Archives of India (Pilot Digital Repository of Government Archives), Computer Aided Registration of Documents (CARD) project of Stamps and Registration Department, State Government of
Andhra Pradesh (Pilot Digital Repository of CARD data), e-District Mission Mode Project of NeGP, and Indira Gandhi National Centre for Arts (Pilot Digital Repository of Cultural Digital Data). As part of this project, C-DAC Noida is also developing the pilot digital repository for e-Court Mission Mode Project.

Since 2011, a team 45 engineers is working on the Centre of Excellence for Digital Preservation project and has accomplished some of the objectives.

Digital Preservation Tools and Technologies
The team of Centre of Excellence for Digital Preservation at C-DAC, Pune has undertaken the development of following digital preservation tools and technologies:

**e-Records Digitālaya™**
e-Record Digitālaya™ is a specialized archival system being designed and developed for the preservation and repository management of electronic records produced through various e-governance applications. In the current scope, this system is being developed for the pilot digital repository at Computer Aided Registration of Documents (CARD), C & IGRS, Hyderabad, Andhra Pradesh, India and e-District pilot project.

**Datāntar™ for Capturing of e-governance Records**
Datāntar™ is a software which can connect with the e-Governance database and capture the electronic records along with preservation metadata as per the eGOV-PID standard.

**Abhilekh Digitālaya™**
Abhilekh Digitālaya™ is a specialized archival system designed and developed for the preservation of digital copies of government records, microfilms, cartographic maps, photographs and various types of documents available with national, state and district level archives. This system is being developed for the pilot digital repository at National Archive of India, New Delhi, India which is expected to received digital records from ministries and central government organizations.

**Datāntar™ for Reformatted Digital Content**
This is another type of Datāntar™ software which provides a library of best practices that can be applied for aligning and reformatting the digital content in terms of images, audio, video, PCDs etc for preservation purpose.

**Sanskriti Digitālaya™**
Sanskriti Digitālaya™ is a specialized archival system designed and developed for the preservation of cultural digital data in terms of images, audio, video, PCDs, 3D models etc. This system is being developed for the pilot digital repository at Indira Gandhi National Centre for Arts, New Delhi, India.
Digital Preservation Planner (PrePlanner)

Digital preservation planner software provides a searchable database of record retention rules and policies formulated by various ministries and departments of Government of India. It also allows the users to define file format specifications for images, audio and video data for estimating the digital storage requirements.

The abovementioned software tools are being developed with adequate compliance with the guidelines provided in various ISO standards. The functional prototypes of some of these tools are currently in the testing phase.


The team of Centre of Excellence for Digital Preservation at C-DAC, Pune along with the Expert Committee has developed the following digital preservation standard and best practices for e-governance to ensure that the electronic records are produced in a preservable manner. The digital preservation standard has been notified and adopted for all e-governance applications by the Ministry of Communications and Information Technology, Government of India.

Best Practices and Guidelines for Production of Preservable Electronic Records

This standard provides the best practices and guidelines for production of preservable electronic records and its management in the context of e-governance. It is applicable for those e-records that need to be retained for long durations (e.g. 10 years, 25 years, 50 years and beyond) and the e-records that need to be preserved permanently. The core concepts of ‘preservability’ are based on the requirements specified in IT ACT, ISO/ TR 15489–1 and 2 Information Documentation-Records Management and ISO 14721 Open Archival Information Systems (OAIS) Reference Model. It introduces 5 distinct steps of e-record management i.e. e-record creation, e-record capturing, e-record keeping, e-record transfer to trusted digital repository and e-record preservation which need to be adopted in all e-governance projects.

Fig. 2: Electronic Records Management Practice as Introduced in the Digital Preservation Standard for e-governance Applications
Standard for Preservation Information Documentation (eGOV-PID) of Electronic Records

It provides the standard metadata dictionary and schema for describing the "preservation metadata" of an electronic record. This standard proposes to capture most of the preservation information (metadata) automatically after the final e-record is created by the e-government system. Such preservation information documentation is necessary only for those e-records that need to be retained for long durations (e.g. 10 years, 25 years, 50 years and beyond) and the e-records that need to be preserved permanently. This standard allows to capture the preservation metadata in terms of cataloging information, enclosure information, provenance information, fixity information, representation information, digital signature information and access rights information. The implementation of this standard helps in producing the valid input i.e. Submission Information Package (SIP) for archival and preservation purpose as per the requirements specified in the ISO 14721 Open Archival Information Systems (OAIS) Reference Model.

Fig. 3: Seven Sections of Preservation Metadata as Per the eGOV-PID Standard

During the public review of digital preservation standard, several suggestions are received as the next agenda items for us to pursue.


As instructed by Department of Electronics and Information Technology (DeitY), the team of Centre of Excellence for Digital Preservation at C-DAC, Pune has drafted the National Digital Preservation Policy and Digital Preservation Rules under the IT Act 2000/2008. DeitY is expected to organize the consultations with various experts and government departments to finalize the policy and rules.
The policy and rules are meant to offer an overarching mandate across various domains and organizations and help in developing the national digital preservation infrastructure in the form of Trusted Digital Repositories in India.

**TRUSTED DIGITAL REPOSITORIES AS NATIONAL DIGITAL PRESERVATION INFRASTRUCTURE FOR INDIA**

Our teams are building the various layers of Trusted Digital Repositories in terms of standards, best practices, tools, systems, audit competencies, etc. Trained human resource will be necessary for managing and auditing the repositories. ISO accreditation of the organizations for conducting the repository audits is also necessary. The national policy will drive the creation and sustenance of the Trusted Digital Repositories across India. The digital preservation infrastructure will include large scale digital storage systems and high speed gigabit network connectivity between the repositories and remote backup (disaster recovery) sites.

![Diagram of Trusted Digital Repository](image)

**ISO 16363:2012 on Audit and Certification of Trusted Digital Repositories**

Fig. 4: Layers of Trusted Digital Repository

**INTERNATIONAL COLLABORATIONS**

The Centre of Excellence for Digital Preservation at C-DAC, Pune is collaborating with Alliance for Permanent Access (APA), InterPARES Trust, Canada and several experts from NDIIPP, USA in order to bring the state-of-art knowledge and best practices to India.
BENEFITS OF DIGITAL PRESERVATION

Development of trusted digital repositories for digital preservation at the national level has following benefits to offer:

Administrative Continuity
The digital information becomes inaccessible or obsolete much before its assigned retention period.

Protection of Digital Intellectual Assets
The wealth of knowledge and intellectual assets of an organization are increasingly encoded in digital formats which requires to be retained and protected. The digital information represents intellectual property which is produced with considerable amount of time, effort and money.

Reuse
Digital preservation can help in immediate, near term as well as long term usability and meaningful use of digital information beyond its primary users. Repositories of e-records and the tools to mine, analyze and re-purpose them represent a society's intellectual capital.

Long Term View
Access to e-records from the past and digital continuity is critical for planning, trend analysis, decision making and research.

Legal Obligations
Digital preservation helps in fulfilling the legal obligations of record keeping and record retention.

Protection from Litigation
The international best practices of digital preservation enable in ensuring availability of electronic records in a legally admissible manner.

Digital Heritage for Future Generations
The information, knowledge, cultural and historical artefacts in the modern digital age are being created, encoded and stored through digital means which forms the digital heritage for future generations.

CONCLUSION

- National Digital Preservation Infrastructure should be developed as a long term goal.
ISO accreditation of the organizations for audit and certification of Trusted Digital Repositories should be taken up on priority.

Training programmes for digital preservation, and audit and certification of repositories should be introduced in India.

Pilot digital repositories across various domains should be developed.

The Centre of Excellence should be funded on continuing basis considering the vastness of the requirements across diverse domains.

REFERENCES

[2] Best practices and guidelines for production of preservable electronic records, Published by Department of Electronics and Information Technology (DeitY), Ministry of Communications & Information Technology (MCIT), Government of India (GoI), December 2013.
[3] Standard for Preservation Information Documentation (eGOV-PID) of Electronic Records, Published by Department of Electronics and Information Technology (DeitY), Ministry of Communications & Information Technology (MCIT), Government of India (GoI), December 2013.
Trust in Electronic Records
Preservation in the Cloud: Towards an International Framework for a Balance of Trust and Trustworthiness

Luciana Duranti
Archival Studies Program, School of Library, Archival and Information Studies, The University of British Columbia, Vancouver, B.C. Canada
E-mail: luciana.duranti@ubc.ca

Abstract—The paper examines the benefits and risks of using Cloud-based Internet services for storing data, records, or archives; discusses the issues raised specifically by the long term preservation of data, records and archives in the Cloud—focusing on the problems presented by location independence and the inability to prove the trustworthiness of the materials held by Cloud providers; and presents the progress made towards the development of solutions based on an international framework encompassing model policies and legislation.

INTRODUCTION

On 30 October 2013, the BBC News reported that India was “planning to impose a ban on the use of foreign cloud-based email services to send official communications, before the end of the year. It would prevent civil servants [from] using Gmail, Yahoo! or Outlook.com. Instead they would be required to use a service provided by the country’s own National Informatics Centre (NIC). The move follows the publication of leaks about US cyber-spying operations…. India is not the only country implementing such measures. Earlier this month Brazil’s president confirmed her country planned to set up its own secure, encrypted email service to ‘prevent possible espionage’.”

There is no doubt that the use of cloud-based Internet services presents many access issues, some apparently insurmountable, and most not even related to espionage or loss of privacy, but is domestic legislation prohibiting or controlling it the best way of addressing them? Furthermore, are the key issues all related to access, or are there other challenges, especially for those who are using the Cloud for preservation purposes? Can the Cloud be internationally regulated? Do we need transparency or would oversight solve most problems? Are cloud providers adversaries to be kept at a distance or potential allies to involve in the search for a solution viable for all?

After a brief discussion of the benefits and risks of cloud-based Internet services, this paper will focus on their use for long term preservation and the key issues that need to be addressed by preservers before they can safely entrust data/records/archives selected for permanent preservation to the Cloud, and it will conclude by outlining the solutions identified by cloud providers as well as academic and professional researchers to ensure that “preservation in the Cloud” will not be an oxymoron.

THE CLOUD: A RISK ANALYSIS

Often the Internet is referred to as the Cloud. However, technically, this is a misuse of terms. The 2001 Budapest Convention on Cybercrime defines Internet providers as “entities providing users the ability to communicate through a computer system that processes or stores computer data on behalf of such communication or users.” Therefore, there are three ‘actions’ related to the definition of Internet provider: communication, data processing and data storage. In contrast, the National Institute of Standards and Technology defines cloud computing as “a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”

Choosing the Cloud is a risk assessment decision. A risk is understood as the probability that an adverse event will occur, multiplied by the impact that such event would have. It is really a question of comparison among available choices: if it is not possible to have all one wishes to have, what is one to give up with the least consequences? Cloud Services Providers (CSP) offer two choices among things that most organizations would consider all non-renounceable. The first choice is between Transparency and Security: they offer “trust through technology” rather than through knowledge, objectivity, repeatability, and verifiability. Security based on technology involves location independence, one of the three core aspects of the Cloud delivery model, because most of its measures, such as sharding and obfuscation cannot happen without spreading the records among servers in data centers.

4Sharding is a type of database partitioning that separates very large databases into smaller, faster, more easily managed parts called data shards. The word shard means a small part of a whole. Jason Tee explains sharding on The Server Side: "In the simplest sense, sharding your database involves breaking up your big database into many, much smaller databases that share nothing and can be spread across multiple servers." See: http://searchcloudcomputing.techtarget.com/definition/sharding. Accessed on November 2, 2013. Obfuscation is the process of intentionally making code harder to read, for security purposes. It protects applications from source code extraction and subsequent reverse engineering by renaming identifiers, making metadata invisible, and altering the code control flow.
around the globe. The second choice CSPs offer is between Control and Economy: they give customers “control on expenditures” rather than on how data/records/archives are managed. However, there is a tension between juridical systems that protect records in a traditional way and the delegation of control without transparency and responsibility. To identify where the tension resides and the breaking point, it is useful to conduct a risk-benefit analysis of the use of the Cloud.

**Benefits of the Cloud**

The most evident benefit of choosing the Cloud is Cost. Organizations do not have to own hardware/software, and can avoid huge upfront costs. In addition, there are reduced information technology personnel costs, as organizations do not have to implement or maintain a preservation system; reduced energy costs; and no cost for state of the art technology, because the shared-tenant system allows pooling of resources to get more for less and better hardware/software and network. Linked to cost is Scalability. Each organization can get whatever platform, infrastructure or services it needs and only pay for what it uses. Furthermore, such use can be tracked and increased or reduced as needed. And CSPs advertise their Reliability, as the material will always be there on demand, irrespective of whether it is data, one record, a file or a whole archives, and available from anywhere in the world, simply using a browser. Security is more robust than any one organization could afford otherwise—both physical and virtual. The reason, in addition to those mentioned earlier, is centralized control on the material. Collaboration among users is enhanced in the Cloud not only because people can access the same material from different geographic locations, but also because all files are in consistent format, viewed in a web browser. Furthermore they can distribute to co-researchers the material they access through means like Google Docs or Dropbox.

**Risks of the Cloud**

Cost represents not only a major benefit but also a significant risk, as proven by a survey of users conducted by the Records in the Cloud project. If one calculates transfer of the material that is not already in the Cloud, implementation of all controls, subscription, and encryption, costs are not insignificant. One can also get unexpected license fees, which may contribute to high variability of costs, as usually there is no set monthly fee. Also, there are significant per-request charges, to motivate access in large chunks. Although some providers allow access to 5% of one’s material each month with no per-byte charge, the details are complex and hard to model, and the cost of going above the allowance is high. There are three different business models for storage: 1) the storage can be rented; 2) storage can be monetized, selling ads against one’s accesses to one’s own material; or 3) storage can be endowed, that is, material can be deposited together with a capital sum thought to be enough to fund its storage. For organizations like memory institutions the latter approach is the most

---

sensible but it is difficult to figure out how big the endowment needs to be to cover the costs that will be incurred in the future.

CSPs’ Reliability constitutes a big issue because a Cloud can go bankrupt, disappear or be sold, and its holdings might be gone with it. Also, the Cloud can lose individual documents or files, and sometimes cannot get them back or the backups may fail.

Security, the biggest benefit of the Cloud, presents also the biggest risk. There can be unauthorized access provided to and by sub-contractors, and hackers tend to attack Cloud environments much more than in-house systems. It is not a matter of if but when a breach will occur. Are Cloud customers told when it does? It is not likely, if even CSPs are denying awareness of breaches by government agencies. Encryption might not be done-in transit whenever documents are moved from a center to another. As the technology is relatively new, non-technical users may not be able to set up their system in a way that their data are protected. Furthermore, documents from different files, folders, and archives could be intermingled within shared servers. And, if servers fail, are sequestered, or become obsolete, all the material in them becomes inaccessible to all institutions or organizations using them.

Control is what organizations lose by putting their materials on the Cloud. In a public Cloud, they would have no control over whom they share their Cloud with or to whom services are delegated by the primary provider. Also the terms of service or privacy policy may change after a contract has been signed, and a contract may be terminated and providers and sub-providers may also be terminated. This would determine lack of sustainability for information and records management and archival functions and lack of portability and continuity for the holdings. As audit is not allowed in the Cloud, it is not possible to know whether back-up is regularly done, whether material is lost or has become inaccessible as a consequence of hardware/software obsolescence, and whether accruals of existing aggregations of materials are integrated with them.

Transparency is the most evident victim of a transfer of data/records/archives into the Cloud. As the legitimate chain of custody which guarantees the trustworthiness of records as reliable sources of past actions and events cannot be demonstrated in the Cloud, authenticity cannot be inferred from known processes and it is easily challenged by the fact that materials in the Cloud can be hacked and tampered with. This also implies that data and records in the Cloud cannot have forensic integrity, in that processing and preservation is not repeatable, verifiable, or objective, as prescribed by digital forensics authentication processes, therefore they cannot be used as evidence. Neither can materials in the Cloud have duplication integrity, which means that, given a data set, the process of creating a duplicate of the data does not modify the data either intentionally or accidentally and the duplicate is an exact bit copy of the original data set. This type of integrity is extremely important to records managers and archivists because they can only maintain and preserve digital records by reproducing them, so transparency about the reproduction and migration processes is essential to trusting records. In fact, process integrity is based on two fundamental principles, the principle of non-interference and the principle of identifiable interference. The former means that the method used to gather and analyse, maintain and use, or acquire and preserve digital material does
not change the digital entities; the latter means that, if the method used does alter the entities, the changes are identifiable. These principles, which embody the ethical and professional stance of digital information and records managers and archivists, are consistent with the traditional impartial stance of such professionals, as well as with their responsibility of neutral third party, or trusted custodians.6

Privacy Risks are also not to be underestimated. The European Union (EU) new Data Protection Directive (which it is expected to be issued in March 2014) aims to create a single set of binding data protection rules across the EU which will replace the 1995 EU directive currently in use. Like the previous one, it does not allow transfer of personal information of EU residents or its processing to countries that do not have similar privacy protection. But this new directive goes much further: “The text sets up a legal framework in an effort to coerce US-based companies from indiscriminately passing on the personal details of EU citizens to US law enforcement and its intelligence agency. A company would face fines, on the basis of the European Union law, if the transfer took place without the legal basis. ‘This article has now been included in the compromises accepted by all political groups in this house,’ said Albrecht. … ‘The right to be forgotten is now called the ‘right to erasure’ but the meaning remains similar to what is outlined in the 1995 directive,’ says Albrecht. ‘The right we are talking about here is the right to deletion and the right to erasure’. If a person asks an Internet giant to remove his personal data, then the company must also communicate the request to others where the data is duplicated.”7

Clearly, at this time, the risks of storing data/ records/ archives in the Cloud by far outweigh the benefits, because the standard of protection in such environment is what legal theory refers to as caveat emptor or ‘consumer beware.’ Yet, increasingly, heritage institutions—especially archives—are entrusting their holdings to the Cloud without realizing that the idea itself goes against the conceptual core of their identity.

Preservation in the Cloud: An Oxymoron?

Heritage institutions and archival units or programs of a variety of organizations consider storing data/ records/ archives selected for permanent preservation in the Cloud for a variety of reasons. In the most technologically advanced countries this happens because many of the data and records these institutions, units or programs are mandated to preserve already exist in the Cloud, having been generated or stored there in the usual and ordinary course of business, or by individuals using cloud technologies as their normal way of interacting both at personal and social level. The advantage these data and records creators see in keeping their materials in the Cloud is easy access from any location to anyone who can use a browser and ability to collaborate and socialise across geographic distant areas. For most institutions and organization in both high and low resources countries, the motivations for resorting to long term storage in the Cloud tend to be that a) a trusted digital repository satisfying ISO standards as well as basic preservation requirements is not affordable, b) the knowledge to

---

deal with data and records produced by complex technologies is not commonly available among preservation professionals and is very expensive, and c) strong protection measures are often confused with preservation measures.

While these motivations are understandable, heritage institutions and among them archives in particular encounter serious issues when they decide to entrust their holdings to the Cloud, issues that reside at the root of what they are and what their purpose is. The most important are location independence and trustworthiness.

Location Independence

A fundamental issue with preserving data/records/archives in the Cloud remains the distinction between the entity responsible for their preservation and accessibility and the entity storing and providing access to them, in relation to the likelihood that the jurisdiction under which either exists is different from that in which the material resides at any given moment in time.

In both Western and Eastern culture the concept of place is at the core of the nature of archives as the trusted custodians of documentary memory. For millennia the place where this memory is kept has been as important to its permanence and quality as the knowledge of the professionals responsible for it. In the Justinian Code, which is the summa of all Roman law and jurisprudence, an archives is defined as locus publicus in quo instrumenta deponuntur (i.e., the public place where deeds are deposited), quatenus incorrupta maneant (i.e., so that they remain uncorrupted), fidem faciant (i.e., provide trustworthy evidence), and perpetua rei memoria sit (i.e., and be continuing memory of that to which they attest)\(^8\). Generally, in the ancient world, an archives was a place of preservation under the jurisdiction of a public authority. This place, public as well, endowed the documents that passed its threshold with trustworthiness, thereby giving them the capacity of serving as evidence and continuing memory of facts and acts.

In 1664 though, a German jurist, Ahasver Fritsch, specified that archival documents did not acquire trustworthiness simply by crossing the archival threshold, but from the fact that 1) the place to which they were destined belonged to a public sovereign authority, as opposed to its agents or delegates, that 2) the officer forwarding them to such a place was a public officer, that 3) the documents were placed both physically (i.e., by location) and intellectually (i.e., by description) among authentic documents, and that 4) this association was not meant to be broken.\(^9\) The “archival right,” that is, the right to keep a place capable of conferring trustworthiness, and therefore authority, to the documentary by-products of activities—the right to hold archives, was in time acquired by all those bodies to which sovereignty was delegated by the supreme secular and religious powers—among these, cities, universities, churches and corporations of every kind.\(^10\)

---

\(^8\) Justinian: Corpus Juris Civilis, Novella 15 De Defensoribus civitatum, Et a defensoribus. Digestum 48, no. 19 De Poenis. Codex I, no. 4 De episcopali audientia. (534 A.D.)


This remained the state of affairs in the Western world until the French revolution, when, with the decree of July 25, 1794, the documents of defunct institutions and organizations, concentrated in the National Archives of France, were declared the patrimony of the nation and made accessible to the public by virtue of its ownership of them. With this declaration, the State recognized its duty to preserve them as historical memory for the subsequent generations and entrust them to archival institutions open to the public.\footnote{Duranti, L.: Archives as a Place. Archives & Manuscripts 24, 2 (1996) 242-255. Republished in Archives & Social Studies: A Journal of Interdisciplinary Research, 1, 1 (2007); Posner, E.: Some Aspects of Archival Development Since the French Revolution. In Munden, K. ed.: Archives and the Public Interest: Selected Essays by Ernst Posner. Washington (1967).}


Similar circumstances in different contexts have been examined by research projects like Records in the Cloud, mentioned earlier, in order to find usable models. Maritime rules of shipping have been closely examined as they center on the recognition of the authority of three separate states: the port state, the flag state and the coastal state. Early international maritime agreements established that the nationality of the transport vessel (the flag state) would establish jurisdiction, and by extension, the laws that would be in effect. Following the abuse of such rule, the port state was given greater control to inspect vessels coming within its territorial waters by the Law of the Sea Convention in 1982. Similarly, coastal states through whose waters the flagged vessels transit, were given authority over the safety and competency of the ship and its crews and allowed inspection and enforcement while the vessel is in the coastal state’s waters, regardless of the flag of either the vessel (flag state) or its destination (port state). On the basis of these rules, an analogy can be made. A Canadian university could place its archives into the care of an American CSP which in turn maintains its data centers in Brazil: the American company would be the ‘flag state’ that would be ‘moving the goods’ through ‘coastal states’ to their ultimate destination in the ‘port state’ of Brazil. This analogy is problematic not only because the Canadian University owning the archives would have no jurisdiction, but also with regards to the rights of the coastal state, in that the ‘pipe’ used to move the records can transit through several countries (coastal states) as they are routed along the way and, traditionally, ‘coastal states’ have not been granted access to inspecting packets of records as they move along the Internet. The rules of conduct then become very difficult, if not impossible, to enforce by any of the parties involved.\footnote{Duranti, L. and Jansen, A.: Records in the Cloud: Authenticity and Jurisdiction. In Digital Heritage 2013, Marseille, France, 28 October-1 November 2013. Conference Proceedings. UNESCO, Paris (2013).}
To address this jurisdictional problem linked to location, other alternatives have been considered with unsuccessful results: the *territoriality principle* is not applicable because it is not possible to know the location of the records at any given time; the *nationality principle* is not applicable because nationality is an attribute of persons, not things, and the principle cannot be used to connect persons to data or records; the *power of disposal* principle, which “connects any data to the person or persons that obtain sole or collaborative access and that hold the right to alter, delete, suppress or to render unusable as well as the right to exclude others from access and any usage whatsoever” was briefly considered, but did not address the issue of the different jurisdictions of custodians and CSPs; and a *power of preservation principle* was conceived that assigns jurisdiction to the institutions controlling the archives as the trusted custodian and the place guaranteeing authenticity, but jurisdiction with responsibility without actual custody defeats its entire purpose.

Thus, the issue of location independence remains open. Ideas about ways of addressing it will be discussed after having presented the other key issue for long term preservation in the Cloud, the continuing trustworthiness of the material.

**Trustworthiness**

Eliminating the long term custody of and provision of access to data/records/archives from the functions and responsibilities of archival institutions, programs or units makes it impossible for them to protect the material for which they are accountable, thus such material cannot serve as trustworthy sources and evidence of facts and actions. Archival preservation is controlled by very strict principles: the principle of *respect des fonds*, according to which data and records created by different bodies cannot be intermingled; the principle of *provenance* (i.e. original order), according to which archival materials must be kept in the same order in which they had originally accumulated; and the principle of the *unbroken chain of legitimate custody*, according to which the custodial history of data and records must be demonstrable at any given time or the material will not be considered authentic. The physical and moral defence of archives is provided by three primary factors: *transparency of preservation, security and stability*. Where archival institutions have evolved over the millennia to be the ideal providers of these three requirements, there are concerns with CSPs’ ability to maintain the trustworthiness of records that have been transferred to them, and to prove it, considering their efforts to remove themselves from liability and responsibility for the contents entrusted to them.

In archival science, trustworthiness comprises three concepts, reliability, accuracy and authenticity. Reliability is the trustworthiness of content, that is, of data/documents/records as statements of facts. It is assessed on the basis of the competence of the author and the controls exercised on records creation. Accuracy is the correctness and precision of such

content and is assessed on the same factors and on the controls exercised on recording and transmission. Authenticity is the trustworthiness of data/documents/records as such in that they are what they purport to be, uncorrupted and untampered with, and it is assessed on their identity and integrity as shown by external attributes—in the digital world, by metadata. In a Cloud environment, the identity metadata (names of person, action, date, etc.) follow the entity to which they relate since creation as they come into being together with such entity, but the integrity metadata, which relate to actions carried out on the data/documents/records through time and to their consequent transformation as they move from creation, use, and maintenance to preservation, often through encryption, conversions, migrations, etc., are added by the CSP storing the material to which the metadata relate.

Thus, the questions that need to be answered are: how does metadata follow or trace data/records/archives in the Cloud from the creator to the preserver? How is this metadata migrated as a preservation activity over time? Who owns the metadata created by the service providers related to their management of the records (integrity metadata)? Is metadata intellectual property? If yes, whose property? How can this metadata be accessed by the public and what are the responsibilities of the provider towards users of data/documents/records preserved over the long term for a community or the public at large? Until such time when we will be able to unequivocally answer these questions there will be no transparency, stability and permanence in the Cloud. There will be no transparency because an unbroken chain of legitimate custody is not possible or demonstrable; reliability cannot be inferred from known management processes; and authenticity cannot be inferred from documentary context and a known preservation process. There will be no stability because archives require that each entity’s context be defined and immutable, with all its relationships intact and this is difficult to demonstrate in the dynamically provisioned environment of the Cloud. There will be no permanence because no provisions are taken or known to be taken about hardware/software obsolescence, records portability and continuity in case of termination of contract, and records sustainability in case of termination of provider.

CSPs are fully aware of these issues and have not stayed idle through the various legal and other problems that have ensued from the lack of regulation in their environment. Academic and professional researchers have developed initiatives to find solutions. And now the two groups are beginning to get together in earnest in an interdisciplinary effort.

MAKING IT POSSIBLE: TOWARDS AN INTERNATIONAL FRAMEWORK FOR A BALANCE OF TRUST AND TRUSTWORTHINESS

At the opening of the International Conference on Cloud Security Management ICCSM 2013, Howard A. Schmidt, technological advisor of a couple of US President, stated that “technology is not the answer to everything,” and that we “need to be prepared to change course” thinking of the future. Following his intervention, Jim Reavis, head of the Cloud Security Alliance, emphasized the facts that at this time the Cloud lacks transparency and

\[For\ the\ International\ Cloud\ Security\ Management\ ICCSM\ 2013\ conference\ website\ go\ to:\ \text{http://academic-conferences.org/iccsm/iccsm2013/iccsm13-home.htm.}\]
visibility, is incompatible with existing legislation, is not regulated by comprehensive standards, and does not make use of true multi-tenant technology. There is a risk of concentration without clear separations and the tendency to maintain logical control losing physical control. He compared the relation between failure of Cloud storage and of in-house storage with the relation between a plane crash and a car crash: although supposed to be reassuring, this is not a pretty picture!\textsuperscript{16} Thus, the Cloud Security Alliance (CSA) is taking a number of initiatives to address the issues identified by Reavis. Although the Cloud will never be 'transparent', there is a willingness to substitute transparency with trusted oversight. As a voluntary industry action promoting transparency, then, the CSA has developed a Security, Trust, and Assurance Registry (STAR), which is a public registry of CSPs evaluations based on a Cloud Control Matrix, a Cloud Control Trust Protocol, and a Cloud Audit on how the CSP makes assertions. This Registry includes the results of an open certification framework (an approach supported by Internet 2) including three steps: self-assessment, third party assessment-based certification and attestation, and continuous monitoring-based certification.\textsuperscript{17} However, the CSA recognizes that all of this will not completely protect their users from the nine top threats, which they identify as: 1. Data Breaches, 2. Data Loss, 3. Account Hijacking, 4. Insecure APIs, 5. Denial of Service, 6. Malicious Insiders, 7. Abuse of Cloud Services, 8. Insufficient Due Diligence, and 9. Shared Technology Issues.\textsuperscript{18} The CSA focus on addressing all these threats is on: global legal issues, privacy level agreements, continuing monitoring, harmonization of various requirements, regulations and standards—especially in relation to smart mobiles and anti-bot, and community policing (the already mentioned trusted oversight). But they cannot do it by themselves. Jim Reavis, in his presentation, made a key understatement “metadata are important.” Indeed, metadata for trust. But what does the CSA mean by “trust” or “trusted”? Is the CSA framework sufficient for trusted long term preservation? How can we build an overarching international framework globally accepted across sectors, disciplines, and governments?

Trust

Trust has been defined in many ways. In business, trust involves confidence of one party in another, based on an alignment of value systems with respect to specific benefits in a relationship of equals. This is in my view the way in which the CSA interprets the phrase ‘trusted oversight’. But, of course, if CSPs want trust from their customers, the government, and society, this interpretation of trust is not sufficient. In jurisprudence, trust is usually described as a relationship of vulnerability, dependence, and reliance in which we participate voluntarily. In substance, trust means having the confidence to act without the full knowledge needed to act. It consists of substituting the information that one does not have with other information. Traditionally, trust in records is based on four types of information about their

\textsuperscript{16}Reavis, J.: Keynote speech, 18 October 2013, ICCSM 2013, Seattle, USA.
\textsuperscript{17}For the CSA Open Certification Framework, go to: https://cloudsecurityalliance.org/star/.
custodian: reputation, which results from an evaluation of the trustee’s past actions and conduct; performance, which is the relationship between the trustee’s present actions and the conduct required to fulfill his or her current responsibilities as specified by the trustor; competence, which consists of having the knowledge, skills, talents, and traits required to be able to perform a task to any given standard; and confidence, which is an assurance of expectation of action and conduct the trustor has in the trustee. These attributes are contextually based in the value systems of a trustor and a trustee who interact within a close social sphere.

The challenge of developing a conceptual understanding of trust—an articulation of its distinctive characteristics—must be addressed before it can be possible to approach the challenge of facing practical issues of trust in the Cloud environment. Can a general model of trust be developed that crosses disciplinary and social boundaries? Some would say that it would depend on harmonizing terminology, definitions, and concepts across disciplines and jurisdictions, and that this is a tall order. Currently, for example, there are more than six distinct definitions of trust within ISO standards alone. If it is not possible to develop a

---


20 This information comes from an oral presentation by Giovanni Michetti delivered at the Peter Wall Institute Exploratory Workshop entitled “Trust and Conflicting Rights in the Digital Environment” on September 24, 2013. See: http://www.digitaltrust.pwias.ubc.ca. Entity X is said to trust entity Y for a set of activities if and only if entity X relies upon entity Y behaving in a particular way with respect to the activities. ISO/IEC 10181-1:1996, 3.3.28. Information technology—Open Systems Interconnection—Security frameworks for open systems: Overview; Degree to which a user or other stakeholder has confidence that a product or system will behave as intended. ISO/IEC 25010:2011(en), 4.1.3.2. Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models I; Confidence; a basis of reliance, faith, or hope; assured reliance on the character, strength, or truth of someone or something. ISO/TR 18307:2001, 3.141. Health informatics—Interoperability and compatibility in messaging and communication standards—Key characteristics; ISO/TR 21089:2004, 3.88. Health informatics—Trusted end-to-end information flows; Quality by which an entity can be said to “trust” a second entity when it (the first entity) makes the assumption that the second entity will behave exactly as the first entity expects. ISO/TS 22600-2:2006, 2.43. Health informatics—Privilege management and access control—Part 2: Formal models; Relationship between two elements, a set of activities and a security policy in which element x trusts element y if and only if x has confidence that y will behave in a well-defined way (with respect to the activities) that does not violate the given security policy. ISO/IEC 13888-1:2009(en), 3.59. Information technology—Security techniques—Non-repudiation—Part 1: General; an entity can be said to “trust” a second entity when it (the first entity) makes the assumption that the second entity will behave exactly as the first entity expects. ISO/TS 22600-3:2009, 3.106. Health informatics—Privilege management and access control—Part 3: Implementations; Generally, an entity can be said to “trust” a second entity when it (the first entity) makes the assumption that the second entity will behave exactly as the first entity expects. This trust may apply only for some specific function. The key role of trust in this framework is to describe the relationship between an authenticating entity and an authority; an entity shall be certain that it can trust the authority to create only valid and reliable certificates. ISO/IEC 9594-8:2005, 3.3.59. Information technology—Open Systems Interconnection—The Directory: Public-key and attribute certificate frameworks—Part 8; ISO/IEC 9594-8:2008, 3.4.64. Information technology—Open Systems Interconnection—The Directory: Public-key and attribute certificate frameworks—Part 8.
common language of trust with shared definitions, we can work toward a common understanding with domain-specific theories, techniques, and metrics, based on an exploration of the concept of trust through a variety of perspectives.

The issues inherent in creating, managing, storing, and preserving records and data in distributed networked environments cross national and international boundaries, and are of interest and concern to all sectors of society. In the digital environment, current policies, practices and infrastructure cannot be readily applied to the technologically-mediated trust in records relying on the four types of knowledge used in the past. Different systems for the assessment of trust are required for different contexts—government, business, personal, etc. The parameters of trust in one cultural context may be very different from those in another context. These parameters must be identified and understood if cross-cultural, or international, trust is to be achieved. This is not easy when, even within the restricted confines of the Western world, the very limited portion of a cultural context which is represented by the legal system is broken down in common law and civil law, and each has a different approach to trust, the former based on observation of action, and the latter on its documentary residue.  

Trust can also be seen as an empirical process in which confidence is built over time by applying known patterns to given situations. In this context, the development of trust is understood to be incremental and informed by personal experience. We apply our own judgment to assess the trustworthiness of an object, actor, or situation. This is contrasted with a scientific assessment of trust, which relies on repeatable actions that produce the same result every time they are carried out. For a scientific assessment of trust, we often place our confidence in experts, thereby shifting the responsibility of trust assessment from ourselves to a recognized authority.

21In the realm of common law, trust involves entrusting some thing to some one—land, money, stock portfolio, etc.—and has a direct relationship with trustworthiness. Standards of trustworthiness may be identified with levels of decreasing expectations of responsibility. The highest level is the fiduciary relationship. This is a legal or ethical relationship of trust between two or more parties. A fiduciary relationship involves a trustee—a person in some position of vulnerability, and a trustor, whose aid, advice, or protection is sought on the basis of its justifiable trustworthiness. The next level is called good-faith/fair dealing. This is the general assumption that people will generally act in good faith and deal fairly with others without breaking their word. A significantly lower level is that of unconscionability, where the trust relationship may be based on terms which are excessively unfair to one party. This is a particular issue with click-through contracts on the Internet, which place the user in a “take it or leave it” position, generally with no ability to negotiate the terms of the contract. We are familiar with such agreements when licensing software online, as we cannot license and use the application without agreeing to the terms of the license. The risk to our privacy and security is minimal. However, service agreements for social media or free online storage carry greater risks to the security of our information. The level with the least expectation of responsibility is that of the ordinary marketplace, caveat emptor, or buyer beware. This level applies in the same social media sites between users, or “friends.” This information comes from an oral presentation by Anthony Sheppard delivered at the Peter Wall Institute Exploratory Workshop entitled “Trust and Conflicting Rights in the Digital Environment” on September 24, 2013. See http://www.digitaltrust.pwias.ubc.ca.
If we decided to entrust our historical documentary memory to the Internet, we must distinguish between trust and trustworthiness, because a trust relationship must be based on the balance between the two. The trustworthiness we should focus on is then not of the trustees but of data/records/archives that are entrusted to them for long term preservation, keeping in mind that historical data and records, a society documentary heritage, always start their life as current data and records and their trustworthiness should be protected from the moment of their creation.

Protecting the trustworthiness of the documentary heritage of society goes well beyond security, as shown earlier in the discussion of what trustworthiness is. For this reason, a large team of researchers involving organizational partners from six continents and thirty countries has begun a new research endeavour called InterPARES Trust.

InterPARES Trust

The goal of InterPARES\textsuperscript{22} Trust is to generate the theoretical and methodological frameworks that will support the development of integrated and consistent local, national and international networks of policies, procedures, regulations, standards and legislation concerning digital records entrusted to the Internet, to ensure public trust grounded on evidence of good governance and a persistent digital memory.\textsuperscript{23}

The objectives of this research are: 1) to discover how current policies and practices regarding the handling of digital records by institutions and professionals affect the public’s trust in them, in light of the exponential growth of and reliance on Internet services; 2) to anticipate problems in maintaining any trust in digital records under the control of entities suffering a waning level of confidence from the public (including government, community and heritage institutions); 3) to establish what significance national/cultural contexts have with regard to the level of trust digital records on the Internet enjoy; 4) to articulate model policies, procedures, and practices for storing, managing, accessing, and preserving records on the Internet, and test them in a variety of contexts so that, from them, international standards, guidelines and best practices can be developed, and 5) to formulate proposals and models for law reform, and functional requirements for the systems in which Internet providers store and manage digital records.

Although the focus is on the relationship between organizations and their particular client groups (citizens, scholars, students, etc.), with client groups becoming concerned about the degree of ‘trust’ they can place on records generated and/or stored and accessed on the Internet and organizations becoming concerned about establishing and maintaining that trust, the same themes are also addressed within the context of organization to organization and

\textsuperscript{22}InterPARES (International research on Permanent Authentic Records in Electronic Systems) is a multinational multidisciplinary research endeavour funded by the Social Sciences and Humanities Research Council of Canada and by multiple partners around the world that began in 1998 and has since gone through three phases. The products of such phases can be found here: www.interpares.org. The fourth phase, InterPARES Trust, started in April 2013, is scheduled to be completed in 2018.

\textsuperscript{23}The website of InterPARES Trust is http://www.interparestrust.org/.
client group to client group relationships (e.g. does an archives trust the creator relinquishing ownership of and access privileges to its records in the Cloud? Or does an historian trust the sources used by another historian if they were accessed from the Cloud?). The project also studies the relationship between an organization and its own employees and the extent to which issues of trust are growing here as well, that is, how much employees can trust their own output (e.g. a virtual inventory of records in the Cloud) or that produced by others in their same organization, especially considering 1) the increasing popularity of “bring your own devise” policies, according to which organizations would not provide employees with information technologies but ask them to use their own androids, laptops, etc. to do their own institutional work; and 2) the expansion of mobile computing within organizations and the proliferation of “apps,” which are substantially changing archivists’ use of software while menacing central control of Information and Communication Technologies (ICT) and the records they produce.\textsuperscript{24}

The research theoretical framework has been adapted from archival and diplomatics theory, in particular the ideas that are foundational to trusting records.\textsuperscript{25} However, the project also uses theories adopted by the information systems management field to understand better the issues and address them. From an organizational point of view (thus, from the point of view of a memory institution), success in the use of the online environment is based on three categories of benefits: strategic, economic, and technological. The business view stresses the fact that choices like that of keeping an organization’s records on the Cloud must be based on capabilities that go well beyond physical and human assets to include “leadership, business-systems thinking, relationship building, architecture planning, contract facilitation and monitoring”, etc.\textsuperscript{26} Thus, the project will use resource-based theory, which focuses on the importance of technical, managerial, and relational capabilities for leveraging resources to maximize competitive advantage.\textsuperscript{27} Resource-based theory shows performance differences among organizations in the way they leverage these resources, and can help the project to identify and capitalize on the resources unique to cultures, societies, and types of organizations to articulate models that can work internationally.

Another framework relied upon by the information systems management field, among many other fields, is that of risk management, an area of study that complements that of trust and in a way represents its counterpart in the context of making decisions in an uncertain environment. Several models of trust exist but few have explored the relationship between risk and trust.\textsuperscript{28} Research and practice in the field of risk management offer both an

operational and a social perspective on trust. Operationally, risk managers reference the ISO 31000 Risk Management–Principles and Guidelines on Implementation framework, and, like record managers, seek to build best practices into business processes in order to support “good governance and accountability processes in organizations.”

Finally, as one of the project’s objectives is to design model policies that can be adapted within cultures, societies, and organizations that are fundamentally different but need to interact through their digital records, it draws upon design theory. This is necessary because these policies will need to address challenges arising from future technological interactions that we can’t yet imagine. Design theorists are adept at taking principled action in situations with many unknowns.

Design theorists Rittel and Webber argue that the best way of dealing with this sort of “wicked problem” is to adopt an “argumentative process in the course of which an image of the problem and of the solution emerges gradually among the participants, as a product of incessant judgment, subjected to critical argument.”

Thus, we are starting a conversation taking into account design perspectives, not just in terms of utilizing design theory, but through direct engagement with the designers of digital information technologies, the members of the CSA, like Amazon, Google, and Microsoft. The combined academic knowledge which addresses the identified questions harnesses expertise at the intersection of archival science, records management, diplomatics, law, digital forensics, visual analytics, organizational culture, information technology, information policy, information systems design, cybersecurity and information assurance. The empirical knowledge for this research project comes from the professions having the highest stake in those questions, i.e.: law and law enforcement, journalism, archives, records and information management, finances and security, health, computer engineering, etc. The partners who participate in this research project bring to it strong academic and empirical knowledge in all these diverse fields.

There is no doubt that the problem addressed by InterPARES Trust together with the Cloud Security Alliance constitutes an enormous and difficult challenge, one which we must attack less we fail to address important cultural, educational, scientific, social, governmental, business and practical needs which depend on access to trustworthy digital data and records. As Nathan and Shaffer point out in the article mentioned earlier, this kind of problem can never be solved for good and one can always do better. There is no singular “solution,” rather there is a continuous series of ever evolving processes based on a combination of theory and empirical inquiry that inform the design of systems, practices, procedures, and

---

32For a comprehensive list of partners see http://www.interparestrust.org/trust/aboutus.
possibly laws. The framework this research will develop will be useful in a variety of contexts, ranging from articulating a theory of online recordkeeping and formulating models for law reform and policy design, to developing specialized bodies of knowledge and skills; planning for and managing online repositories; developing and implementing strategies for particular sets of information objects; and defining the need for better recordkeeping technologies, guiding their development, and evaluating the relevance and adequacy of specific maintenance techniques.

CONCLUSION

To establish a “balance of trust” requires enabling Cloud providers to develop trustworthy technologies, procedures, and contractual conditions that can support the long term preservation endeavour by identifying the changes needed in our paradigms of trust in data/records and archives, and developing an internationally shared trust framework that providers, memory institutions and users can live by, because the current framework within which law enforcement operates and security concerns are addressed is inconsistent within and across jurisdictional boundaries.

Only then we can require and expect from Cloud providers transparency, compliance and accountability, in addition to security and economy of costs, and develop a degree of trust in the Cloud that will make the expression “Preservation in the Cloud” sound no longer as an oxymoron, but as a viable option based on trustworthiness as well as trust.
Authenticity in Records Systems:
Emerging Research in Digital Preservation

Adam Jansen
University of British Columbia, InterPARES Trust Research Team, Vancouver, Canada
E-mail: adamj@mail.ubc.ca

Abstract—Assessing and documenting the authenticity of digital records as they are transmitted from the record’s creator, received by the custodian and preserved over time is a challenge facing every digital repository. There exists a need for a base set of criteria and related implementation procedures for digital repositories that will support the acquisition, processing and preservation of authentic digital records. This paper discusses the concept of authentic digital records within records systems and presents an emerging research project analyzing the technological contexts of records systems that support the authenticity of records.

INTRODUCTION

The rapid advance of the technologies employed in information systems over the previous two decades, combined with the increased use and adoption of Internet-based resources, have created a dramatic shift in the way that records are created, managed and maintained. As these records systems decrease in price and increase in capability, more powerful and user-centric software interfaces are providing records creators the ability to create, store, modify, distribute and preserve digital records through organizational intranets, the Internet and disparate networked resources. This ‘information explosion’ poses a series of technical and administrative challenges for the custodians who must acquire the knowledge and the necessary technological infrastructure to acquire and maintain these digital records in such a way as to be able to attest to their authenticity over the long term.

Records have always possessed an innate value in that they serve as evidence of the activities in which they participate. In order to maintain this evidentiary capacity that is originally inherent in the records, records must be demonstrably authentic, that is, the record must be “intrinsically able to be proved that it is what it purports to be [13].” As such, the content within these records must not be substantively changed with the shift from one stored format to another (such as analog media (e.g. paper) to digital media or from one digital

---

1A record is a “document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference.” The InterPARES Project. Terminology Database. Retrieved from: http://www.interpares.org/ip2/ip2_terminology_db.cfm
format to another) or when moved from one computer system to another. The challenge for
the records custodian is that the skills, procedures and equipment required to properly
preserve these records in such a way as to be able to attest to their authenticity has changed
dramatically. Information Technology units further complicate determination of a record’s
authenticity within any given digital environment due to the great diversity of both the
computer products and services deployed in any given organization, as well as how the IT
work itself (e.g. migration, upgrades, new system installations) is planned and implemented.
The need to maintain the authenticity of digital records through time and space is a
necessary, unquestioned attribute of the records, yet the method with which this can be
accomplished is not as clear.

**THEORY OF AUTHENTICITY FOR THE DIGITAL AGE**

Extending the traditional concepts of diplomacy--The study of the nature of records and their
relationship to the facts represented in them--developed by Mabillion in the late seventeenth
century, InterPARES integrated traditional diplomatics with concepts and principles from
archival science to form what is known as archival diplomatics: “based on jurisprudence, the
history and theory of administration, and an extensive and centuries old body of written
reflection and experience about the nature of records and record-keeping practices in
bureaucratic organizations [11].” This concept of authenticity encompasses not only the
attributes contained within the digital records themselves, but also takes into account the
whole of the procedural contexts involved in the creation, transmission, maintenance and
preservation of those records throughout their entire lifecycle. Thus, the determination of the
authenticity of any given record is contingent upon the known facts regarding the creation
and handling of that record. To be considered authentic and capable of providing faithful
witness to the activities in which it participated, it is essential that the record be:

1. What it purports to be and
2. Free from manipulation, substitution, or falsification.

The authenticity of a record is a factor of establishing the record’s identity while
demonstrating the integrity of that record. The identity of a record is derived from the whole
of its attributes that when taken together characterize and distinguish that record from others
like it. Integrity refers to the record’s wholeness, in that it possess all of its necessary parts to
be effective, and its soundness, in that the condition of the record is unimpaired. Integrity is a
reference to the degree that the record is complete and uncorrupted in all essential respects;
that is, the record is capable of delivering the message it was intended to communicate in
order to achieve the purpose for which it was created. Where the identity of a record is based
on known facts, the assessment of integrity is not an absolute; rather it is expressed in relation
to its intended purpose in the environment in which it was created, managed and
maintained. Therefore, in order to have a high level of integrity, the record must have been
created, transmitted and maintained through an unbroken chain of custody from its point of
creation to the present day through legitimate successors.
Authenticity in Records Systems: Emerging Research in Digital Preservation

Records that are created in the usual and ordinary course of business are afforded the presumption of authenticity when they are created in the usual and ordinary course of business through regular, controlled procedures. This presumption, however, is not permanent; it must be maintained through a legitimate chain of custody; for once lost, it is improbable that the record can be deemed trustworthy again. The record may still have value and interest as a historical document, but its evidentiary value with regards to the activities in which it participated will be suspect in comparison to records that have been continuously maintained throughout their life in an authentic state. It falls to the records creator and custodian to defend the authenticity of records through the implementation of policies, procedures and technologies that will ensure that the record has not been altered either its transmission or its maintenance and preservation, and maintain the ability to demonstrate that the records remain free from the taint of special interests. As authenticity is an inference that is based on the processes, procedures and technologies used to create, manage and maintain the record, the stronger that the chain of custody, procedures of creation, and attestation are, the greater the presumption of authenticity. In the digital age, the implementation of technologies that create, transmit, store, retrieve and preserve records has a strong influence on a record’s authenticity.

When those technologies, policies and procedures are outside of the control of the custodians, or the records themselves are stored using third parties, it can be exceedingly difficult to produce sufficient documentation to support the authenticity of those records. In an increasingly networked world, where records can be created, managed and maintained exclusively over the internet, the authenticity of records is reliant, at least in part, upon the Internet Service Providers (ISPs). Two major projects ([9] and [14]) out of the University of British Columbia are looking to expand the understanding of ways to support the authenticity of Internet-based records. [14] is seeking to address the questions surrounding the control and authenticity of Cloud-based records in the context of an integrated legal, administrative, and value system. [9] is furthering the research of the InterPARES project by seeking to generate the theoretical and methodological framework that will support the development of integrated and consistent networks of policies, procedures, regulations, standards and legislation that can be applied to digital records that are created, stored and accessed using the Internet.

RECORDS, VALUE AND LOSS

Despite the legal and the business value of maintaining records in an essentially unaltered state, there is a general ennui amongst records custodians about preserving these records in such a way as to be able to attest to their authenticity. This attitude was evidenced in a survey of twenty-two digital repositories conducted by [3], “the majority of which felt that ensuring authenticity and integrity represented a low priority compared to increasing access and preserving content.” As [10] observed, this de-emphasis of the priority of ensuring the authenticity of records in digital repositories has a corresponding negative effect on perceived trustworthiness of the repository:
We are somewhat vague about what we expect, and we do not always understand what a given system can actually do. Our lack of clarity produces both overly optimistic (trusting) and overly pessimistic perceptions. In the world of digital information, the tools and mechanisms of ensuring integrity are complex and exotic, and our unfamiliarity with these tools leads us to distrust their efficacy. Thus we regard the digital information environment as basically lacking integrity.

The courts, however, have stated a very different point of view on the importance of maintaining the authenticity of records, and as a result, are increasingly holding records custodians to higher standards of conduct. Given the perceived lack of trust in digital environments, records custodians often required to establish the identity and integrity of the records within their records system, yet find it an immense, if not impossible challenge due to the continual process of upgrades and migrations. As a result, this inability to establish a record’s authenticity negatively affects the value of that record, as [15] states: “digital objects that lack authenticity…have limited value as evidence or as an information resource.” Digital records are one of the principle assets of any given organization and are purposefully set aside to serve as faithful witnesses to the activities that they document. In a digital age, with its rapid advances in technology, custodians are continually presented new challenges to maintaining the authenticity of records within their repositories, and thus, to maintaining the value of these assets to the organization.

THE ROLE OF RECORDS SYSTEMS IN AUTHENTICITY

A records system is defined by [7] as an "information system which captures, manages, and provides access to records through time." Records systems are not only responsible for the content of the records, but as stated by [4] are also responsible for preserving a record’s structure, business context, and association with other like records. It preserves a record’s authenticity (it is what it purports to be), reliability (accurate representation by a knowledgeable source), integrity (complete and unaltered) and usability (can be located, retrieved, presented, accessed, interpreted, and understood over time).

As most records produced today are created using some type form of a records system, the way that these systems were implemented within the organization and the controls exercised over these systems have a direct impact on the underlying authenticity of the records contained within these systems. During the course of the InterPARES Project, [4] concluded: “digital information produced by and maintained in most of the systems presently used cannot be considered trustworthy and is easily lost in a self-perpetuating and expensive cycle of obsolescence and incompatibility.”

Therein lies the conundrum of the digital age: while these records systems have the capability to create, manage and maintain authentic digital records, few of them have been implemented to do so. [12] found that: “electronic systems are still being designed to manage data rather than records.” The issues of technological obsolesce and cross platform/cross-generational incompatibility aside, procedural controls over the creation, management and maintenance of records can ensure that the records remain complete and unaltered; and the
technology exists to enforce such procedures while reinforcing them with sufficient metadata to identify the record and its context of creation and use. Records systems can, in theory, create records as part of a documented, consistent, repeatable process that imbues them with the necessary qualities to be considered authentic when moved through time and space—IF the system and procedures are properly designed and implemented AND this process of preservation begins at creation [8]. Records that are created, managed and maintained in such a defined and formulaic way are capable of serving as faithful witness to the activities in which they participated, while also holding their creators accountable. This accountability, in turn, increases the trustworthiness of those creators through providing documentation trails and a degree of transparency into the organization’s operation thereby “open[ning] the door to more sophisticated methods of control [2].”

Accountability is of particular importance at the junctures of a record’s lifecycle when they are most susceptible to alteration or technological change. As stated by [10]:

The authenticity of electronic records is threatened whenever they are transmitted across space (that is, when sent to an addressee or between systems or applications) or time (that is when they are in storage, or when the hardware or software used to store, process, communicate them is updated or replaced).

[1] refers to these instances of movement across time or through space as ‘moments of risk.’ As stated by [1], digital records are at their point of greatest risk in “moments when they are transitioning between states, e.g., when control is being passed to different systems.” [6] describes the issue thusly:

The authenticity of digital resources is threatened whenever they are exchanged between users, systems or applications, or any time technological obsolescence requires an updating or replacing of the hardware or software used to store, process, or communicate them.

The movement of records between systems—be that between two states of the same system, two disparate systems on the same network or two systems interacting with other from different organizations—entails movement through technological boundaries that require a transition of those records from one state to another. This transition, if not carefully managed and documented, can result in significant, unintended changes to those records that can adversely affect their identity or their integrity. Such changes to the underlying attributes of the records results in a negative impact to the authenticity of that record and its value as evidence.

The extent to which a record system can help or hinder the authenticity of the records that move through it depends on the way in which the underlying technology and processes have been implemented, documented and tested, as well as the extent to which the procedures controlling the creation, management and maintenance of those records has been embedded within the system. The greater the level of control that the system exhibits over the records, the greater the ability that system has to support the authenticity of the records that are on, or move through, that system. Those systems that lack sufficient operational ability to
establish the identity or demonstrate the integrity of the records compromise the authenticity and trustworthiness of those records. While the concept of what constitutes an authentic digital record has been studied extensively by the InterPARES research project over the past decade and a half, the extent to which the technological contexts of record systems affects this conceptual understanding of a records authenticity bears further study.

EMERGING RESEARCH ON TECHNOLOGICAL CONTEXTS OF RECORDS SYSTEMS

Given the many and diverse factors involved in records systems, there exists a need to establish criteria and related implementation procedures centered on technological contexts that support the authenticity of digital records contained within that system. The goal of the author's ongoing research is to establish such criteria through analyzing the traditional archival concepts of authenticity as they are implemented within records systems of several large scale digital archives through ethnographic observation, interviews of subject matter experts, analysis of implementation documentation, and the operation of the records systems. [5] views exploring this intersection between archival theory and implementation of records systems as increasingly important to the development of expanding theoretical understanding in archival science:

Exploration of the interaction between theory and practice is a crucial part of archival systems research, especially in the pursuit of requirements for digital recordkeeping within ever-evolving technological frameworks. Systems development methods may play an increasing part in research regarding this interaction as we conceptualize and build the tools needed to support recordkeeping processes in digital and network environments. The advantage of using systems development as a research approach is that it not only develops the practice, but it also serves to deepen theoretical understandings and ultimately ensure that new technologies can be made to serve archival science.

The research has been broken into four thematic phases centered on the acquisition and preservation of authentic digital records. The first phase concerns the development the organization's conceptual view of authenticity of records within its records system, with the subsequent three phases focusing on the specific 'moments of risk'—two dealing the movement of digital records across space (the transmission of records from the creators to the custodians and the process of accepting these records into the records system) and one dealing with the movement of records through time (i.e. long-term maintenance in the records system). To achieve the project's goal, four central research questions are being explored at each of the case study sites:

- What does authenticity mean to the organization accepting digital records?
- What technological contexts of the records system support the assessment of the authenticity of records as they are transferred to the records system?
- What technological contexts of the records system support the documentation of the authenticity of records ingested into the records system?
- What technological contexts of the records system support the maintenance of the authenticity of records over the long term?
The specific methods and strategies each case study employed in their implementations of the four thematic phases will be analyzed in order to identify and understand the causational relationship between the records systems and the authenticity of digital records contained within. These results will then be evaluated against the concept of authentic digital records as understood by current archival theory and archival diplomatics. The expected outcome of this research includes a new conceptual model of authenticity within records systems and an expanded understanding of the requirements to establish and maintain authentic records in records systems.

CONCLUSION

Maintaining authentic digital records through time and space requires a unified, standardized and regulated records system to create, manage, and maintain the records. Organizations are facing significant challenges in sustaining the authenticity of their records through the moments of risk due to the diverse and rapidly changing nature of technology. The goal of the author’s research project is to establish criteria and related implementation procedures for the technological contexts of records systems that support the authenticity of digital records. This analysis of the technological contexts is focusing on the entire technological environment involved in the capture and preservation of digital records; from the hardware and software architectural design, to the deployment of specific technologies, and finally, the construction of the system workflow processes. To accomplish this goal, existing records systems will be compared with traditional archival concepts and archival diplomatics to determine which aspects of the records system supports authenticity of its contents. Such analysis is expected to produce a new conceptual model of authenticity in digital records systems and an expanded understanding of the requirements to establish and maintain authentic records in digital systems that can be applied to existing and evolving records systems to better enable them to maintain authentic records through time and space.

REFERENCES

Electronic Records Capturing using eGOV-PID Standard

Sourabh Koriya¹, Jayshree Pawar², Suman Behara³, Srinu Naik⁴ and Dinesh Katre⁵

¹, ², ³, ⁴, ⁵Centre for Development of Advanced Computing, Pune–411007, Maharashtra, India
E-mail: ¹ksourabh@cdac.in, ²pjayshree@cdac.in, ³sumanb@cdac.in,
⁴srinub@cdac.in, ⁵dinesh@cdac.in

Abstract—For long term preservation of electronic records (e-records), it must be brought in preservable form called Submission Information Package (SIP) as per Open Archival Information System (OAIS). The electronic record producing agencies should implement the Digital Preservation Standard metadata eGOV-PID (e-Governance Standard for Preservation Information Documentation), duly notified by Ministry of Communications and Information Technology, Government of India. Datāntar Software helps to capture electronic records in the form of SIPS from the e-governance databases. The Preservation metadata contains basic information about the e-Record divided in several parts such as Cataloging, Enclosure, Provenance, Representation, Fixity, Digital Signature and Access Rights Information. The SIP and all files inside the SIP should be named with unique accession number which follows predefined naming convention. Then Datāntar Software transfers valid SIPS to archival system for preservation.

WHY E-RECORD CAPTURING SOFTWARE (DATĀNTAR)?

The national and state level initiatives of e-governance, e-service delivery, computerization and digitization across various domains are producing enlarging volumes of e-records which must be preserved as per the retention rules and to fulfill various legal obligations. The e-records can be quickly lost much before the assigned retention period due to obsolescence of file format, storage media, database, software and vendor lock-in as result of dependence on proprietary solutions.

The electronic records which are going for long term preservation must be final and have some fixed object form. The preservation metadata of e-record should also be captured at the time of capturing the e-record itself. As per case studies, various kinds of e-governance systems are storing their electronic records in the Databases. So to bring the e-governance records in a fixed object form along with its preservation metadata, e-Record capturing Software (Datāntar) is required.
Background Study
The study of e-District Lucknow, AMTRON Assam, and Mission Mode Projects such as CARD (Computer Aided Administration of Registration Department) Hyderabad, and MCA21 (Ministry of Corporate Affairs) has been done. These different Projects provide multiple types of services which produce the certificates or documents in different forms. In order to make all the digital data preservable, it needs to be brought in the format which is not obsolete and follows open standard along with its metadata in the form of xml. After the study of the database, it is concluded that the database register which is the basis of creation of the e-record should also be preserved along with the e-record itself as the record can be reproduced from the database register.

Different types of e-records such as birth certificate, driving license, passport, court cases, property documents, etc are being studied in order to optimize the parameters which are usually appearing and important for the e-record to be identified and those parameters are incorporated in the eGOV-PID standard.

After the study, it is clear that the huge amount of digital data is generated which needs to be preserved. Hence, if the process for capturing of e-records is simple and automated, it is more likely that the system will reduce the risks that arise when records are not captured.

WHAT IS DATĀNTAR SOFTWARE?
Datāntar is a software used for capturing e-records according to e-Governance standard for Preservation Information Documentation (eGOV-PID) for Electronic Records. It has a capability to connect with various kinds of databases. It automatically captures the e-records (database register xml) along with preservation metadata (eGOV PID) xml from the e-governance database. It produces the valid Submission Information Packages (SIPs) with unique accession number (as per the naming convention). It also automatically transfers the SIPs to the archival system for perseveration.

DIGITAL PRESERVATION STANDARD

eGOV PID
The e-Governance standard for Preservation Information Documentation (eGOV PID) defines metadata of content information (digital object) which is categorized in following parts:

- **Cataloging:** It contains the basic details of the digital object. It is the handle of the record.
- **Enclosure:** It holds the details about the supplementary documents.
- **Provenance**: It is the origin and migration details where the digital object was produced.
- **Representation**: It stores all the hardware and software details which can be helpful to render the digital object.
- **Fixity**: It is the hash value of digital object which is used to check the integrity.
- **Digital signature**: All digital signatures associated with the object can be mentioned. It contains details about signer, issuer, etc.
- **Access rights**: It defines the degree of confidentiality and user level permissions to access the object from digital repository.

![Fig. 1: Datantar Software](image)

eGOV PID of Electronic Records provides standard metadata dictionary and schema for describing an electronic record. Most of the preservation information (metadata) can be automatically captured using this schema after the final e-record is created, as most of the required information is already present in an e-government system. The implementation of this standard helps in producing a valid input i.e. Submission Information Package (SIP) for
archival and preservation purpose as per the ISO 14721: 2012 Open Archival Information Systems (OAIS) Reference Model. The elements of eGOV PID are shown in Fig. 2.

Fig. 2: eGOV PID Digital Preservation Standard

Benefits of Digital Preservation Metadata (eGOV PID)

Preservation metadata is the information about the preservable object. It is the background information that describes how and when and by whom a particular set of data or a record was created, collected or received and how it is formatted. Especially when data is computerized, without appropriate background essential information it is impossible to understand. It also tells about the migration of e-records format.

Metadata serves many important purposes, including:

- Protecting records as evidence and ensuring their accessibility and usability.
- Ensuring the authenticity, reliability and integrity of digital records.
- Enabling the efficient retrieval of digital records.
- Providing logical links between records and the context of their creation, and maintaining the links in a structured and reliable way.
- Allowing timely destruction of temporary value records when business use has ceased
- Providing information about technical and technological dependencies, to help ensure their long-term preservation and usability.
- Identifying records.
- Authenticating records.
- Capturing in a fixed way the structural and contextual information needs to be preserve the record’s meaning.
- Administering terms and conditions of access and disposal.
- Linking the attachment to the records for authenticity.

**WORKFLOW OF DATĀNTAR**

Datāntar is e-record capturing software. It captures the e-record from e-governance system as per Digital Preservation Standard (eGOV PID). The detailed workflow is explained below.

**Architecture of e-Records Capturing Software**

Datāntar Software is divided into the three parts as prepare, capture and transfer:

- **Prepare:** Prepare the e-records for capturing process by connecting to database, uploading the xml schemas for citizen services and inputting the preservation metadata.
- **Capture:** Capture the e-records (database register xml) along with preservation metadata (eGOV PID) xml by mapping the specific service schema elements to the respective database table columns. The e-records are captured with unique as valid SIPs.
- **Transfer:** Transfer the valid SIPs to the respective location of Open Information Archival System (OAIS) for preservation.

**Overall Functional Outline of Datāntar Software**

Datāntar software offers user login with password along with biometric authentication for more controlled access to the system. It is essential as the e-governance records can contain sensitive, valuable and private information pertaining to citizens and therefore it should be accessed only by authorized persons. Datāntar software allows the user to connect with heterogenous databases as observed in various e-governance implementations. The software allows you to upload the XSDs (XML schemas) pertaining to specific types of e-records such
as birth register record, death register record, domicile record, etc. It provides access to various tables and columns in the database in order to map the existing values/ information available in the database related preservation metadata with the various sections in the standardized eGOV-PID XSD. Datāntar software also allows to input the missing information externally which has to be part of the e-record. One has to do these settings only for the first time which can be saved and reused throughout the e-record capturing process thereafter.

Datāntar software also allows to generate the submission information packages(SIPs) as per the standardized eGOV-PID XSD. Submission information packages(SIPs) contains digital object and associated enclosures along with preservation metadata XML. If digital signature is available in e-governance databases, Datāntar software can capture and encode digital signature as base64 automatically. This software allows to transfer the submission information packages to the respective location of Open Archival Information System for preservation. Datāntar software is also capable of scheduling the electronic records to transfer as daily, weekly and monthly.

Datāntar software is capturing e-governance records as per eGOV-PID which contains cataloging information, enclosures information, provenance information, representation information, fixity information and access rights information.

**Cataloging Information**

Cataloging information contains the basic information of electronic records. It is the handle of the record. Almost all the information is automatically captured from e-Governance database...
but the information which is not available can be inputted from the software. Some of the basic details which are covered in cataloging information are mentioned as follows. It holds the unique identifier of the record. It allows to input State Recognized Official Language Code to be mentioned for describing the languages used in the e-record as per Officer of Registrar General of India (ORGI). All the names of persons related to the e-record can be mentioned with their role and identification document no. such as UUID, PAN, Voter ID, etc.

The duration for which e-record should be retained can also be mentioned in the cataloging information as Retention through Datāntar software.

The final e-record (e.g. the certificates issued by e-governance systems) is generated on the basis of various documents, proofs and correspondence which are enclosed with it. The enclosure information is needed for establishing the context in which the e-record was produced. The list of enclosures can be included in the eGOV PID if applicable. The domain specific metadata not covered in eGOV PID can also be linked as a separate XML as other descriptive metadata type of enclosure. The accuracy of the final e-record can be verified and validated on the basis of the enclosed documents.

**Provenance Information**

It includes the office address of e-governance systems which issued the final certificate and the device address of the system where the request was processed and final certificate was issued. As per the IT act 2000/2008, Datāntar software automatically captures the device address which contains the IP address and MAC address.
Digital migration is the transferring of data to newer system environments. This may include conversion of e-records from one file format to another (e.g., conversion of Microsoft Word to PDF or Open Document) or from one operating system to another (e.g., Windows to GNU/Linux), so the resource remains fully accessible and functional. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.

**Representation Information**

Representation Information holds the software and hardware details of the system where the final e-record was issued. Software details include the names, version and license information of software, operating system, compiler, API Library, application, tools, web browser, database, etc. which was used for creating the final e-record and the software necessary for reading it. Hardware details can be automatically captured by the software. Representation Information is helpful to render, understand and interpret the digital object content in future.

![Fig. 6: Representation Information](image)

**Fixity Information**

It includes the checksums of the final e-record. It helps to check the integrity of the record at any point of time during or after preservation.

![Fig. 7: Fixity Information](image)

**Digital Signature Information**

The digital signature metadata needs to be captured so as to establish the authenticity of the e-record at a later date. Datántar software automatically captures the digital signature from egovernance database if available. The Digital Signature information holds the details of the person who signed the e-record, the issuer of digital signature, certificate authority, public key, etc.
Access Rights Information

Access rights information identifies the access restrictions pertaining to the e-record. It covers the address and contact details of e-record holding agency. It allows to define the different permissions such as discover, display, review, extract, print, duplicate, delete on e-record for various kinds of users. Also the custom permissions can be added as "other permission" such as bulk sharing, republish, download, etc. It also provides the classification of disclosure for "public" or "private" e-records. The degree of secrecy associated with the e-record can be defined as confidentiality as per the e-Office Procedure.

Access Rights Information is useful when the user has to access electronic record from digital repository. The user can access e-records according to the permissions defined in the Access Rights block at the time of e-record capturing process.

In this way the electronic records are captured in the form of database register xml because xml is a simple, extensible, interoperable and open standard format. In the future the records' data can be regenerated from the xml. The preservation metadata file is generated as explained above. The enclosures and digital signature certificates are extracted if available. The valid SIPs are generated by Datāntar software. As shown in Fig. 9, Datāntar software transfers the SIPs to the OAIS (e-Record Digitalaya) for further process of preservation.
Fig. 9: Transferring e-records from Capturing System to e-Records Digitalaya (OAIS) being Developed by C-DAC Pune

CONCLUSION

The Datāntar software follows the Indian Information Technology Act 2000/2008 to ensure its legal admissibility. All the processes in the software are automated as a batch process which is helpful in capturing and transferring millions of electronic records in valid preservable form. Hence manual efforts and time consumption is reduced while accuracy and performance is increased:

- The Datāntar software captures the electronic records in the form of database register xml from the e-governance databases.
- It captures the preservation metadata xml along with the main preservation object in order to create valid Submission Information Package (SIP).
- The SIPs are created with unique accession numbers following the predefined naming conventions. The supplementary documents are captured in SIP.
- If the digital signature is available in the database then it can be extracted in the form of bas64.
- The software generates the fixity of main preservation object in order to retain the integrity.
- The valid SIPs captured in Datāntar software can be transferred to Open Archival Information system (OAIS) for further preservation.
In the current scenario Datāntar software is a prototype for legacy e-governance records. In the future it can be made generic and used for various citizen services provided by e-governance system.

**ACKNOWLEDGEMENT**

The authors of this paper acknowledge the support received from the project engineers at C-DAC Pune namely Sujata Aher, Madhuri Dhone, Priya More, Piyush Patel, and Hemant Anjana. C-DAC acknowledges the funding support and encouragement received from Department of Electronics and Information Technology (DeitY), Government of India for the Centre of Excellence for Digital Preservation project, which is entrusted with Human-Centred Design & Computing Group at CDAC, Pune.

**REFERENCES**

[4] Best Practices & Guidelines for Production of Preservable e-Records (PRoPeR) of eGov-PID.
Long-term Preservation, Management and Utility of EHR

Gaur Sunder
Centre for Development of Advanced Computing (C-DAC), Pune, India
E-mail: gaurs@cdac.in

Abstract—Electronic Health Record (EHR) of a person is collection of records generated during clinical encounters, preserved electronically. Since EHR captures data about individual, diseases, treatments and their efficacy; it is a treasure trove of information when used for historic analysis as well as statistical and correlational studies. To be effective, EHR needs to be preserved during the life time of individual and also retained for later use to understand new details as science and technology opens more opportunities to understand the knowledge contained therein. This paper introduces some ongoing work and presents ideas for preservation of EHR, management and utility of preserved records.

INTRODUCTION

Electronic Health Record is a repository of information regarding the health of a subject of care in computer processable form [1]. Broadly, it is longitudinal electronic record of patient health information, generated by one or more encounters [2]. As is evident, EHR is a lifetime worth of healthcare records not specific to any provider or location but of an individual (subject of care). Medical data captured in EHR is of great importance in understanding the medical history of the individual in order to treat illness later. It also has significance from the point of medical research. Hence, there are several incentives in that the medical data constituting the EHR must be preserved.

PRESERVING MEDICAL DATA

The DOHaD study [3] has established the importance of biomedical records to study the historical context in order to understand various associations of disease and early medical and related history. Medical practice has evolved several processes that emphasize the need to preserve the medical history of a patient. Generally, better the historical records, more specific the treatment can be.

Need

There are already various studies being undertaken to study the effect of EHR on reducing cost of medical care [4]. The reduction in cost of medical care itself is only one of the goals of preserved medical record. Its biggest contribution is in epidemiology and medical research.
The promise of available medical data captured across a region or entire country will speed-up the medical processes and significantly reduce the unwarranted cost associated with lack of such data when needed. Since preserving the medical data is unarguably important, the need to preserve its electronic form is unquestionable.

Several countries are running national programs to interconnect various sources of medical records together with the view to provide medical data wherever the patient is. Around eleven countries have a national program at various stage [5] of implementation. India is also catching up by placing important pieces of such a system in place. Recently notified EHR Standards for India [6] is an important piece in achieving an interoperable medical data system leading to common accessible EHR. The requirement to enable preservation of medical records is already mentioned under the Guidelines of Software Standards.

<table>
<thead>
<tr>
<th>Country</th>
<th>National Healthcare IT Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>HealthConnect</td>
</tr>
<tr>
<td>Austria</td>
<td>ELGA</td>
</tr>
<tr>
<td>Canada</td>
<td>EHRS Blueprint</td>
</tr>
<tr>
<td>Denmark</td>
<td>MedCom</td>
</tr>
<tr>
<td>England</td>
<td>Spine</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>eHR Infrastructure</td>
</tr>
<tr>
<td>India</td>
<td>EHR Standards for India [6]</td>
</tr>
<tr>
<td>Netherlands</td>
<td>AORTA</td>
</tr>
<tr>
<td>Singapore</td>
<td>EMRX</td>
</tr>
<tr>
<td>Sweden</td>
<td>National Patient Summary (NPO)</td>
</tr>
<tr>
<td>Taiwan</td>
<td>Health Information Network (HIN)</td>
</tr>
<tr>
<td>USA</td>
<td>EHR Meaningful Use</td>
</tr>
</tbody>
</table>

**Coverage**

Like all data, medical data (in other words clinical information) is in two forms; that recorded on paper and that born digital. Obviously, the approach to preserve the two have to be different. However, in current time preserving the paper form may seem like waste of space (especially) and money.

Preserving the born digital data is also a challenge in itself. With upwardly mobile population, the chances of medical records being spread across geographies has increased significantly. Coupled with ubiquity of computing devices and interconnected networks, the aggregated EHR from all sources is worth preserving. However, how to aggregate the data from geographically spread locations and in a format that is interoperable with various systems and format is a monstrous challenge.

**Barriers/ Considerations**

There are several barriers or considerations in preservation of EHR. Some of these barriers are technical and some are due to practices and perceptions. One of the work groups of HIMSS has suggested several mechanisms that will increase adoption of the EHR itself [7]
and its benefits in medical practice. In order for medical data to be available, first of all the medical professionals have to move to using electronic modes of recording of data. Reluctance of the medical professional and lack of familiarity with technical tools are the biggest impediments in implementation of electronic data collection [8]. Some of the barriers in preservation of medical data through EHR are:

**Existing Non-electronic Data**

The medical practice in the country is largely paper based. Right from registration of patient to final discharge (in IPD) or medical note (in OPD) is carried out using paper based forms and notes. For an effective birth-to-current EHR, bringing this historical medical data to electronic format will be a major activity running the cost up.

**Modes of Preservation**

One of the difficult questions to answer is how the data will be preserved and accessed. The medical data is being created locally however, the purpose of preservation is to make it retrievable from anywhere. Also, if the data is preserved in a format that is not understandable or usable by others (people or computer program), it will defeat the purpose of preservation. Medical data is known to exists in all possible format; image, text, waveform, binary, etc. Also, standards wherever existing and adopted also evolve.

**Size and Duration of Preservation**

With increasing life expectancy of people in country, for people being born today and all records going digital, it will still be roughly sixty-four / sixty-seven years’ [9] worth of medical records at the time of death. Obviously, the records needs to be available for quick access for another two decade for analysis and medico-legal requirements. With this it clearly will be about a century before the first medical record (or entire EHR) of a person could be archived to slower/seldom accessed medium. For a population of India, average size of medical records, and accounting for rate of birth today, the data-size could be in zeta-bytes already and increasing.

**Access, Security and Privacy**

The preserved medical data is protected and private in nature. The ownership of medical record is still unsettled legal question. The data needs to be accessible to medical professionals, researchers, patient and her/his guardian. However, each access must be controlled, secured and also privacy of data has to be ensured. These are competing requirements; how do you secure the data yet make it accessible wherever required for legitimate purposes? Privacy cannot be defined in absolute terms as perception varies from person to person. The constant threat of malicious access poses a whole set of different challenges.
Hardware and Software Obsolescence

The underlying hardware and software technologies keep changing. Despite best efforts, the data preserved in one medium may become unusable tomorrow. The understanding of format in which data is preserved may be lost in time. With electronic formats and medium, the loss of usability and access can be quick and silent one.

Other Technological Aspects

There are few other issues of how to interconnect the healthcare systems, ensuring fail safety, performance, scalability and support for multiple standards.

Strategies

With advancement in technology and our understanding of the medical ecosystems, few strategies have emerged that attempt to provide workable approach to the problems posed above. In reality the above mentioned problems can never be truly ‘solved’ as the real world is dynamic and rightly so. Progress in technology, format, and methods is an added cause of these barriers and hence, cannot be written off. However, some mitigating strategies could be used to lessen the impact of change.

For the inclusion of non-digital data, there are several possible ‘compromises’ that can be considered:

- **Start from a particular date:** Select a date for preservation and start collecting all data from that date onwards. Medical data available before this date, if available in usable digital format, could be assimilated slowly in phases. This way, some historical context may not be covered but system will be effective from then on. This will also save cost of conversion.

- **Convert all non-digital data to digital:** This could be a massive task but it must be weighed along with benefits it will bring to the effort of preservation. The conversion can be taken up in phases and cover records up to a period each time.

- **Need based conversion:** Another way to do it could be to make the conversion a need-based one rather than explicit requirement. This way, if a past record is required, it can be converted then.

Data existing in digital formator not, still have to be assimilated for preservation in a fixed or evolving set of formats. Many national programs have adopted or defined a particular format in which all medical data is to be created for preservation. India has chosen to allow a set of standards, one for each specific purpose [6]. It appears prudent to select a nationally (preferably internationally) accepted standard for creation or conversion of data. This will ensure that no proprietary or esoteric format is selected that will defeat the purpose of preservation. However, a standard itself evolves over time and newer version/edition of standards are released from time to time. This means that data created using standard selected today may become unusable when newer version/edition is selected later. There are three possible strategies for this scenario:
- **Convert all preserved data to new standard**: This approach will ensure that all preserved data is always usable. However, this is easier said than done. The size and variety of data will be enormous and attempting to convert all data will be time consuming and costly. Also, if newer edition/version of altogether new standard will be selected, the data itself may not be easily convertible requiring extra effort or possible loss. This strategy, on one hand will ensure data usability but on other will be expensive both time and cost wise.

- **Ensure preserved data is always usable**: In this approach, data must be preserved along with its standardized reader, writer, and visualization technologies. These must be available as a national service or as reusable (and preferably open source) software components that can be freely integrated in existing or new application. Since ability to read/write and visualize the preserved data will never be lost, the data will continue to be useful.

- **A mixed strategy**: A mix of both the above mentioned methods could be used. The data should be preserved along with ability to read/write and visualize it as a national service or component. Along with that a national data translation service be made available that converts the data to newer version/edition. This way the cost can be centralized while local and user side applications can progress to newer version/edition. Also, after a long time and several revisions in format or standard, a pipeline of previously created translational services can be used to access preserved data in new format/standard.

The enormity of data size and their storage is a significant issue. There are few simple and few complex working solutions for the data size problem:

- **Local storage, common registry**: Few countries have adopted mechanism of leaving preserved data at local end or where it is created. They have implemented central registry or index service that allows searching and locating the data and then actual transfer taken place from place of actual storage. This allows for natural distribution of data storage. However, this strategy suffer major possibility of an unrecoverable outage leading to total loss of data. Also, if a health service provider winds down its operations, there is no clarity of what will happen to preserved data.

- **Central storage**: The strategy is to centralize all data storage and ensure preservation at all times. This is the safest option as these central storage points can preserve, backup, archive, and process data independently. This approach will make the central storage system very active one and necessitate creation of large data centers. Central storage may slow down the end-point systems. With falling data storage media cost and increasing reliability it is a workable approach.

- **Scalable distributed system**: A mix of above approach is to let end-point data systems store and manage its own data but keep a copy in chosen standard format in a system that distributes it across low-cost storage nodes that are fail-safe and redundant. This will ensure that data exists independent of point where it was created.
and is accessible from anywhere. The access can be load-balanced across various nodes to ensure performance. Moreover, in this approach application vendors are free to use and design the storage system that best suits the need of the user and yet ensure that data in proper format is preserved. It gives benefit of central storage and yet is more scalable approach. India has developed a prototype framework for distributed system that can aggregate EHRs from diverse sources [10].

- Hierarchical or multi-level storage techniques can significantly reduce data storage cost. As a rule, faster the storage devices, more expensive it is. Hence, spreading the preserved data across storage mediums of varying speed could lower the cost. Also, data not being used actively could be archived temporarily or permanently to offline storage mediums.

There is no silver bullet for security concerns. Thankfully, a set of technologies and mechanisms are available that can mitigate the threats considerably and yet ensure safe and secure access. Incidentally, a Role Based Access Control may not be sufficient, by its own, to implement over regional or nationally preserved medical data for several reasons. To use one example; a person in role doctor cannot be allowed access to all medical records. Also, it may not be possible to define access rights of each user to each record every time a user or a new data is created. Some ideas for security and privacy of preserved medical data are:

- One of the safe and viable approach is to use fine-grained policy based access control. EHR Standards for India [6] has listed some standards for use in access control that can be used for access of preserved data as well.

- Use of PKI and Digital Certificates to identify the user and her/his role reliably will providesafety from impersonation.

- All communication should preferably be conducted over secured channel, e.g. encrypted using SSL/ TLS as the case may be.

- All data in storage could always be kept in encrypted form using some key specific to machine or user.

- Although these measures coupled with access control will thwart most malicious attempts, still need to audit and log each access is required to bring in legal action, if required.

All hardware and software ultimately become obsolete and are replaced by newer (and probably better or cheaper) technologies. Several approaches have been in vogue to lessen the effect of the change. Some ideas are:

- The software when developed using open technologies and standard have some protection from obsolescence.

- Periodic software upgrade and migration, although are time consuming, but safeguard against data loss.
• For hardware, one good strategy could be to use various levels of hardware specification as per need of the operation. As opposed to standard specification causing whole datacenter to become obsolete at same time, different specifications and staggered purchase cycle could reduce the cost and turn it into cyclic mode.

• Distributed system model allow scaling dynamically and hardware can be renewed periodically without affecting entire system or usage.

For all other issues, a good, robust design coupled with periodic review will ensure that system is always functional and usable.

MANAGEMENT OF PRESERVED MEDICAL DATA

Control and Governance of the preserved data is an important aspect of preservation. Without clear approach to ownership of data, policy or access, modes and mechanisms of its utilization, entire preservation would be meaningless.

Retention rules/ law for medical records vary and there is no clear guideline for preservation of medical data. The table below lists some of the guideline/ rule in this regard in India:

<table>
<thead>
<tr>
<th>Authority</th>
<th>Rule/ Guideline</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Council of India</td>
<td>Physician should maintain medical records for period of 3 years from the date of commencement of treatment</td>
<td>Regulation 1.3.1 of Medical Council of India (Professional conduct, Etiquette and Ethics) Regulations, 2002</td>
</tr>
<tr>
<td>Directorate General of Health Services, India</td>
<td>Records to be maintained for period of 5 years for out-patient, 10 years for in-patient, and medico-legal register to be maintained for 10 years</td>
<td>Letter No. 10-3/68-MH dated 31-Aug-1968 Also, Chapter 12-Medical Record Services of Hospital Manual by the Directorate General of Health Services, MOHFW, GOI, 2002</td>
</tr>
<tr>
<td>State Directorate of Health Services, Punjab</td>
<td>Medico-legal records to be maintained for period of 12 years</td>
<td>Punjab Medical Manual, 1934</td>
</tr>
</tbody>
</table>

Similarly, several states may have their own local laws related to maintaining of medical records. There are several other laws which call or rely upon medical records for various purposes. Most important of them all is the Indian Evidence Act, 1961. Using this law, the records can be called/summoned/subpoena by courts or legal authorities. Once, listed as evidence in any on-going legal dispute, the record is to be preserved until final decision is reached or legal permission to destroy the record has been obtained. Destruction of such evidence is punishable under Section 201 of Indian Penal Code (IPC). The IT Act 2000 of India has made electronic record admissible as evidence following some conditions. There are views [11] that medical records pertaining to a minor must be preserved for at least 25 years to cover for time to attain age of maturity and subsequent civil limitation of 3 years in case the minor would pursue legal action for any reason.

Even leaving the medico-legal cases, there is no single understanding of how long the records should be preserved. In such a case, it is all the more important to preserve the records for a very long time. As indicated above (see Barrier/Consideration section for
discussion of data sizing), preserving medical records for over a century is not a very radical requirement. Medico-legal and medical research requirements alone make the preservation of records in perpetuity a valid requirement.

However, in order to enforce such a rule for all record could be a tall order, if not impractical. Few strategies that can be considered for deciding duration of preservation could be:

- If not technically impossible or impractically costly, preserve all medical records in perpetuity.
- Preserve all records for a specific time period (a century as suggested above) or at least 25 years after the death of the individual.
- Categorize medical records based on their importance or requirement for preservation and assign a specific time period for each category. For example, records pertaining to case or treatment of common-cold could be preserved for very short time, while records of major invasive procedure should be preserved for a very long time.
- In any case, homogenize the time requirement to preserve medical records across entire range of laws and rules. For medico-legal case, records can be marked as non-destructible till final decision of the case or order/permission of court to destroy them is obtained.

Another issue in management of preservation is Ownership of records. There is no clear rule/guideline for ownership or medical records. The Medical Council of India rules leave the ownership issue unresolved but indicate that records must be provided to patient upon request within specific time period. The widely accepted theory about medical records is that the caregiver or hospital owns the records, but the patient has the right to the information included in the record [10]. In recent time, the EHR Standards Committee of India has recommended a radical shift from this standpoint, stating that Patient is the owner of the medical record and healthcare service provider is custodian of the record on behalf of the patient [6].

It can be argued that State has legitimate interest in preserving medical records of individuals as long as it is secured and reasonable privacy is maintained. The legitimate interest could be for purpose of providing better service to individual in particular and society in general. Also, furtherance of science (a Constitutional Directive Principle) by way of medical research, better drugs, response to epidemics and out-breaks could be argued to be legitimate use allowing for preservation of medical records of all. In such a case, State could be legitimate custodian of preserved medical records with appropriate legislative restrictions upon their legitimate use.

Under Indian Constitution, Health is a subject in the State List given in Seventh Schedule. This allows each State in Indian Union to be able to create laws and rules governing universal healthcare within its territories. This could pose a problem in
implementation of a centrally administered EHR preservation scheme. The National Rural Health Mission (NRHM) program of the Union Government could be a role model on which EHR preservation program could be started. A program under Union grant requiring each state to follow a certain set of rules to be eligible for funds could be used to administer a common program. Once the basic contours of program such as; following a fixed set of standards for creating, storing, sharing, and preserving medical records are in place, a national program can be started and administered.

MAXIMIZING UTILITY OF PRESERVED MEDICAL DATA

The most visible purpose of preservation of medical data is to support empirical and correlational research to understand the variables of healthcare and wellness of individual and group. Preservation of healthcare records is based on the long held premise of medicine that past medical condition has a bearing on the current and future health of a person or society. This premise was confirmed by the DOHaD study [3]. Nonetheless, from an individual’s perspective, the preservation could have following utilities:

- Access to medical records when and where needed: With a mobile population being a norm in the Economically connected society, availability of the proper medical records anywhere and anytime is a boon. This could be life-saving in emergency conditions as medical service provider does not have to second guess or lose time in running tests to confirm or deny a medical condition before proceeding with required treatment.

- Reduced time and cost in medical diagnostics: The medical diagnostics processes are currently unavoidable cost because, the needed records are not always available at the point of treatment. This forces the medical practitioner to order the tests and observations again. This causes time delay and unnecessary cost escalation in medical treatment. With all past medical records being available quickly, some delays and costs can surely be avoided.

- Evidence-based medical care: A natural side-effect of active preservation of medical records is that the records build over lifetime and diagnostics and medical care can become evidence based and personalized. This can help practitioner arrive at well-informed decisions on line of treatment than experiment with medicine and their efficacy over an individual.

- Better insurance handling: One of the biggest issue of Health Insurance is incomplete or unavailable medical history. This is also one of the reasons for prolonged litigation in Insurance services. With medical records available, Insurance Companies will be able to better assess individual’s need for coverage and provide tailor-made scheme. Although it may cause short term increase in cost of insurance in particular cases, it will lead to better cost of coverage spread across population thereby bringing cost down for everybody.
As a society and healthcare provider, there are several utilities to preservation of EHR:

- **Availability of medical history**: For effective treatment and to avoid medical mishap, availability of medical history is very important. Once preserved medical records are available, practitioner can make informed diagnosis and choose most effective line of treatment. This can also reduce exposure of practitioner to medico-legal prosecution.

- **Patterns and correlations**: With medical records of people in a certain region or suffering with similar afflictions being available, it is possible to study them empirically and in correlation to arrive at better line of treatments. This can assist in better prediction of medical epidemic and out-break as pointers, and alerts can be raised based on available data.

- **Drug efficacy**: One of the challenges in drug development and improvement is lack of scientifically collected data to study short, mid and long term effects of drugs on individuals. With different genotype-phenotype interactions with drugs the task to study drug efficacy is challenging. With such interaction data being available through preserved medical data, better drugs can be designed and used.

- **Management information**: For any public healthcare system, it is important to cover specific intelligence in order to generate management information for policy and execution effectiveness. With specific information being made available on the fly from the preserved data in real-time, the effectiveness of policy and program could be observed.

Above utilities being only indicative of information and possibilities that preservation of EHR will offer, one can automatically come to theorize what may become possible due to availability of usable and reliable medical information. This could be gold-mine of data-warehousing and studies/ findings of future healthcare.

**CONCLUSION**

It is beyond doubt that preservation of medical records as embodied in EHR is a complex yet required activity. The benefits far out-weigh the challenges, which themselves are not unsurmountable. To be effective, EHR needs to be preserved during the life time of individual and also retained for later use to understand new details as science and technology opens more opportunities to understand the knowledge contained therein. Several nations have already begun the process and many have achieved effective operationalization. India has begun to understand the possibilities that this exercise can offer. India has moved on various fronts from defining common set of standards to developing prototype for aggregation and storage of EHRs.

As Information and Communication Technologies make more inroads into the healthcare domain in the country and worldwide, preservation of the medical data will become an imperative. As is the saying; a work well begun is half done. It is in our favor to understand the need and the task at hand, and commit ourselves to it.
ACKNOWLEDGMENT

The author acknowledges the support of Dr. Pradeep Sinha, and his team members, especially Manisha Mantri, in guidance and assistance in writing this article. He is also thankful to Dr. Dinesh Katre for providing a platform to put forth these ideas towards a national healthcare data repository.

REFERENCES


Digital Preservation Infrastructure
Towards IT Support for Sustained Access and Use of Digital Content

Natasa Milic-Frayling
Microsoft Research, Cambridge, United Kingdom

Abstract—Information in the digital form is amenable to sharing, aggregation, and analysis at unprecedented speeds and scale. That creates new opportunities for innovation, increasingly requiring development of new computing technologies and systems. However, the current approach to the adoption and management of information technologies raises concerns about our digital legacy: computing systems and software become obsolete and our digital assets become inaccessible. Digital content can be used only through computation, i.e., by instantiating the software application to present the content in the form that can be consumed by users, most often by rendering on the screen or providing audio and haptic output. The content is accessible only while the application is running. Prior efforts to secure access to digital content focused primarily on the persistence of data and program files and less on ensuring that the software can run in the contemporary environment. We wish to emphasize the importance of computation. We argue that the task of preserving digital assets is the task of enabling computation by which digital content can be used in the contemporary computing environment.

The problem of instantiating digital content is difficult because of the inherent dependence on a sophisticated computing infrastructure that includes stacks of supporting software and a variety of enabling hardware. For the sake of the discussion, we look at a specific case of the digital content produced by office productivity tools such as Open Office, Microsoft Office suite, and similar. Assuming that the adequate hardware is available, we discuss three approaches for ensuring that the content is accessible to users. These approaches to various degrees rely upon the original software application and data files and thus raise the issue of authenticity. We expect that all three will be required to cover a range of scenarios and requirements. At the same time, we are aware that each of them will require on-going investment in IT support. Thus, it is critical to consider ways to lower the cost of IT support, software development, and data processing.

In order to explore this issue, we consider three aspects related to the ICT industry: the value chain in the digital information production, the need for computing architectures, such as the cloud platform, to achieve the economy of scale and viable models for hosting digital content, and principles of software engineering to make computation sustainable at lower costs. A pre-requisite for long life of digital is making the IT support of computation affordable and developing methods to integrate the legacy with the contemporary computing ecosystem.
Building a Federated Infrastructure for Preservation of and Access to Research Data in the Netherlands: The Front Office-Back Office Model

Peter Doorn\textsuperscript{1}, Ingrid Dillo\textsuperscript{2} and Paula Witkamp\textsuperscript{3}

\textsuperscript{1,2,3}Data Archiving and Networked Services (DANS), The Hague, The Netherlands
E-mail: \textsuperscript{1}peter.doorn@dans.knaw.nl, \textsuperscript{2}ingrid.dillo@dans.knaw.nl, \textsuperscript{3}paula.witkamp@dans.knaw.nl

Abstract—A federated data infrastructure is emerging in The Netherlands on the basis of the collaborative model proposed in the Riding the Wave report as a framework for the scholarly information system of the future. This federated model is elaborated as a layered front office – back office model, in which university libraries, national data services and basic technical e-infrastructure organizations work together. The responsibilities and functions performed by the various stakeholders involved in the federated infrastructure are clearly complementary. The costs and benefits are distributed efficiently over the stakeholder. The introduction of the model is timely, because many research organizations are currently developing data management policies. Therefore the federated infrastructure is attractive for all parties involved.

Keywords: Research Infrastructure, Data Archiving, Data Access, Data Management

The Importance of Storing and Sharing Research Data

Sharing research data is important. Data availability enables replication of research and thus enhances scientific transparency. Reusability of data also contributes to scientific progress and the advancement of knowledge [1].

Research funders acknowledge the importance of data sharing and require a high degree of open access. The Dutch funding organization for research (NWO) and the Dutch Academy (KNAW) argue that research data obtained with public funding should be publicly accessible as much as possible[2, 3]. The European Commission wants to increase the impact of publicly funded research by making its results available to open access. In the context of the Horizon2020 program the EC is examining ways in which open access to research data can best be realized [4].
There are all sorts of snags to sharing data. The data should be searchable and accessible to others than those who collected them, and they must be stored in a consistent and sustainable way. Sometimes the privacy of individuals needs protecting. Additionally, conflicts of interest may arise when researchers collaborate with companies that want to exploit the data obtained. Also, many researchers do not want others to publish about the data collected by them before they themselves have had the opportunity to do so [5].

Towards a Federated Data Infrastructure

Good, timely data management and secure data storage, both during and after completion of a research project, are essential prerequisites for sharing those data. It is therefore very important for universities and research institutions to formulate a clear data management policy for their organization. In order to implement this data policy good support and an adequate technical infrastructure are indispensable.

In the influential report Riding the Wave, which has been embraced enthusiastically by European Commission Vice-President and European Commissioner for the Digital Agenda, Neelie Kroes, such a “Collaborative Data Infrastructure” is touted as a framework for the future [6] (Fig. 1). The figure suggests, in the broadest possible terms, how different actors, data types and services should interrelate in a global e-infrastructure for science. Data generators and users gather, capture, transfer and process data – often across the globe, in virtual research environments. They draw upon support services in their specific scientific communities - tools to help them find remote data, work with them, annotate or interpret them. The support services, specific to each scientific domain and provided by institutes or companies, draw on a broad set of common data services that cut across the global system; these include systems to store and identify data, authenticate them, execute tasks, and mine the data for unexpected insights. At every layer in the system, there are appropriate provisions to curate data - and to ensure their trustworthiness.

![Diagram](image)

Fig. 1: The Collaborative Data Infrastructure: A Framework for the Future

Source: Riding the Wave [6]
In the Netherlands, a federated data infrastructure is emerging on the basis of the reference model outlined above, with three layers of roles and responsibilities for the various stakeholders[7] (Fig. 2).

The foundation is a basic technical infrastructure, which facilitates data storage and back-up.

Above that is a layer of back-office data services, providing facilities and support for long-term archiving and accessibility.

The highest level includes the front-office services. They provide for the first-line contacts, supporting, advising and training researchers and students in responsible data management. The front office can rely on the expertise of the back office.

**STAKEHOLDERS IN THE FEDERATED DATA INFRASTRUCTURE**

In this federated and layered data infrastructure the various stakeholders have specific responsibilities stemming from their respective positions and competencies.

The basic technical infrastructure is provided by data centres, an area where e-infrastructure providers like SURFsara and Target have a coordinating role on a national or regional level.

---

[7] SURFsara is the Dutch national ICT infrastructure organization for research and higher education, providing services in the areas of computing, data storage, visualization, networking, cloud and e-Science; Target is one of the largest ongoing public-private projects in the Netherlands in the area of large-scale data management and information systems.
The back-office functions are carried out by organizations with a national role to play in the field of long-term accessibility of data in trusted digital repositories, such as Data Archiving and Networked Services (DANS) and the data center of the three collaborating technical universities (3TU.Datacentrum), collaborating in Research Data Netherlands. Together they have expertise on data from the humanities, sciences and social sciences.

The front offices are located at universities (libraries, local data centres), research/knowledge institutes, institutes of applied science, national and international research infrastructures (ESFRI and National Roadmap). All these organizations have in common that they are primarily responsible for the quality assurance of the data produced and processed by them or for them.

**SERVICES, ROLES AND RESPONSIBILITIES IN THE FO-BO MODEL**

In the federated data infrastructure roles can be divided according to the Front Office – Back Office model (FO-BO model). The services provided in this model are all related to data management and storage. They fall roughly into three groups:

- Awareness raising and information provision.
- Training (focusing on data librarians/experts and researchers).
- Data curation, management and storage during and after research projects.

The focus of the *front office* is on supporting its own research organization. In the area of data management the front office takes care of awareness raising, providing information and training its researchers.

In addition, the front office features so-called Virtual Research Environments or Data Labs, which offer research tools and secure temporary storage facilities (Sharepoint, Dataverse, etc.) for the organization’s researchers. In consultation with the back office, the front office also facilitates the transfer of data to a trusted back-office digital repository after a research has been completed. Facilities that are shared by several universities, including Dataverse, can be hosted and supported by the back office.

Data acquisition within its research community is another front-office duty. In all its tasks the front office will, if necessary, maintain contact with the back office.

The focus of the *backoffice* is on the expertise surrounding data governance and data stewardship\(^2\), including long-term storage and accessibility of the research data.

---

\(^2\)Data stewardship is the management and oversight of an organization’s data assets to provide business users with high quality data that are easily accessible in a consistent manner. While data governance generally focuses on high-level policies and procedures, data stewardship focuses on tactical coordination and implementation. Data stewards can also be responsible for carrying out data usage and security policies as determined through enterprise data governance initiatives [for more information, see http://searchdatamanagement.techtarget.com/definition/data-stewardship].
The back office is responsible for training the data librarians/experts employed by the front office and providing the front office with substantive support via dedicated contacts. Back-office employees may act as experts and contribute to the front-office training activities for researchers.

Where needed, the backoffice also provides consultancy services to the front office. In other words, the backoffice acts as a centre of expertise and innovation.

Furthermore, the back office ensures the sustainable and secure storage and retrieval of data upon completion of the research project. For this purpose the data are transferred through the front office to the back office.

In the acquisition, support, consultancy and training services, the duties of front and back offices may overlap. Coordination and definition of responsibilities will therefore be necessary. Specific responsibilities will vary from organization to organization, but it is important to have clear agreements on e.g. data acquisition\(^3\), and the use of data management plans.

**COSTS AND BENEFITS: THE FO-BO BUSINESS MODEL**

The FO-BO model offers benefits for all stakeholders because it provides for an optimum division of labour based on the respective expert competencies of those stakeholders and their various roles in the data infrastructure. The bullets below summarize these benefits for both back and front office and the research community:

**Benefits for Researchers**
- Researchers benefit from increased data curation knowledge at the Front Office (FO).
- Researchers have better access to sustainable storage of their data.
- Researchers gain time by engaging in data management in an early stage.

**Benefits for Front Office (FO)**
- FO benefits from data storage facilities at Back Office (BO).
- FO benefits from BO’s knowledge on data curation.
- FO supports institutional responsibility in data management.

**Benefits for Back Office (BO)**
- BO benefits from well-trained FO contact persons for researchers.
- BO benefits from direct contact with their target audience through FO.
- BO acquires more data that are better prepared for archiving and reuse.

\(^{3}\)The Dutch Funding organization for research NWO uses data contracts for this.
As a trusted digital repository DANS performs back-office functions, supplemented by a number of national and international front-office duties. In its business model, DANS aims at institution-wide, formal agreements with the front offices at the universities. Such framework agreements set out the details of data management and data storage services based on the FO-BO model outlined above. They may involve the distribution of responsibilities, but also e.g. the establishment and maintenance of the required technical infrastructure.

DANS will charge the basic data storage costs including back-up. Storage is currently supplied by Vancis, a subsidiary of SURFsara. There is a Service Level Agreement with Vancis in place, which guarantees a very high level of security and data availability.

In return for a one-off payment of these costs for five years in advance, DANSensures conservation of the data “indefinitely”. This will enable the long-term safekeeping of data from projects with temporary funding.

This arrangement assumes that data and metadata are supplied in the agreed format. Where this is not the case, DANS will charge costs for processing data and organizing documentation.

REFERENCES


SCIDIP-ES: A Sustainable Data Preservation Infrastructure to Support OAIS Conformant Archives

Shirley Crompton¹, David Giaretta¹,², Brian Matthews¹, Holger Brocks³, Felix Engel¹, Arif Shoan* and Fulvio Marelli⁵
¹Scientific Information Group, Science and Technology Facilities Council, UK
²Alliance for Permanent Access, UK
³InConTec GmbH, Germany
⁴FTK, Germany
⁵European Space Agency, Italy
E-mail: ¹shirley.crompton@stfc.ac.uk, ²david.giaretta@stfc.ac.uk/director@alliancepermanentaccess.org, ³brian.matthews@stfc.ac.uk/holger.brocks@incontec.de, ⁴arif.shoan@stfc.ac.uk, ⁵felix.marelli@ esa.int

Abstract—The SCIDIP-ES project provides an e-infrastructure to support long-term preservation and use of the knowledge encoded in scientific data. The infrastructure offers a set of generic, sustainable services and toolkits based on the CASPAR prototypes to support efficient preservation planning and management along with usability and access needs. The SCIDIP-ES services are specifically designed to augment existing repositories as well as to facilitate the development of new ones, so that they can be properly OAIS conformant as described in section 1.4 of the OAIS Reference Model. In particular, the infrastructure enables repositories to support the OAIS Information Model and to fulfil several of the mandatory responsibilities which are required for conformance yet which are rarely discussed.

Keywords: Digital Preservation, OAIS, e-infrastructure, Earth Science, Services

INTRODUCTION

The EU regards long term preservation of scientific data a public interest in view of the societal and economic gains from improved access to and continuous exploitation of reliable research output [1]. SCience Data Infrastructure for Preservation–Earth Science (SCIDIP-ES) [2] supports this vision by offering generic, sustainable services and toolkits for efficient preservation planning and management along with usability and access needs. By helping to build a broad Earth Science (ES) user community of significant mass, sharing and using each other’s data over the long term, implementing the software in a simple and robust fashion, and linking to the Alliance for Permanent Access (APA) [3] beyond the project, we aim to guarantee sustainability of these services. Besides being driven by requirements from the Reference Model For An Open Archival Information System (OAIS) [4], the services offered are those which have been shown by CASPAR [5] as being effective for preservation
and by PARSE.Insight [6] as being widely recognized as needed by diverse communities. To ensure the SCIDIP-ES infrastructure will indeed be used (i.e., we wish to avoid a “build it and they will come” approach), the detailed requirement specifications and customisations of these services and toolkits have been defined using various exemplars from ES and other domains.

Science data infrastructures, e.g., within the Earth Science (ES) domain, are inherently heterogeneous and fragmented due to the proliferation of completely different data capture instruments and data formats used. In addition, most of these existing science data infrastructures are not OAIS conformant, i.e. not preservation-aware. Shaon et al. [7], for instance, have reported on the main preservation challenges and barriers that science data infrastructures need to overcome. SCIDIP-ES can help data archives to become preservation-aware through adopting its OAIS-conformant services and toolkits. In this paper, we describe the design and functional specification of the SCIDIP-ES infrastructure and, in specific, how we built-in support for the OAIS standards. We begin with a brief characterization of the state of OAIS conformance within digital preservation (Section 2), followed by a review of the mandatory responsibilities required for conformance (Section 3). Section 4 describes the three high level use cases developed using results from analyses of ES exemplars and responses to a requirement survey [8] conducted by the project. The analyses were aligned with the OAIS Reference Model, focusing on identifying the roles, responsibilities and functionalities in archival information preservation. The resultant use cases and their requirements were then used to guide the functional definition of SCIDIP-ES services and toolkits. Their intended application is illustrated according to these use cases (Section 5). Finally, we conclude with a report on our development plan and the evaluation approach for testing the effectiveness of the infrastructure.

THE STATE OF OAIS CONFORMANCE WITHIN DIGITAL PRESERVATION

The field of long-term digital preservation has witnessed marked progress over the last few years. This has predominantly included newer preservation tools and systems being developed at both academic and commercial spheres to address the technical challenges of digital preservation. However, most adhere to the mantra “emulate or migrate”, ignoring the more general aspects of the OAIS Information Model, in particular the use of Representation Information (RepInfo) to help encapsulate knowledge and describe digitally encoded information. In this respect, semantic RepInfo is sadly neglected in terms of preservation.

The currently on-going SCAPE project [9] is building a series of preservation tools and services to address the scalability challenges of various preservation operations including ingest, characterization and migration while the ENSURE project [10] focuses on emulation in virtual machines. More recently, there has also been an increased focus on a few conceptual areas of work, particularly in terms of preservation policy models (e.g. SCAPE), and, to some extent, preservation cost estimation models (e.g. ENSURE). Unsurprisingly, the majority of these endeavors are inspired or influenced and, in some cases, even underpinned by the OAIS Reference Model. It should be noted that the Model is a conceptual model of the
functions and responsibilities of an archive viewed as an organization or other entity charged with the long term preservation of digital contents. Although it articulates the preservation processes, roles and responsibilities in an OAIS, it does not actually describe how to achieve this. Thus, an issue with the trend of preservation tools and services (both established and emerging) adopting the Model is that developers have different and, often, partial interpretations of the Model. For example, both the Safety Deposit Box (SDB [11], a commercial preservation system developed by Tessella [12]) and Roda [13] (an Open Source and research-based preservation repository developed by the Portuguese National Archives) claim to be OAIS-compliant. However, neither seems to support a RepInfo Network (see Section 5.1) linked to a defined Designated Community (DC)-an approach adopted by SCIDIP-ES (see Preservation Archive Creation). In the following sections, we present how SCIDIP-ES addresses OAIS requirements.

**Responsibilities of an OAIS Conformant Archive**

OAIS defines a set of mandatory responsibilities (OAIS Section 1.4) that digital preservation archives should fulfill in order to be conformant. These include:

- A conforming OAIS Archive implementation shall support the model of information described in 2.2 [of OAIS]. The OAIS Reference Model does not define or require any particular method of implementation of these concepts.

- A conforming OAIS Archive shall fulfill the responsibilities listed in section 3.1 [of OAIS].

For the latter we focus on:

a. Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.

b. Ensure that the information to be preserved is *Independently Understandable* to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.

c. Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.

d. Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.
The SCIDIP-ES e-infrastructure provides 13 generic services and tools (see Section 5) that facilitate fulfilling the above requirements. The remaining two OAIS responsibilities concern the operational policy of a preservation archive and are therefore outside SCIDIP-ES’ scope:

a. Negotiate for and accept appropriate information from information Producers.

b. Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.

The SCIDIP-ES infrastructure is designed for use by any organization involved with long-term digital preservation. However, our primary focus is to showcase its use in the context of ES organizations working with non-ES ones concerned with data preservation to confirm SCIDIP-ES’ broad effectiveness in helping to improve, and reduce the cost of, the way in which they preserve their ES data holdings. In the next section, we describe the SCIDIP-ES use cases which encapsulate the OAIS responsibilities 3(a-d) identified above and which have been contextualized with reference to ES community-specific requirements.

**The SCIDIP-ES Use Cases**

To ensure conformance, the OAIS Reference Model is consistently used in the design and functional specifications of the SCIDIP-ES services and toolkits. For instance, we aligned the three SCIDIP-ES high level use cases to the Reference Model to capture functional requirements in the long-term preservation of ES data. These use cases reflect the three distinct phases in the OAIS lifecycle model of data preservation:

- **Preservation archive creation**: To support the processes for setting up an archive or a new collection of data within an existing archive. These include identifying what kind of information need to be preserved to ensure the long-term usability of the ES data by the target DC; determine the user roles and preservation objectives; define a cost-effective preservation strategy and the correct procedures needed to implement the archive. For existing archival systems, this would also need to address the efficient integration of preservation processes within the existing system architecture.

- **Archived data access**: This relates to the access and exploitation of the preserved data. The functionalities required are search, discovery and retrieval of associated information in the archive, to allow data consumers to access, interpret and use the preserved data efficiently, correctly, and ideally within their familiar tools.

- **Archive change/evolution**: To protect the preserved data against changes which could range from the DC, technology, policies to funding issues etc. In the case of a changing DC, an archive may choose to update or augment the RepInfo associated with the preserved data to ensure that it is still intelligible and usable by the new DC. In the extreme case of an archive ceasing operation, it would need to identify and prepare to handover its holding to a suitable successor. Whatever the nature of the challenge, an archive should have the capability to plan responses to changes in a safe, cost-effective and sustainable manner.
MAPPING SCIDIP-ES USE CASES TO INFRASTRUCTURE SERVICES AND TOOLKITS

Figure 1 gives an overview of the SCIDIP-ES services and toolkits aligned to the familiar OAIS Functional Model. Although ES is used as the pathfinder for building the infrastructure, it should be stressed that the component services and tools are designed to support the full preservation life-cycle irrespective of the data domain as defined in the use cases above. In the next sections we motivate the key services and toolkits through the use cases and by reference to the needs of OAIS conformance.

Preservation Archive Creation

To conform to OAIS an archive must implement the OAIS Information Model and also fulfill, amongst others, the mandatory requirements about independent usability by its identified DC [see 3(a, b, d)], and the ability to hand over its digital holdings to a successor [see 3(c)]. In SCIDIP-ES, we see the ability to construct complete Archival Information Packages (AIPs) as key to the above OAIS requirements. We must of course bear in mind that an archive is unlikely to completely change the way in which it stores its data holdings and the related Preservation Description Information (PDI), both of which, experience shows, archives do possess. However, there is commonly no definition of the DC for which the data is being preserved nor of adequate RepInfo to support its usage. While this is marginally acceptable for rendered holdings, this is unacceptable for long term data preservation because the latter requires that data be process-able and can be combined with other, perhaps newer, data. Therefore we provide toolkits to create AIPs using functionalities provided by the Packaging Toolkit and Storage Service. For an existing archive, elements in the AIP would reference locations within the archive, and to RepInfo.
The amount of RepInfo required is measured against the perceived skills, resources and knowledge base available to the target DC. An archive may wish to broaden exploitation of its data holding by providing additional RepInfo for a wider group of users with different knowledge and resource bases. Given the potential diversity and quantity of RepInfo involved, an archive cannot by itself be expected to capture and manage all the RepInfo that it might require. To support these key requirements and help share the burden and efforts for preserving data long term, SCIDIP-ES provides the RepInfo Registry Service which would be used to query, retrieve and manage RepInfo required by a group of preservation archives. Note that we fully expect that repositories will locally cache the RepInfo they need. To facilitate the use of Rep-Info, the Registry will contain a special type of RepInfo called RepInfo Labels (RILs) which are pointers to multiple RepInfo objects. We note that a RepInfo Registry must itself be an archive, and so, logically, the actual RepInfo data objects would be preserved as OAIS AIPs which include PDI and its own RepInfo objects to facilitate interpretation. The latter construct gives rise to RepInfo Networks (RINs) and the Registry would also enable users to navigate a network of RepInfo objects to explore the knowledge represented. In SCIDIP-ES, an RIN represents the chosen solution for fulfilling a specific preservation objective (see next section).

The Gap Identification Service (GIS) is defined to help assess if a data consumer with a particular knowledge profile can ‘understand’ the preserved digital objects by identifying “gaps” in the corresponding RIN stored in the Registry [14]. Preservation planning and the creation of RepInfo are also ubiquitous activities for this use case and these are discussed in more details in the Archive Change Evolution section (5.3).

To assess if the defined archive is OAIS conformant and to identify potential areas for improvement, archivists may use the Certification Toolkit which implements the ISO16363 standards for Trustworthy Digital Repositories Audit and Certification [15] to perform a self-audit.

Archived Data Access

To meet users’ need to discover, access and use data from different sources [see OAIS responsibilities described in 3(b)] across multiple domains, SCIDIP-ES defines a Finding Aid Toolkit (FAT) to support the many existing domain search facilities by building on an archive’s current search and find capabilities. As highlighted in 3(d), preserved information should be disseminated with sufficient provenance information to provide quality assurance. SCIDIP-ES offers an implementation of the Authenticity Toolkit (AT) for associating provenance evidence and other PDI-related records on ingest. As in the previous use case, the Registry and GIS services could ensure that adequate RepInfo is available given the user’s registered DC knowledge profile and the requested data.

It is envisaged that the Data Virtualisation Toolkit (DVT, also see next section) may also be used in conjunction with specific types of RepInfo to facilitate data access. The toolkit would allow users to inspect the contents and structure of a digital data object using the associated semantic and structural RepInfo, e.g. viewing a NetCDF-based [16] file in tabulated
format without a dedicated NetCDF viewer. In this way, users are able to bring together and analyze data from multiple sources without having to use multiple dedicated software systems [see OAIS responsibilities described in 3(b)]; or preview data content of the preserved digital object before making the effort to obtain all the RepInfo needed to use the data.

Archive Change/ Evolution

Changes are inevitable given the timescale involved in preserving data long-term and an OAIS conformant archive must ensure that information is preserved against all reasonable contingencies [see 3(c)]. To monitor changes to technology or DC that might affect the long-term accessibility and usability of the preserved data, SCIDIP-ES provides the Orchestration Service (OS) to act as a knowledge broker which preservation archives may subscribe to receive notifications on specific topics of interest. The source of such information must come from experts in the various fields.

Those experts may also be asked to create additional RepInfo and so we provide the RepInfo Toolkit (RIT), a user-friendly GUI that is actually a collection of tools to facilitate the creation of RepInfo objects and interactions with the Registry. Some sub-components of this toolkit, e.g. the Data and Process Virtualisation Toolkits (PVT), are aimed at describing the data in more “virtualized” terms to help integrate the data into other software or interoperate data from multiple sources.

In terms of preservation, there are a number of basic strategies for preserving digitally encoded information. Besides using RepInfo (which includes emulation software), one could migrate (Transform in OAIS terminology) the data into different formats. CASPAR developed the Preservation Network Model (PNM) [17] which helps an archivist to plan and evaluate different preservation strategies, balancing factors such as costs against efficacy and risk against tolerance for given specific preservation objectives. The Preservation Strategy Toolkit (PST) provides an intuitive simple to use GUI for archivists to build PNMs as network diagrams, to capture the relationships between the PNM objects together with their attributes such as location, preservation objectives, risk, cost, tolerance and quality assurance, etc.

As an illustration of the notion of PNM and PST, a digital object might depend on multiple information objects (RepInfo) for a particular function or operation that could be performed on or using that object. The relationships between the objects could be either composite or alternate, where the former signifies a dependency on a combination of digital objects, while the latter represents optional dependencies. For example, to use a piece of scientific data, like combining it with newly acquired data, there are two basic approaches:

- Transform the data into a format compatible with the analysis tools used for the new data—this has associated computational and storage costs, and potential loss of information.

  OR

- Describe both sets of data in such a way that software can access and combine information from both. For example, describing the bit location and encoding of temperature measurements from the same geographical location for each dataset.
Whichever alternatives is used, both will require that the user is provided with Semantic Representation such as the units of the temperature measurements, whether they refer to surface or sub-surface temperatures and so on. Thus, recording these types of relationships is vital as they play an important role in preservation strategy formulation and evaluation in an archive. Both approaches have its own costs and risks, and these may be different for different archives for the same digital holdings. PST supports evaluation of alternative solutions within a PNM by the user defined risk and or cost profiles. PST could also be used to monitor the stability of an implemented PNM solution (in the form of an RIN in the Registry). For example, technology changes may render a ReplInfo obsolete leading to a breakdown in the RIN. PST could be used to re-evaluate the original PNM in view of the new information to identify a new solution which may simply involve using an alternative relationship already defined in the PNM.

Long-term data preservation requires long-term commitment. It may come the point when an archive needs to handover ownership of its data due to a change in policy or funding. To facilitate the changeover process, OS could also play the role of a generic knowledge broker that could be used to identify a suitable successor. In SCIDIP-ES, both ReplInfo and data objects are preserved as OAIS AIPs. This approach ensures that the tacit dependencies and knowledge are captured in a standard format [see OAIS responsibilities described in 3(c)] to guarantee the continuous accessibility and usability of the preserved data.

**SUMMARY AND FUTURE WORK**

In this paper we described the objectives of the SCIDIP-ES project which include delivering a sustainable OAIS conformant infrastructure for making data archives preservation-aware. In particular, we discussed in details how the SCIDIP-ES services and toolkits meet mandatory OAIS responsibilities through illustrating their usage within the data preservation lifecycle as represented by the three high level SCIDIP-ES use cases. SCIDIP-ES is now in its final year and has already delivered two releases of the software. The focus of the remainder of the project is to conduct intensive testing of the services and toolkits to collate feedback and refinement requirements to feedback into the last iteration of the product scheduled for release in February 2014. To ensure that the services and toolkits are useful for long-term knowledge preservation as defined by the OAIS Reference Model and can be handover as well as sustained, we are developing customized implementation with different configurations of the services and toolkits to integrate with selected ES partners’ data archives in different countries, including ESA [18], DLR [19], BGS [20] and BADC [21], etc. These custom implementations will serve as test beds for us to test the usability of the SCIDIP-ES software in different contexts with reference to functional and quality requirements. For example, ESA wishes to run their own Registry and a more complete set of SCIDIP-ES services and toolkits while BGS and BADC are willing to share the use of an external Registry and the efforts of maintaining commonly used ReplInfo objects. Finally, we re-iterate that the SCIDIP-ES infrastructure, though initially targeted at ES, is intended to have wider application across scientific disciplines. To achieve this, we designed the services and toolkits to implement the OAIS Information Model and provide generic functionalities to support the relevant mandatory OAIS responsibilities—which are domain agnostics.
ACKNOWLEDGEMENT

SCIDIP-ES is funded by the Seventh Framework Program (FP7) of the European Commission (EC) under the Grant Agreement 283401.

REFERENCES

[18] European Space Agency (ESA): http://www.esa.int/ESA
[21] British Atmospheric Data Centre (BADC): http://badc.nerc.ac.uk/home/index.html
Blank Page and Duplicate Detection for Quality Assurance of Document Image Collections

Roman Graf\textsuperscript{1}, Ross King\textsuperscript{1} and Sven Schlarb\textsuperscript{2}

\textsuperscript{1}Research Area Future Networks and Services Department Safety & Security, Austrian Institute of Technology
\textsuperscript{2}Austrian National Library
E-mail: \textsuperscript{1}roman.graf@ait.ac.at/ross.king@ait.ac.at, \textsuperscript{2}sven.schlarb@onb.ac.at

Abstract—Digitization workflows for automatic acquisition of image collections are susceptible to errors and require quality assurance. This paper presents an automatic expert system for long term preservation that supports decision making for blank page and accurate duplicate detection in document image collections. The important contribution of this work is a definition of the expert rules with associated severity level and its automatic computation. Our goal is to create a reliable inference engine and a solid knowledge base from the output of an image processing tool that detects blank pages and duplicates based on methods of computer vision. We employ artificial intelligence technologies (i.e. knowledge base, expert rules) to emulate reasoning about the knowledge base similar to a human expert. In order to improve analysis accuracy we use OCR tool for blank page and duplicate detection. The novelty of this approach is an application of OCR method for this task. A statistical analysis of the automatically extracted information from the image comparison tool and the qualitative analysis of the aggregated knowledge are presented.

Keywords: Expert System, Digital Libraries, Image Processing

INTRODUCTION

During the last decades, libraries, archives and museums have been carrying out large-scale digitization projects at different scales. New digital collections comprising millions of books, newspapers, journals, and other digital object representations have been created, and internet-based access to a wide range of cultural heritage resources that so far have only been available in analogue form is now possible. Depending on the type of digital object, a typical book, newspaper or journal collection item can contain hundreds or even more document images and other related information entities. Due to the scale of digital information that has to be managed, memory institutions are facing a paradigm shift in the way how preservation, maintenance, and quality assurance of these collections have to be addressed. For that reason, automated solutions for data management and digital preservation are absolutely necessary.
This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137)

In a typical book digitization workflow, the ability to update a digital copy is a frequent requirement. Either a new digital copy is created by scanning the original analogue resource again, or a new digital derivative based on the raw digital object is produced. The new derivatives are either created to get an improved representation of the digital image by removing page borders or skew, or to enrich the digital object by adding related information like full text, layout or semantic representations, etc.

In this context, image analysis and comparison technology can help to align and document changes that occur from one version to another. The detection of blank pages (see Figure 1) in a digital book and the selection between the old and the new version of the associated documents (see Figure 2) are basic operations in this regard. Based on this information a decision support system can automatically make a recommendation if a digital object can be safely overwritten or if human inspection is required. The Expert System proposed in this article is based on image processing and OCR methods, which implement image analysis and comparison for digitized text documents.

The main contribution of this paper is the development of a rule-based Expert System for the analysis of digital document collections, and for reasoning about analyzed data and for assessment regarding blank pages and duplicated images.

The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the duplicate and blank page detection process and also covers image processing and expert rules definition issues. Section 4 presents the experimental setup, applied methods and results. Section 5 concludes the paper and gives outlook on planned future work.
**RELATED WORK**

In artificial intelligence, rule-based Expert Systems support the techniques of quality assurance for digital content and replace a human expert regarding the decision-making process in a particular domain. An expert system comprises the inference engine that is employed for reasoning about the knowledge base. The knowledge base contains expert knowledge in form of data or rules.

![Fig. 2: Sample of Book Scan Sequence with a Run of Eight Duplicated Pages: Images 10 to 17 are Duplicates of Images 2 to 9 (Book Identifier is 151694702)](image)

The implementation of an expert system for color retrieval described by Yoo et al [12] proposes an image retrieval system using color-spatial information from the content-based image retrieval applications. In contrast to our Expert System approach, the described system does not implement rules, although it does perform similarity computation. In our duplicate detection approach the similarity computation task is provided by the image processing techniques. One of the possible tools is a matchbox tool [5], which is a modern quality analysis tool based on SIFT [7] feature extraction. In contrast to SIFT descriptor matching, the matchbox tool makes use of a bag of visual words (BoW) [3] algorithm. The SIFT descriptor approach is very similar to the matchbox tool, with the difference that we do not use BoW and structural similarity matching in order to increase performance, whereas the matchbox tool guarantees better accuracy employing these techniques. Typically, approaches in the area of image retrieval and comparison in large image collections make use of local image descriptors to match or index visual information. Near duplicate detection of key frames using one-to-one matching of local descriptors was described for video data [13]. A BoW derived from local descriptors was described as an efficient approach to near-duplicate video key frame retrieval [11]. Local descriptors were employed for the detection of near-duplicates [6].
Several authors mention that the use of optical character recognition, which is an obvious approach for the extraction of relevant information from text documents, is quite limited with respect to accuracy and flexibility [2], [9]. But employing of OCR methods in our approach meant just as a complimentary support for image processing methods and is just one expert rule in a pool of other duplicate detection rules. And innovative OCR methods application for blank page detection does not require high accuracy, since we employ a threshold metric for result estimation.

The rule-based system presented by Bernard [1] is designed for process and power control in a power plant. In order to evaluate a control action the relevant parameters are measured. Actions are specified in rules. Given the current state of the plant and the desired objectives, the fuzzy logic and inference engine is used to search through the knowledge base in order to identify those rules that are applicable. This approach is very similar to our Expert System organization, with the difference that we have a different application field and another input parameters. Compared to existing systems the proposed system is more efficient due to the use of SIFT features instead of color signatures and filtering, and it is more simple without the use of linguistic variables for fuzzy logic. The proposed system is unique for the given domain.

**DUPLICATE AND BLANK PAGE DETECTION PROCESS**

Due to huge number of images and text documents in modern digital collections the quality assurance plays an increasingly important role. Decision making process for quality assurance in digital preservation requires deep knowledge about image processing, file formats and regular library processes. The manual search for such knowledge is very time consuming, requires an expertise in the domain of digital preservation and image processing skills. A consistent collection should not contain duplicates or blank pages. Therefore we aim at providing automatic image duplicate and blank page identification and verification methods in order to support decision making regarding the collection cleaning. An additional challenge for manual analysis is that existing information often is either not structured or is only partly structured.

The Knowledge Base shown in Fig. 3(a) is required in order to collect information and to perform automatic document assessment and duplicates detection. A basis for accurate reasoning is information aggregated from digital documents in image collection and from knowledge provided by human experts.

**Expert Rules Identification**

To organize the Knowledge Base we must structure the information that has been obtained from the domain experts of digital preservation and from conducted experiments. We define typical scenarios and identify the parameters used by library experts for collection handling. Then we define the linguistic labels to classify measured values of each parameter and associated ranges. Finally, we determine the conditional rules that relate these linguistic labels to specific consequences. Information retrieved from the image collection is processed by the
customized domain model. This model enables structured and maintainable handling of analyzed data. If necessary, the data could be stored in a database for further treatment. A user communicates with the Expert System by sending a request query and receives an advice in response. A user could leverage these rules according the requirements and circumstances for a particular book or collection for example if a file name has a semantic meaning or if the file size is of interest for analysis. Possible actions according to the advice provided by the Expert System include removing of a document, ignoring the advice and performing a new scan for the particular image or a collection including similarity analysis.

The previously defined rules should be organized in order to process input statements (assertions) and to infer appropriate advice and conclusions. Forward rule chaining for duplicate or blank page detection is presented in Fig. 3(b).

Forward chaining is the process of moving from the “if” patterns (antecedents) to the “then” patterns (consequents) in a rule-based system. We consider the antecedent as satisfied when the “if” pattern matches the assertion. Assertions are depicted by black rectangles on the input side and by the white rectangles on the output side, respectively. The rules are presented by blue half-spheres. A specific rule is triggered if all of its antecedents are satisfied. A triggered rule is considered as fired if it produces a new assertion or performs an action on the output (white rectangle). Since our Expert System is focused on duplicate and blank page detection there is no need for any conflict resolution procedure to resolving possible rule conflicts. The rules are weighted according to their severity and can be extended. The weights are customizable and can be adjusted according to user requirements.
In Fig. 3(b) we present rules distinguishing blank pages from non-blanked ones. The rule-base system starts blank page identification with the rule D1. Suppose that SIFT features score of particular document is over particular adaptive threshold. Then if the antecedent pattern matches that assertion, the value x becomes “is a blank page candidate” and the rule D1 fires. Because the document contains text (OCR size) and file size is not null, rule D5 fires, establishing that the document “is a blank page”. Similarly we go through remaining rules. The final conclusion of the rule-based system is whether there is a blank page observed. The inference engine performs conditional rules and blank page analysis, infers appropriate action and formulates advice using relation of linguistic labels to specific consequences. The OCR score values could be leveraged for duplicate detection 3(c) as additional method to the image processing methods and metadata analysis.

![Fig. 4: Evaluation Results Samples from Book Identifier 151694702 for Duplicate Detection with SIFT Feature Matching Approach: (a) Similar Pages with 419 Matches, (b) Different Pages with 19 Matches](image)

**Image Processing**

Application of different digitization methods for the same document might result in information significantly differing at the image pixel level. This depends on performed geometric modifications as well as filtering, color or tone modifications. Therefore, we used interest point detection along with local feature descriptors, which have proven highly invariant to geometrical and radio metrical distortions [7][10] and were successful applied to a variety of problems in computer vision. To detect and describe interest regions in document images we used the SIFT approach. The key point locations are identified from a scale space image representation. In our approach we make use of a direct matching of SIFT descriptors in contrast to the procedure [8], where all descriptors for all images of the same category are clustered independently and subsequently appended to the BoW. Our approach differs also from the matchbox approach, where a list of clustered descriptors is constructed and in the second step this list is clustered in order to obtain a dictionary for the whole book. The similarity score between two documents is obtained from the comparison of corresponding SIFT features followed by OCR output comparison.
Duplicate Detection Workflow

Collection analysis is conducted according the quality assurance workflow shown in Figure 3(c). The user triggers a complete collection analysis, the results of which are stored in a text file. In order to detect duplicates we aggregate collection specific knowledge and analyze collections using SIFT feature extraction demonstrated in Figure 4, filtering and matching, as well as the OCR analysis. Local feature descriptors are extracted from SIFT key points. Robust descriptor matching employs the RANSAC [4] algorithm which is conditioned on an affine transformation between key points locations. In the next step we compare images by matching consistent local features with each other. Finally human expert should validate the list of duplicate candidates. Additionally, duplicate candidates contained in a shortlist can be validated by OCR comparison, which requires additional computation time and also is limited for documents with printed text written in supported language.

Evaluation

The goal of evaluation is an application of different methods for collection analysis for duplicates and blank page detection resulting in its cleaning, i.e., a collection with no duplicates and blank pages. Additionally, a statistical overview of evaluated data and characteristics like performance and accuracy is delivered. The suggested Expert System processes reasons on found blank pages and duplicates and generates advice on how to clean up the collection.

Hypothesis and Evaluation Methods of the Collection Analysis

The presented two evaluation use cases find duplicate pairs and blank pages and present them for additional manual analysis and collection cleaning. Our hypothesis is that automatic approach should be able to detect blank pages and duplicates with reliable quality. We consider two use cases. First one is a duplicate detection. The OCR analysis should prove the results of image processing methods and OCR scores for similar files should have similar OCR scores. The second use case is blank page detection. For this use case OCR scores should be null or near to null as well as SIFT descriptors score should be very low. We also aim to evaluate whether file size of blank page could be a reliable parameter for blank page analysis. If described hypothesis is true then this methods would be a significant improvement over a manual analysis. The considered collection with identifier Z151694702 is provided by then Austrian National Library and contains 730 documents corresponding to a single book. Manually created ground truth was available.

We assume that employing of SIFT feature comparison and calculation with an OpenCV 2.4.3 based python workflow and OCR analysis will demonstrate good performance by sufficient good accuracy. Evaluation takes place on an Intel Core i73520M 2.66GHz computer using Java 6.0 and Python 2.7 languages on Linux OS. We evaluate duplicate candidate pairs, calculation time and calculation accuracy for each evaluation method. For OCR analysis we use Tesseract 3.02 tool.
Experimental Results and its Interpretation

The threshold value 0.9 was determined using statistical approach and robust estimators. The threshold can be adapted dependent from the content. In the first instance we apply “SIFT features” rule. For this rule we compute average similarity score over the all similarity scores provided as an output of the SIFT descriptors matching analysis. In conjunction with similarity threshold rule we are able to isolate most of the duplicate pairs. The number of pages between the original and the new version of the duplicated documents in the collection is an additional help to find duplicates, since duplicates often appear in a sequence. Some of the detected duplicates have a dominating color and relative high similarity score like documents 2-9. These documents should be verified manually and independent from the average similarity score and offsets. OCR scores comparison provides additional help for working with these documents.

The manual analysis of the test collection shows eight duplicate pairs. The automatic approach of duplicate search did not find three duplicated pages (3, 5 and 6) which were identified as duplicates by manual analysis. The reason for that is the computed average similarity score was higher than the scores of pages 3, 5 and 6. In this specific case we have to deal with nearly empty pages with dominating white color, which makes it difficult to identify these pages as a pair of duplicates. The pages in the range 108 to 115 and pages 117, 124 are detected as false positives by the automatic analysis. In contrast to the dominating color case similarity scores are in range here. Manual checking of mentioned pages reveals that there are no duplicates. The reason for detecting false positive is a high structural similarity of digital image data. But this high similarity does not always mean semantically text similarity that can be validated only by human expert. The SIFT features method scores with five true positives. The calculation times of SIFT method is 95940 seconds. Experiment shows that SIFT feature matching method detected 13 false positives, whereas OCR validation method demonstrates three false positives. The OCR method for duplicate detection used for validation of SIFT analysis results demonstrates sufficient accuracy with seven correct detections among eight possible and three false positive results. The total calculation time for OCR analysis is 11418 seconds. The results of OCR analysis are very dependent on printed text and image quality, threshold setting and OCR tool quality. Texts of duplicate files extracted by OCR method can differ and require further analysis. The manual search for blank pages in the test collection shows 18 blank pages with four cover pages among them that are not fully blank and are brown colored. The automatic approach of blank pages search successfully detected all blank pages and one false positive. The OCR output score for blank pages mostly is 0 or three. Manually we also detected 13 pages with large empty areas that take approximately the half of the document. Seven of them we were able to detect automatically but this analysis is not very reliable, since the definition of such pages is very difficult. One page (index 634) was mistakenly tagged as a blank page, whereas it is a normal text page. The reason for that could be that the quality of the text was not sufficient and OCR output size was 0. A typical text document image in SIFT analysis workflow contains up to d = 2,000 descriptors. Direct matching of feature descriptors requires d2 = 4 • 106 descriptor comparisons for a single pair of images. For a sample book with n =
730 pages $n \times (n - 1) \approx 532.170$ OCR score comparisons are necessary. The text, resulting from the OCR evaluation could not be regarded as a reliable evaluation parameter for duplicate detection, due to strong dependency on image quality. But the size of this text can be successfully employed for blank page detection. The advantage of the OCR method in comparison to SIFT method is that we analyze each file only once. Therefore the OCR method presents more reliable results for blank page analysis and can be applied for quality assurance of digital collections. All of these approaches help to automatically find out duplicate and blank candidates in a huge collection. Following this, manual analysis of duplicate candidates separates real duplicates and blank pages from structural similar documents and evaluates resulting duplicate or blank pages list. Presented methods save time and therefore costs associated with human expert involvement in quality assurance process. Therefore our initial hypothesis is true. But further research is required to improve performance and accuracy metrics of mentioned methods.

![ROC Space Plot](image)

**Fig. 5: ROC Space Plot**

The duplicates and blank pages search effectiveness can be determined in terms of a Relative Operating Characteristic (ROC). Similarity analysis divided the given document collection (book identifier is Z151694702) in two groups “duplicates and “single or “blank” and “text/picture” by associated thresholds. The inference engine detected five true positive TP duplicates, 712 true negative TN documents, 13 false positive FP duplicates and three false negative FN documents. The main statistical performance metrics for ROC evaluation are sensitivity or true positive rate TPR and false positive rate FPR (see Equation 1).

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

Therefore the sensitivity of the presented approach is 0.625, the FPR is 0.018.
The associated ROC values for OCR blank page and duplicates detection are represented by \((0.007, 1.0)\) and \((0.3, 0.875)\) points respectively. The ROC space demonstrates that the calculated FPR and TPR values form all these points are located very close to the so called perfect classification point \((0, 1)\). These results demonstrate (see Figure 5) that an automatic approach for blank page detection is very effective and it is a significant improvement compared to manual analysis. The best possible classification is represented by the point \((0, 1)\). The distribution of collection points above the red diagonal demonstrates quite good classification results that could be improved by refining of rules. Therefore, OCR analysis can be suggested as an effective method for blank page detection and as a verification step for duplicate detection.

**CONCLUSION**

We have presented an automatic expert system that supports decision making for blank page and accurate duplicate detection in document image collections. This system uses automatic information extraction from the image processing tools, performs analysis and aggregates knowledge that supports quality assurance process for preservation planning.

An important contribution of this paper is the definition of expert rules and creation of reliable inference engine with the solid knowledge base from the output of the image processing tools that detects blank pages and duplicates based on methods of computer vision and OCR. We employed AI technologies (i.e. knowledge base, expert rules) to emulate reasoning about the knowledge base like a human expert.

The experimental evaluation presented in this paper demonstrates the effectiveness of employing the artificial intelligence techniques for knowledge base design and for generating reasoned suggestions. The Expert system reliably detects image sequences containing duplicated or blank images for typical text content. An automatic approach delivers a significant improvement when compared to manual analysis.

The expert system for document image collections presented in this paper ensures quality of the digitized content and supports managers of libraries and archives with regard to long term digital preservation.

As future work we plan to extend an automatic quality assurance approach of image analysis to other digital preservation scenarios. The rules could be combined with different subject categories in order to meet requirements for different use cases.

**REFERENCES**


Scaling Up and Scaling Out: Leveraging Preservation Infrastructure and Experience to Benefit the Community

Kate Wittenberg¹, Amy Kirchhoff² and Sheila Morrissey³

¹Managing Director, Portico, 151 E 61st St. New York, NY 10065,
²Archive Service Product Manager, Portico, 100 Campus Drive, Suite 100, Princeton–08540, NJ
³Senior Research Developer Portico, 100 Campus Drive, Suite 100, Princeton–08540, NJ
E-mail: ¹kate.wittenberg@portico.org, ²amy.kirchhoff@portico.org, ³sheila.morrissey@portico.org

Abstract—Now well into its eleventh year as a preservation archive and service, Portico’s preservation infrastructure—hardware, software, and key data and metadata models and definitions—has been subject to a continual process of review and revision that makes it possible for us to leverage our work to benefit the broader community. Since 2012, Portico has been delivering e-journal content to the British Library as part of their legally mandated deposit program, enabling the British Library to benefit from existing preservation expertise to manage the normalization of e-journal content. In this paper we will discuss Portico’s development, the scaling up of our preservation capacity, the challenges and opportunities in our partnership with the British Library, and the value of leveraging and sharing existing preservation infrastructure, skills, and experience for the good of the broader community.

INTRODUCTION

Portico is a digital preservation service provided by ITHAKA¹, a not-for-profit organization with a mission to help the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Created in 2002, Portico was founded to build a sustainable digital archive to serve the academic community to enable publishers and libraries to be secure in the long-term preservation and accessibility of the e-journals being licensed, and to realize tangible benefits as they transitioned to greater reliance on digital content. Portico began as a project funded by The Andrew W. Mellon Foundation to further its seminal E-Journal Archiving Program. Today, with over 25 million articles preserved, Portico is among the leading digital preservation services in the world.

¹http://ithaka.org/
Our approach to digital preservation is comprehensive—combining long-term content management and organizational commitment with a philosophical dedication to addressing the needs of tomorrow’s scholars. Portico preserves content through a format-based archive management strategy. The key points of this strategy are: identifying significant preservation metadata at the initial point of preservation, and conservative and pragmatic migration of content at the point where such an activity is both safe and necessary. While the archive is dark, publishers and libraries are provided with audit privileges that allow them to review the status of content. Content in the archive becomes “light” for faculty, staff, and students at participating libraries whenever a trigger event occurs (cessation of a publisher’s operations; discontinuation of a title by a publisher; back issues no longer offered by a publisher; or catastrophic and sustained failure of a publisher’s delivery platform). In addition, the majority of the titles in Portico are available for post-cancellation access if needed. Upon receipt of a claim from a participating institution and confirmation of the past subscription status by the publisher, campus-wide access is provided to the requesting participating library.

For Portico, the key goals of digital preservation include:

- **Usability**: The intellectual content of the item must remain usable via the delivery mechanism of current technology
- **Authenticity**: The provenance of the content must be proven and the content an authentic replica of the original
- **Discoverability**: The content must have logical bibliographic metadata so that it can be discovered by end users
- **Accessibility**: The content must be available for use by the appropriate community

To meet these goals, we have defined and follow exacting standards and processes for content management, maintenance, and replication of the archive; we conduct self-checks and third-party archive certifications to ensure quality and security; and we maintain a delivery system and services to provide access to users in a manner that is easy to use and integrated with other online resources.

Three years ago, we began a substantial transformation of Portico’s internal environment with the goal of building a premier content management organization with robust and flexible platforms and processes, with increased efficiencies and economies of scale. Part of this development involved expanding the archive’s capability to handle new content types. Portico extended its content model to accommodate e-books and digitized materials (newspapers, books, documents, pictures, etc.), modified its preservation metadata schema, and migrated roughly 15 million METS\(^2\)-based metadata files to a new preservation metadata format. In addition to these internal infrastructure and technology developments, Portico underwent a successful audit that made it the first CRL certified Trustworthy Digital Repository. Portico also submitted the winning proposal to become the legal deposit service for the British Library, initiating the development of a new type of preservation service, and

\(^{2}\text{http://www.loc.gov/standards/mets}\)
demonstrating the viability of Portico preserved content outside of the Portico system. As a result of this work, Portico is positioned as a trusted and sustainable service in the rapidly changing scholarly communications environment, and is able to use its infrastructure and experience to benefit the broader preservation community.

**THE CURRENT ENVIRONMENT**

We recognize that Portico works within a rapidly-changing environment, and that we must be aware of these changes in order to respond effectively. Scholarly communication, particularly book publishing, continues to evolve in the digital space, along with an associated need for new approaches to preservation of the material (in both old and new content types) coming from traditional scholarly publishers, but also from libraries and individual scholars involved in the creation of digital research. In addition, there has been a steady rise in the pressure on scholarly publishers for open access to scholarly research, and a need for innovative and effective approaches to the sustainability, persistence, and preservation of this content.

Research libraries are also undergoing important changes. They have for some time been pressured by the combined challenges of shrinking resources resulting from the stressed macro economy, and by the disintermediation of the role of the librarian caused by new digital modes of scholarly publishing and research. The scholarly literature increasingly acknowledges a professional change in focus to “connections, not collections.” Libraries are, jointly and individually, developing strategic plans whereby they take the lead in providing infrastructure, tools, and support for faculty research workflows. Libraries have taken the initiative in implementing several projects focused on large-scale, substantive collaboration. Hathi Trust is perhaps the most fully developed exemplar of this trend toward resource pooling replacing local work. They have made great strides toward an infrastructure that includes digitization, aggregation, and preservation of non-digital content, acquisition, access, and preservation of e-journals, and provision of a university press digital publishing platform.

Another change is the increasing engagement by faculty in digital initiatives in the scholarly domain, including challenges to authoritative channels of scholarly communication, collaboration, research, and instruction. Digital humanities centers have developed rich corpora of digitized and born-digital literary and historical source documents, both primary and secondary, often enriched with mark-up and hosted within a framework of tools for discovery, for visualization, for both “close” and “distant” (machine assisted) reading, and for crowd-sourced enrichment of the collection. These initiatives raise critical questions concerning the long-term accessibility of non-traditional scholarship such as computer models, data, and the human interaction with these materials and software. Increasing numbers of scholars are trying to alter the “sociology” of scholarly communication, collaboration, and evaluation (including the factors affecting tenure decisions), and are looking to associate scholarly distinction not with publication, but instead with impact, including reception via social media. Faculty and researchers are actively seeking collaboration with the research library community to establish permanent homes for access to, and preservation of, these collections.
As we think about our role as a preservation organization, we see many opportunities to engage usefully and productively with these developments. In order to do so, however, we will need to think innovatively about our internal systems as well as our external relationships with the community as a whole, and with outside initiatives that intersect with our work. We must now see the possibility of productive partnerships and collaborations that will provide critical services to an increasingly complex community of content.

**Scaling Up Infrastructure: Technical Challenges of Preservation at Scale**

As noted in the call for papers for this conference, the digital universe in India alone is expected to increase by more than an order of magnitude by 2020. The capacity and capabilities of the digital preservation community will have to keep pace with this accelerating demand. As has been the case for many of our institutional colleagues in digital preservation, we at Portico have found that this reality of ever-increasing scale means a continuing process of reflection and re-invention, both of our technical and of our organizational infrastructure.

What are the technical challenges of such dynamic growth? As Clay Shirkey said, in reporting the results of the United States Library of Congress-sponsored Archive and Ingest Handling Test,

Scale is a mysterious phenomenon–processes that work fine at one scale can fail at 10 times that size, and processes that successfully handle a 10-times scale can fail at 100 times. […] Institutions offering tools and systems for digital preservation should be careful to explain the scale(s) at which their systems have been tested, and institutions implementing such systems should ideally test them at scales far above their intended daily operation, probably using dummy data, in order to have a sense of when scaling issues are likely to appear. [7]

Portico is now well into its eleventh year as a preservation archive and service. As part of its institutional mission and definition, Portico’s preservation infrastructure–hardware, software, and key data and metadata models and definitions–has been subject to a continual process of review and revision, consonant with a disciplined understanding of normative software system lifecycles, and of the ever-changing information environment in which we all find ourselves. In the course of that process, Portico has undertaken two sorts of major system migrations centered on issues of scale: a refactoring of the workflow system (“ConPrep”) that ingests publisher content streams to scale it up by a factor of 60 to a current capacity of approximately one million files per day, [5] and the migration of over 15 million preservation metadata records to a new, semantically richer, syntactically leaner metadata model. [4]

---

3http://www.ndpp.in/APA-DPDTR-2014/
In scaling up the capacity of Portico’s workflow system, a good deal of analysis went into determining what optimizations, or refactoring, of the original system design were necessary to enable us to take advantage of hardware scaling with newer, faster, and less expensive systems and storage devices. A number of factors placed pressure on the original design [5] [1], including:

- A performance bottleneck in the system component responsible for persisting workflow information to the database.
- The total size of the content, which caused performance degradation of disk drives as the size of disk volumes grew, and consequently the length of time it took to perform fixity checks on archive content.
- The sheer number of files that make up individual archival components (a journal article, an e-book), which resulted in significant overhead in per-file reading and writing, and in the overhead of loading files into the content management system.
- The lack of standard submission packaging for electronic journal and other content, resulting in the need for batches with different characteristics (for example, batches comprising a single journal issue with a large number of files, and other batches with multiple issues, each with smaller numbers of files), with different processing requirements, making varying capacity demands, and necessitating the intelligent automated management and balancing of batch and process scheduling.

Portico learned many lessons about what it understands to be the ongoing challenge of system refinement and refactoring to keep pace with increases in scale during the process of system migration in 2007. We learned about the need to gather data, to analyze and experiment, and rigorously to test our changes—also at scale—both before and throughout the migration process. We learned to look for the consequence of optimization of one part of the system on other parts, and to balance all parts (hardware, software, people and processes) of the system as a whole. Increasing the capacity of one part of the system required that we create tools to increase the scope and capacity of automation of such system administration tools as cleanup and maintenance, logging and log management, and, perhaps most importantly, in the automation of quality assurance of the content being manipulated, ingested, and managed by the archive.

These were all lessons that stood us in good stead a few years later, when, in anticipation of new content types (electronic books, digitized collections) coming into the archive, and aware of refinements to, and maturing of, standards and practices in the digital preservation community (for example, the publication of the PREMIS data dictionary), Portico undertook a revision of its content model. This in turn resulted in new schema for Portico’s preservation metadata, and the consequent migration of over 15 million already-existing metadata files to the new content model. As we noted in describing this migration, preservation metadata plays a key role in the archive:
The archive’s preservation activities are made manifest through the preservation metadata generated and collected throughout the life cycle of a preserved object. In Portico’s case, these data can be generated during processing in ConPrep, at ingest to the archive, and as preservation activities take place thereafter.

This meant that nearly every part of the system was likely to be “touched” in some way by the metadata migration. It meant as well, as indeed Portico’s experience in scaling up ConPrep had demonstrated, that the migration would need to be carefully thought through, documented, managed, and coordinated amongst staff who would also be engaged in other work. [4]

Again, we undertook a careful process of analysis and planning, testing and tuning, to accomplish the migration.

**Scaling Up Infrastructure: Organizational Challenges of Preservation at Scale**

The Portico experience of scaling up its system capacity in particular gave us a view of the human impact of scale. Particularly affected was the Portico production staff, who were responsible for superintending the loading of content, for performing quality assurance on new content streams, and who performed problem resolution during processing. [5] Portico was challenged to accomplish an order of magnitude increase of system capacity, without necessitating an order of magnitude increase in the number of production staff required to handle this scaled-up capacity. Accomplishing this meant developing more automated tools, redesigning the graphical user interface, and refining and redefining staff roles and responsibilities.

The organizational challenge of preservation at scale is something that Portico shares with many others working in digital preservation. Realizing the growing awareness of, and experience with, the complexities of scale, Portico, along with colleagues from the National Library of the Netherlands, organized a workshop on preservation at scale to be held as part of IPRES 2013. Over thirty people from seventeen countries attended this workshop, including representatives from national libraries of Sweden, the Netherlands, France, the United Kingdom, China, and Germany, as well as from many university libraries, NGOs, and digital preservation consortia. As indicated in the report on the workshop [3], there were a number of broad categories addressed by the workshop speakers and participants, including:

- The technological adaptations to collect and preserve ever-increasing amounts and varieties of digital content while taking advantage of new advances in both hardware and software.
- The institutional adaptations, sometimes including new institutional self-definition, necessitated by the increasingly large role that the preservation of digital, as opposed to analog, objects has come to play in memory institutions.
- Quality assurance at scale and across scale, often at cross-purposes with the large-scale automation required to handle increasingly large amounts of digital content.
• The scale of the long tail of unpreserved digital content.
• Economies and the sometimes unexpected diseconomies of scale.

It was particularly interesting to hear reports on the institutional challenges and responses to those challenges from the National Library of the Netherlands [8], the British Library [6], and Harvard University Library [2]. For these institutions, the sheer scale of the acquisition, management, and preservation of digital content has moved digital preservation from what might be called a satellite activity, to the core of each library’s mission and self-definition. This in turn has often resulted in dramatic organizational restructuring. As the workshop report notes,

This shift--from relegating the preservation of digital content to an organizational sub-unit, to ensuring that digital preservation is an organization-wide endeavor--is challenging, as it requires changing the mindsets of many in each organization. It has meant making choices and reallocation of resources from other activities, recognizing that the organization cannot do everything. It has necessitated strategic planning and budgeting for long-term sustainability of digital assets, including digital preservation tools and frameworks—a fundamental shift from one-time, project-based funding. It has meant comprehensive review of organizational structures and procedures, and has entailed equally comprehensive training and development of new skill sets for new functions. [3]

SCALING OUT: COLLABORATIONS AND CENTERS OF EXCELLENCE:
THE BRITISH LIBRARY PARTNERSHIP

As part of our mission, Portico explores new opportunities that address the emerging preservation needs of the broader community. These opportunities may include forming partnerships with key players and stakeholders to advance an overall, community preservation agenda and agreed-upon goals. As one step in this area, Portico is engaged in research related to the emerging preservation needs of researchers, including the preservation of data that supplements and supports journal articles and monographs, and the preservation of the links between published content and data.

We recognize that as an established, trusted, and sustainable preservation partner for publishers, libraries, and individual scholars, Portico can and should play a key role in leveraging our work in order to support the print-to-digital transition for the global community of libraries and individual scholars. We need to recognize opportunities and initiate projects that address current and emerging preservation needs of the community as a whole.

As part of this work, we are involved in a partnership to provide preservation services for the British Library’s legal deposit program. In 2012, Portico and the British Library jointly worked on a pilot project to identify the tools needed to automate the transfer of content from Portico to the British Library and the communication channels needed to identify and correct problems in the content transfer. In early 2013, the U.K. Parliament approved legal deposit legislation requiring the deposit of all content electronically published in the U.K. with
the national deposit libraries, and in April 2013, Portico began to deliver preservation files for e-journal content to the British library as part of this ongoing deposit program. Through 2013, Portico will be delivering content for 3 publishers and nearly 1500 journals (Portico has delivered over 108,000 articles to the British Library between April 2013 and November 2013).

The relationship is highly collaborative and allows the British Library to leverage the existing Portico technical infrastructure and staff expertise to manage the normalization of e-journal content to meet their legal deposit obligations. In addition, the British Library may also leverage the existing Portico publisher relations infrastructure and staff expertise to manage the content conversations with the publishers. The publishers that choose to participate in the program make arrangements with the British Library and sign an addendum to their existing agreement with Portico or, if they are not a Portico publisher, sign an agreement specific to the British Library arrangement with Portico.

The British Library project has influenced some changes in the Portico processes. Due to the long-term nature of Portico’s preservation model, there has historically not been an urgent need to process and ingest content into the archive immediately upon its publication. However, the British Library needs to ingest the content into its preservation system quickly upon publication in order to make that content available in the Library reading rooms and to deliver the content to the other deposit libraries in the U.K. For the British Library publishers, therefore, Portico has increased the speed and flexibility of its receipt and ingest process. As with any electronic system, managing the limited set of special cases (issues published without an issue number, issues published and never delivered to Portico, etc.) requires more time from Portico staff than the large amount of content that simply processes straight through. As Portico is providing this service simultaneously for the content it is preserving and the content it is delivering to the British Library, the overall costs to the community are reduced.

Portico and the British Library have collaboratively designed a delivery system that relies on a number of standards, including:

- BagIt\textsuperscript{4} for packaging and transmission of content.
- Journal Article Tag Suite (JATS)\textsuperscript{5} for e-journal XML mark-up.
- Dublin Core\textsuperscript{6} for metadata mark-up.

We have learned a great deal about the challenges and the benefits of this arrangement, which we can now share with the broader community. Our collaborative work with the British Library has involved defining and architecting the Portico systems and exploring various content transformation and delivery options, and then making final choices together. We have addressed challenges around coordinating project management, working with different software development styles, and communication between the geographically

\textsuperscript{4}http://sourceforge.net/projects/loc-xferutils/
\textsuperscript{5}http://jats.nlm.nih.gov/
\textsuperscript{6}http://dublincore.org/documents/dces/
distant teams. We believe that the lessons we have learned through this project will be useful
to the community as many consider how to manage new preservation needs while realizing
economies of scale.

First, it is important that we have identified ways to ensure that the library community is
not paying to develop the same tools multiple times. The Portico/British Library partnership
allows each party to focus on separate activities and leverages work each has already done.

Second, we have learned that communication, rather than technology, is the most
common challenge for this sort of collaborative project. Regular communication and close
cooperation between the Portico and British Library team members on a geographically and
organizationally complex project has been a key factor for success. In addition, the teams had
to agree on secure and efficient protocols for moving content back and forth between the two
organizations. With a clear workflow process in hand, both organizations could independently
develop their work in parallel.

Finally, we have come to understand that by leveraging and sharing Portico’s existing
infrastructure, experience, and skills, it is possible for us to create preservation solutions that
provide significant value to others in our community.

REFERENCES

Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at
Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at
Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at
Migration at Scale: A Case Study. Presented at iPRES 2012, Toronto, October 1-5, 2012. In iPRES
Moore, Kevin Ashley, and Seamus Ross (eds.), pp. 97-105. Available at
https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Pr
ocedings%20Final.pdf
[5] Owens, Evan, Cheruku, Vinay, Meyer, John, Morrissey, Sheila: Digital Content Management at Scale:
Presented at The Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at
and flexibility. Presented at The Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6,
2013. Available at http://ipres2013.ist.utl.pt/workshop5_presentations/Finding%20the%20balance-
iPRES2013.pdf
Digital Repository on Cloud Infrastructure: Issues & Challenges

Mukul K. Sinha
Expert Software Consultants Ltd., New Delhi, India
E-mail: mukulks@yahoo.co.in

Abstract—This paper first elaborates the challenges faced by the digital repositories for the long term preservation of their digital assets against various kinds of threats, and their vulnerabilities against perpetual technological obsolescence. It then expounds the needs, and consequently, the emergence of the concept of trusted digital repositories, to provide long term digital preservation services against a fee. Further, it elucidates the suitability of cloud infrastructure service for long term preservation of digital repositories, and then, examines few salient experimental approaches taken by different repositories for long term preservation of their digital assets, manifesting the future convergence of digital repositories and cloud infrastructure services. Finally, it visualizes that, in the near future, few cloud service providers would offer digital preservation as a service, to cater for the niche but promising market of trusted digital repositories.

INTRODUCTION

Advancement in digital technology compounded with internet has brought a qualitative impact on the way an individual, an organization, a company, or a government manages its documents. In most of the activities, documents in paper form are being progressively replaced by digital documents mainly due to latter's extremely attractive characteristics like its high storage density, ease of access and update, and convenience to operate on and manage, etc. But we now face a big paradox. While the eligibility of paper documents can be guaranteed for few centuries, a digital document that is stored on a digital medium may not survive even for a decade, and some time even shorter than that. The vulnerability of a digital document is of extreme kind: it can be lost without any trace, and once lost it cannot be recovered at all. In 90s, many organizations, without being aware of the vulnerability of digital media, lost their digital data completely, and once they realized, it was too late to recover, causing them huge amount of credibility as well as financial loss.
The vulnerability of digital documents are due to technological obsolescence, viz.:

1. Degradation and obsolescence of digital media on which they are written.

2. The proprietary format, owned by specific company which decides to withdraw from that segment of business, making it unreadable.

3. Its inaccessibility, as the software required to access it is either withdrawn, or it runs on specific hardware which has become obsolete, and is not being maintained, etc.

Thus, by mid-90s it was getting universally recognized that it is imperative for all organizations to preserve their digital documents for long term eligibility, i.e., ensuring long term maintenance of the byte stream, and continued accessibility of its content (called Long Term Digital Preservation), both for business needs as well as for statutory and legal requirements. In other words, all digital materials of an organization are, in essence, its digital assets that must be preserved to be accessed in distant future regardless of changes in technology, standards, and its user community.

In last twenty years the challenges of long term digital preservation are being thoroughly discussed, and being a live challenge different approaches are being experimented and evolved to handle the same. It has been realized that long term digital preservation tasks has two distinct aspects:

1. Handing various types of digital assets for long term preservation against various kinds of threats and risks and

2. Managing the data centre holding its digital repositories against continuous technological obsolescence of all types of their digital resources, viz., storage media, computer hardware, operating systems, databases, standards, and application software etc.

While for handling the former aspects needs with expertise of archival and information science, the latter aspects needs the expertise of computer science. As the main business needs are related with digital assets preservation, this paper explores whether the task related with digital repositories be off-loaded to some other type of service providers whose main business is to provide digital resources infrastructure, viz., the cloud service providers.

The Section 2 elaborates various challenges of long term digital preservation, how various digital repositories are tackling those challenges, and consequently, the evolution of the concepts of Open Archival Infrastructure System, and Trusted Digital Repositories. While the Section 3 explores the characteristics of storage service offered by cloud service providers and its suitability for digital repositories the Section 4 examines the few salient experimental approaches that are unraveling the possibility of future convergence of digital repositories and cloud infrastructure services. The Section 5 visualizes that in near future when the networks of Trusted Digital Repositories make their appearances some cloud service providers may emerge who would provide storage of archival information packages as a service (viz., Digital Preservation as a Service, i.e., DPaaS) offering primitives to facilitates preservation as well as audit activities.
Long Term Digital Preservation: New Challenges Due to Digital Media Vulnerability

Long Term Digital Preservation tasks comprises of bit preservation tasks and logical preservation tasks. Bit preservation implies ability to preserve the bits of digital record in face of media degradation, media obsolescence, destruction due intentional attacks or unintentional mistakes, or any natural catastrophe like fire, flood, earthquakes, etc. On the other hand, logical preservation implies preserving the readability and understandability of the content in long run against technological obsolescence of computer hardware, operating systems, storage formats, data management formats, or application software.

Bit preservation is managed by data replication, data refreshing, and data migration (for media obsolescence). In contrast, for logical preservation of digital assets different strategies would be followed for different kinds of technological obsolescence and it may result into bit mutation of the digital contents as well.

Archives and Digital Repositories

Archives

For an organization, the task of maintaining its digital asset is its own responsibility, which is managed either by its administrative or archive section that usually maintain multiple copies of their digital archives and backups in another site(s) having independent failure mode, viz., disaster recovery (DR) sites. This approach protects them from media degradation and destruction due intentional malicious attacks and unintentional mistakes. For protecting their archives from technological obsolescence of media, the archive section needs to be alert and should periodically migrate their digital assets from earlier storage media to new media that may have a life of at least next ten years.

In brief, bit preservation of digital assets is good enough for those organizations who have to maintain their digital assets for at most around a decade or so but completely inadequate if they have preserve it for far longer.

Digital Repositories

This new task of long term digital preservation requires logical preservation, in addition to bit preservation, of digital assets and also needs periodic audit for the integrity, authenticity, confidentiality and reliability of the same to assure the organization that its assets are completely secure. These activities require additional technical and administrative knowledge that may be beyond the normal and prevalent level of knowledge and experience of an organization’s in-house archival section. For an organization the task of digital preservation requires, not only to commit additional fund for periodic up gradation of its technological infrastructure deployed for the preservation of its digital asset but also to encompass associated systems, viz., policies, administrative procedures, and staff with appropriate professional education and training to manage the same from all aspects.
Open Archival Information System (OAIS) Reference Model

In a study conducted by [14] for those organizations who were trying to grapple with their long term digital preservation problems of their repository, it was discovered that 75% organizations felt that irreplaceable information would be lost unless their digital assets are not appropriately preserved, and 42% organizations felt that they completely lacked the technical and operational capabilities for the handling their digital preservation tasks.

Consequently, there emerged a broad consensus that for an organization, in general, it would be financially too costly and unaffordable to have in-house technical and organizational capability for long term preservation of their digital asset themselves. Thus a need was felt to outsource its long term digital preservation task to a third party, that can provide digital preservation services against a fee that its subscribing customers can afford, and on whom a subscriber can repose trust to preserve and manage its digital assets.

This need led to the emergence of the concepts of Open Archival Information System (OAIS) reference model [20] that was proposed for characterizing attributes and operations of digital repositories that assures organization for the long term digital preservation of their digital assets from all aspects.

According to OAIS reference model, to fulfill the commitment of the long term digital preservation to its customers, a digital repository must assure authenticity, integrity, confidentiality, and availability of the stored digital assets.

The basic unit of OAIS reference model is the Archival Information Package (AIP). An AIP is composed of zero or one Content Information and one or more Preservation Description Information (PDI) components. The Content Information is composed of the Content Data Object (the raw data to be preserved) and its associated Representative Information (RepInfo) where the RepInfo contains all information needed to make the raw data understandable to the user community. A PDI has additional metadata related to content’s logical preservation, viz., Reference (its persistent identifiers), Provenance (its origin, the chain of its custody, and the processing it has gone through), Context (reference to other related APIs), and Fixity (assures that the Content Information has not been altered).

The OAIS is a reference model, and it does not say how it should be implemented. It must be realized that for any digital asset to be logically preserved, its context, its provenance tracking, and fixity checking become core functional requirements that have no corresponding requirement in any other type information/ database applications.

Community Based Digital Repositories and Trusted Digital Repositories

Community based Digital Repositories

Realizing inadequacy of an individual organization to cope with the challenges of long term digital preservation many organizations possessing similar types of digital assets started to form community to evolve Community Based Digital Repositories to preserve their assets [7] [19]. As most of the national and cultural repositories and libraries are funded by the state or federal government community started to emerge based on the common funding agencies,
and they are most prevalent now [25] [18] [13]. Most of these community based digital repositories started with limited features, but they are continuously evolving and enhancing additional features periodically. They are usually maintaining their digital assets in Community Managed Distributed Data Centre’s.

**Trusted Digital Repositories**

In parallel, efforts were invested to characterize the attributes and operational characteristics of a specialized service organization that offers long term digital preservation service to its customer, viz., any digital asset possessing organization, and on whom its customers can repose trust as well. Thus evolved the concept of OAIS compliant Trusted Digital Repositories [26], those organizations that understand the potential threats and other risks related with long term digital preservation requirements of a range of digital materials of its customers, and equip themselves to convert them into manageable risks. The framework and attributes of TDR is elaborated in [17].

As there is always a finite possibility of organizational closure and/or financial failure of a Trusted Digital Repository in long run, and hence, it must be coupled with another Trusted Digital Repository which, in case of its failure or closure, would function as a trusted inheritor of its data. And hence, the Digital Preservation Infrastructure must be composed of a cooperative network of sufficient number of Trusted Digital Repositories.

Each Trusted Digital Repositories must go through a Process of Certification to create a climate of trust among its customers. Trustworthiness is the capacity of a system to operate in accordance with its objectives and specifications, i.e., it does what it claims to do. Later, the Criteria and Checklist for Audit and Certification of Trusted Digital Repositories [5] [27] has been formalized, and it has been recognized as ISO Standards as well [6].

**Digital Repository & Data Storage on Cloud**

Unlike a private digital repository that preserves data for its own organization, a Community Based Digital Repository or a Trusted Digital Repository that preserves data of its various subscribers for long term digital preservation must have a data centre with very large data storage capability. Further, for any subscriber the volume of data to be preserved steadily grows over time, and hence, the data storage capability of a repository also needs to grow over time.

The core business activity of a community/trusted digital repository is associated with the digital preservation activities, visualized in OAIS reference model, for all its customers. In order to fulfill the commitment of the digital preservation to its customers, a digital repository has to additionally manage its data storage centre.

Can a digital repository be relieved of the responsibility of managing the data centre so that it can concentrate on its core business activity, viz., the digital preservation activities? Can the responsibility of managing the data centre be offloaded to a cloud service provider? Can this offloading be economically viable?.
Cloud Computing Services

Cloud technology offers on demand network access to a shared pool of computing resources, (viz., storage, computing, and application services) that can be rapidly provisioned and released to its customers, with minimum management effort and service provider interaction [15]. The main advantages of computing as a service are ‘on-demand availability of resources through internet’, ‘on-the-fly rapid provisioning of computing resources, i.e., elasticity’, ‘measured services, i.e., pay as you go’, and ‘no up-front commitment by customer’, etc.

The offering of ‘storage as a service’ and ‘compute resources as a service’ by a cloud service provider presents a very attractive as well as cost-effective alternative to the traditional in-house data centre. This approach further relieves an organization of the task of managing data centre infrastructure, viz., data centre space, air conditioning, power supply, and associated human resources, etc. In addition, it converts them from an ‘owner of computing resources’ to a ‘customer of computing services’ and transforms their data centre expenditure from ‘capital expenditure’ to ‘revenue expenditure’.

In other words, a cloud computing service provider offloads the digital repositories the responsibility of not only managing their data assets professionally, it also relieves them from managing the associated infrastructure, and that also at a lower cost with no upfront commitment. Further, a cloud service provider offer its clients compute services of various configurations, without or with OS platform of customer choice, and storage services of various levels of reliability to choose from. It further permits a customer to scale up or down its hired resources on demand, relieving him either from over-provisioning or under-provisioning of resources.

Cloud Storage Services

The major activity of a digital repository is to transfer digital asset to/ from its repository to its customers. Consequently, the performance and the cost of managing the repository would depend on the volume of its storage, and the amount of data transfer from its repository. The offering of storage as a service by Amazon [1] is discussed here as a representative case.

Amazon offers storage as a service in two forms:
1. Simple Storage Service (S3)–for storing arbitrary size byte string through application program interface (API) with options of two levels of reliability, reduced reliability of four nines, and standard reliability of eleven nines, and
2. Elastic Block Storage (EBS) Service where a customer can mount its file system that can be accessed by a virtual machine instance running on compute service. The performance characteristics of these storage services have been experimented in [Rosenthal 12] and are numerated as followings:
   - The transfer speed between S3 and outside world is same as that of a local computer and its disk.
   - The transfer speed between compute service and S3 is far slower than that of local computer and its disk, or disk on LAN.
   - The transfer speed between compute service and EBS is same as that of a local computer and its disk.
In brief, the transfer speed with cloud storage service is either slower, or at the most as fast as that of local computer and its disk.

The charging model is based on two components:

- The volume of storage used by the customer, and the time for which it has been used.
- The number of times and the amount of data transferred to and from the storage service.

In short, it is better suited for applications that have less frequent access to storage service.

Further, some cloud service providers including Amazon have their data storage distributed across multiple sites in different continents, and they offer customers storage service at site(s) of their choice that they can dynamically verify.

Suitability of Cloud Storage Service for Digital Repositories

We compare the characteristics and needs of digital records of a digital repository with that offered by storage service of cloud service provider from following aspects:

- **Data integrity**: To assure data integrity basic checksum is provided by almost all service providers. Very few service providers additionally provide user run checksums as well [10].

- **Data confidentiality**: To assure data confidentiality many cloud storage service providers’ offers data encryption through pair of access key and secret key.

- **Data availability**: For most of the cloud service providers data stores are kept on online disks, and are accessible to users through internet by simple API command assuring good data availability.

- **Scalability**: On the fly rapid provisioning of resources and pay as you go are basic features of cloud service that is most attractive from economic aspect for digital repositories.

- **Data reliability/ fault tolerance**: Data reliability needs for long term digital preservation are supposed to be totally fault tolerant, even against disasters of all kinds. Most of the cloud storage service has very high data reliability (e.g., eleven nines for Standard S3 of Amazon) yet it falls short of preservation needs. A subscriber has to manage itself the fault tolerance of its data assets by storing them in multiple copies and synchronizing them periodically.

Off late a concept of cloud broker has emerged that facilitates customer to manage its assets on multiple sites, and is discussed in some detail in Section 4.

- **Data portability/ interoperability**: A digital repository would like to have its administrative independence, either for economic or operational reasons, to move its
data from one cloud storage service provider to another. Most of the cloud storage service providers do not offer interoperability, and a subscriber has to invest large amount of time and effort to do the same, if and when it chooses to migrate to another service provider.

Here again, cloud broker services can facilitate a customer in migrating its digital assets from one service provider to other, and is discussed in Section 4.

- **Data audit**: Any repository assuring long term digital preservation requires to be audited by a third party audit to confirm the authenticity of the digital object, and also that the object is undamaged. Full audit of all digital assets by reading it makes audit unaffordable. Presently, most the storage service providers do not offer any efficient method to conduct the same.

In future, some cloud service providers aiming to capture long term digital preservation market has to provide some primitives on the object meta-data [9] to facilitate third party auditing of the repository.

- **Operational performance**: As digital preservation is a latency tolerant application, the present data transfer speed from/to any cloud storage service would be acceptable for the same.

- **Operational cost**: As preservation data are mostly fixed in size, get never modified, and are accessed infrequently, the volume of data transferred across and the number of times transferred from/to repository would be very less. And hence, the data transfer component of operational cost would also be relatively of low volume.

So, the main operational cost depends upon the rate and the volume of data of digital repository stored on the cloud.

**Cost Effectiveness of a Community Digital Repository on Cloud: An Experimental Result**

The LOCKSS (Lots of Copies Keep Stuff Safe) [9] is a system that helps community of libraries to maintain a network of distributed digital repositories, each comprising of a PC server with large local disk storage, to collect and preserve contents of e-journals published on the Web. Each node, called LOCKSS box, continually audits its content and repairs any damage by coordinating with other nodes having the replicas. The LOCKSS system is presently covering more than 200 libraries.

In 2012, an experiment [8] was done to explore the possibility of LOCKSS boxes using cloud storage instead of their local disk storage, and assess their technical and economic cost effectiveness with respect to the cost of local disk storage, for long term preservation. The experiment was done on Amazon running LOCKSS box on Amazon Machine Instance and ingesting data on EBS (Elastic Block Storage), and periodic snapshots on S3 (Simple Storage Service).
Their conclusion indicates that rate of price drop in cost of storage dominates all other aspects. As the average rate of local disk is falling by 20% per year against 7% of Amazon S3, and by simulating the future cost the experiment clearly shows that for the long term preservation, the cloud storage is around five to six times costlier than that of local disk, and hence, cloud storage is not at all cost effective with respect to local disk. Conversely, if the non-hardware cost of maintaining a local LOCKSS box is around five to six times more than its hardware cost then only the option of cloud storage is attractive.

In their analysis, they have found that if periodic audit is required to be done on the entire repository, then the operational cost of data transfer shoots up. Consequently, the author has also suggested that a minor modification in Amazon’s HEAD command, for recomputing checksum of the digital content, can reduce the audit cost considerably, making cloud storage cost competitive.

In other words, the Amazon’s cloud storage service has yet not visualized the potential of digital repository market, and once they do realize the same, they would minor modifications in few of their primitives to make the cloud storage cost competitive.

**Evolution of Digital Preservation & Cloud Technology Towards Mutual Adaptability**

The digital repository community have realized that the needs for long term digital repository differs widely from that of traditional storage applications in terms of threats, characteristics, and operational and functional requirements. Consequently, they started to explore the distinguishing requirements for long term preservation, viz., the bit preservation as well as the logical preservation, for digital repositories.

**Storage System Architecture for Long Term Preservation of Digital Repositories**

For digital repositories, Baker et al [3] have enumerated following distinguishing operational requirements:

1. Needs of storage replication over multiple sites, having independent failure modes (as its digital assets should not get lost at all, irrespective of any kind of threats and operational or organizational failures, as their data life cycle may expand for ever).

2. Low per-site engineering cost, (each site can afford to be less reliable, slow and without any need of snapshot backups, as data is stored at multiple sites and accessed infrequently without any possibility of local hotspots), and

3. Capable of handling vast heterogeneity of storage systems (as repository slowly keeps scaling over time, technology, and vendors).
These operational features of long term digital preservation, mainly the bit preservation, of the data repository would help to design appropriate system architecture, most suited for long term preservation.

**Preservation-aware Storage System for Long Term Logical Digital Preservation**

For long term logical preservation of digital repositories the essential functional requirements are followings:

1. Fixity computation (after any alteration of AIP).
2. Migration of preserved data across storage devices (to protect digital content against storage technology obsolescence) and
3. Tracking and maintaining the provenance (for the authenticity of the preserved content).

As these functional requirements for logical preservation are compute intensive, the preservation application would need multiple number of data transfers to maintain long term preservability of a repository. This led to the evolution of the concept of preservation aware storage system [11] that provides built-in support for digital preservation.

Preservation-aware Storage System offloads preservation related functions associated with repositories from their preservation applications to itself, decreasing the number of data transfers (related with maintenance and audit of data assets), reducing the operational cost, and in turn, simplifying the preservation application. For logical preservation, different kinds of implementations have been reported in literature [12] [11]. Here we do discuss the latter for its rich features.

Preservation DataStore [22] have been developed as an infrastructure component of European Union CASPAR (Cultural, Artistic, and Scientific knowledge for Preservation, Access and Retrieval) project [4]. While ingesting Preservation DataStore (PDS) encapsulates raw data with its large amount of metadata (e.g., provenance, fixity, and RepInfos) forming the main AIP (and associated RepInfo AIPs), and then transforms the AIPs (the logical information objects) into physical storage objects, on which it can perform preservation related functions (e.g., fixity checking, transformation, provenance updates, media migration, etc.) within the storage itself. Thus, it offers a very efficient preservation aware storage system.

PDS has its own object storage device (OSD) where it stores the ingested AIPs. The OSD comprises of a (scalable) network of disks with secure object level access control. PDS also offers connectivity to existing filing system for accessing already stored archival data.

**Preservation Aware Digital Repository and Cloud Storage Service Broker**

Preservation DataStore (PDS) the community based repository that administers and maintains its own object storage device (OSD) to serve the whole community. As there are so many cloud service providers offering storage and compute resources at competing cost, to offload the OSD to one or more cloud storage looks a natural choice. With this approach PDS Cloud [23] is being developed for European Union ENSURE Project to preserve medical and financial data.
Compared to an organizational storage system a preservation aware digital repository like PDS would like to have storage system with better tolerance against organization failure, provision for multiple copies distributed across multiple sites, higher long term reliability, long term scalability with least amount of over provisioning, and with no vendor lock-in. In addition, the application is access latency tolerant as well. Considering all these technical and operational aspects, availing the storage services offered by various cloud vendors for offloading their OSD looks an attractive option for the PDS.

In PDS, user understandable digital content is encapsulated in AIPs while residing in OSD, which needs to be extracted before rendering, and encapsulated while ingesting. Therefore, this component of the depository application and the OSD has to be offloaded together. As cloud vendors do provide compute as well as storage service, the application component would run as a virtual appliance on cloud compute service and the OSD would be residing on the cloud store.

Accordingly, PDS Cloud has been designed as an intermediate layer that functions as a broker to interconnect user requests for preservation services on one end, and multiple cloud vendors on other. AIPS of OSD can be stored on multiple cloud platforms for sustainability, and also to avail relative cost advantages. The two components of the PDS Cloud is Preservation Engine and Multi-Cloud Service. Each customer of PDS Cloud, called tenant, has its logical information structured hierarchically into aggregation, docket and objects. Aggregation is the configuration profile of a tenant defining policy and managing data in physical storage of various cloud vendors. Docket is grouping of objects like a file directory. Objects reside in cloud container. Preservation Engine implements various preservation services to be operated on different AIPs stored as an object in different cloud store.

In short, PDS Cloud is a multi-cloud broker that provides preservation services to its customers and avail storage services from multiple cloud vendors.

Public Preservation Vendors with Cloud Storage

On the similar line of PDS Cloud some digital archival vendors have started limited offering of preservation services with back end storage on multiple public cloud vendor sites with independent failure modes assuring total fault tolerance. Notable among them are DuraCloud [10], and Preservica [24].

**DuraCloud**

DuraCloud takes storage services from Amazon, Rackspace, and SDSC Cloud storage service providers. It provides smooth scalability and rapid availability and offer high speed data streaming with additional charges. For data reliability it offers to store multiple copies at different cloud storage sites, and keep these copies automatically synchronized. For portability it offers customers data to migrate from one cloud store to another. For checking data integrity it offers several ways to verify checksums. It also provides health checkup service to determine checksums of all contents by recalculating at a specified time, and gives the report on the dashboard.
Preservica

Preservica asserts that data archiving and long term preservation is its niche. It provides storage scalability and offers high data integrity by creating multiple copies and storing at multiple cloud data centers that are cross-checked among themselves by comparing content checksum at customized intervals and by cyclic redundancy checks. User can store data using metadata and data tags. User has access to both data and the metadata.

As preservation activity it does an automatic background integrity check and recovering the corrupted and missing file. It offers a set of tools for data migration from obsolete formats to newer formats.

These efforts indicate that that slowly the broker market is covering all the data preservation requirements of digital repositories through multi-cloud storage services.

TRUSTED DIGITAL REPOSITORIES AND DIGITAL PRESERVATION-AS-A-SERVICE (DPaaS) BY CLOUD

As few certifying agencies to certify trusted digital repositories [2] have been recognized many archiving organizations would now try to get them certified as trusted digital repositories. To maintain trust of their customers, a certified trusted digital repository is needed to go through periodic audit by a third party auditor.

Further, to protect customers against organization failure/ closure of a trusted digital repository it needs to be part of a digital infrastructure composed of a network of trusted digital repositories so that one or more trusted repositories can function as a trusted inheritor of its data.

Needs of a Trusted Digital Repository

Need of audit is and essential feature of a trusted digital repository as this is the main basis for its customers to repose faith on them. Audit of a repository is very compute intensive and data transfer intensive operation. Auditing a large volume repository by accessing every record is neither time wise or cost by feasible. Various suggestions have been made in research articles [16] [21] [9] about what additional features a cloud service provider should offer so that auditing would be cost effective.

Presently, digital archive vendors like PDS Cloud, DuraCloud, Preservica, etc. does provide preservation tasks to very much extent, and have multiple cloud vendors in back-end for storing multiple copies. It does protect a customer from organizational failure of any of the cloud service providers but what happens if these preservation vendors themselves are threatened with organizational failure/ closure?.

Cloud Service Provider and Digital Preservation as a Service (DPaaS)

There are three service model for cloud service providers to offer, viz., Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Storage as a service, and Compute as a service is part of IaaS. In the present cloud service market, apart
from major players who are mainly concentrating in Infrastructure and Storage as a service, there are large section of cloud vendors that are catering to niche but generic enough market, by offering specialized services through Software as a Service. Rackspace was the first to enter such a specialized niche market segment.

I strongly feel, that digital preservation is a large enough niche market which has few very specific needs, mainly catering to long term logical preservation activities related with OAIS compliant Archival Information Packages (AIPs), and third party audit activities that cannot be solved in a cost effective way existing archival service vendors.

And hence, there is a strong need that few cloud service providers should offer Digital Preservation as a Service (DPaaS) for storing AIPs, and offer a set of functionalities that should facilitate periodic third party audit in a cost effective way.

REFERENCES

Digital Preservation of Scientific Data
Exa-scale Data Preservation in High Energy Physics

J.D. Shiers
CERN, Geneva, Switzerland
E-mail: jamie.Shiers@cern.ch

Abstract—This paper describes work in progress in the High Energy Physics (HEP) community to provide sustainable services and solutions for the preservation of data for long-term (re-)use. The key issues that differentiate HEP from other communities will be explained, along with services and technologies from HEP that we believe can be applied equally well to other disciplines, either as part of a tool-kit or else as an e-infrastructure service. A small number of high level Use Cases will be outlined, along with their associated Business Cases. These Business Cases will be compared to known and foreseen costs and cost models, building a clear case for the preservation of HEP data, together with the full environment for the re-use of this data, for a number of decades. Beyond this period, digital preservation—i.e. the bits—can be foreseen for a much longer duration at reasonable cost. The motivation for maintaining the full ability to re-use this data versus just the data will be used to perform a cost-benefit analysis. It is well known that long-term data preservation is a journey—not a destination. We will summarize by concluding that it is an affordable journey, provided that one sets a few reasonable boundary conditions (just as in life itself).

Keywords: Exabyte, Bit Preservation, Use Cases, Business Cases, Costs and Cost Models

OVERVIEW

The High Energy Physics community has been building and operating particle accelerators and colliders, together with increasingly complex and large detectors for a period that overlaps more or less completely with the digital age. Whilst computing has always been an integral part of the analysis of the acquired data, non-digital recording was used for many experiments (film) for at least a fraction of experiments for several decades. This era is now over with “born digital” data dominating since many years. The total volume of digital HEP data now exceeds 100PB and is expected to increase to the EB scale within the next 15 years. This bulk of this data is, or will be, from CERN’s Large Hadron Collider (LHC), which is expected to run, with a number of significant upgrades, until around 2030.

Although there is valuable data from previous colliders, much of the discussion below will refer to the LHC experiments, given that they are still in the relatively early phase of data taking and analysis. We will none the less consider how (potential) solutions to the preservation of LHC data can be applied to the data from these other experiments. A strong
expression of thanks is due to these recently “terminated” experiments [1], which have done much to raise the awareness of the need for data preservation in HEP in recent years and without whom this current work would probably not be carried out.

THE LARGE HADRON COLLIDER

In the 1980s, following discussions in the previous decade, CERN constructed the Large Electron-Positron Collider (LEP) in a 27km long tunnel some 100m under the French-Swiss border. During the early days of the LEP discussions, it was proposed that the tunnel be large enough to additionally house a hadron collider, in the event that the LEP machine be approved. This hadron collider the LHC as it is now known became operational about 30 years after these initial discussions, with the collaborations (also know as “experiments” within the HEP community) being approved in the early 1990s. One of the key scientific goals of the LHC was the confirmation or not of a boson matching the predictions of Higgs et al. This has now been confirmed, with a Nobel Prize being bestowed on Englert and Higgs for their theoretical work in this area.

It is hard to put a value on such a discovery although this is likely to become clearer in the long term but the “cost” can be calculated relatively easily. This includes the cost of construction of the machine, the detectors, the computing infrastructure and the long-term operation of all of these.

The duration of a project such as the LHC is about the same as the whole history of HEP to this date. This is clearly much shorter than other disciplines such as astronomy from which it is also differentiated in that HEP works with “factories” that generates large number of events, the analysis of which is statistical, as opposed to unique, unrepeatable observations in the former case.

KEY USE AND BUSINESS CASES

In the scenario of the LHC, but equally applicable to previous HEP programs, a number of key Use Cases can be identified.

These Use Cases cited are high-level and, together with their associated Business Cases, are the primary motivators for attempting long-term data preservation for the HEP community.

In summary, the motivations for preserving the data are exactly those that led to its being acquired, given a strong suggestion that the same funding agencies could or should be targeted to ensure this preservation.

Other Use Cases, such as the preservation for an indefinite period for unknown future reuse, are much less clear.

It is also important to note that at least some host laboratories may change their core business over such a period and so alternative partners for providing some of the key infrastructure, such as one of the data repositories, may be needed.
Table 1

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Business Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Continued Ability to Perform Analysis 10-15 years after end of data taking.</td>
<td>Experience shows that a significant number of publications and conference presentations are made in the 5-10 years following the end of data taking. However, this period also sees a significant drop in or end of funding to directly support the experiment(s) in question. The business case for continued funding is to ensure the maximum scientific potential of the experiment(s) / facility. Lack of funding may result in a “loss” of some 20% of the potential output.</td>
</tr>
<tr>
<td>2 Ability to Reanalyze past data in the light of new theoretical models / in-sights.</td>
<td>Improved and / or new theoretical models can have a major impact on the interpretation of data. Re-analysis of past data using such new insights has, in the past, led to significantly improved results. As opposed to Use Case 1, the time scale involved is much less clear: there is no guarantee that new models will appear in 10, 50 or even 100 years. On the other hand, if the data–including the full capability to reanalyze them are not preserved, then the “outcome” is assured.</td>
</tr>
<tr>
<td>3 Ability to Compare Results from a Future Facility with those of a Current / Past one, including full re-analysis if the new results justify it.</td>
<td>Machines (accelerators, colliders) used in HEP often follow a “discovery machine” followed by “precision machine” pattern. There is strong scientific motivation to retain the data from the “discovery era” to the “precision era” to perform comparisons and, if necessary, new analyses of the old data. The duration of the preservation period is typically known, although potential delays in funding / construction / commissioning of the new facilities needs to be taken into account. A “successor” machine to the LHC may be operational in the 2030s, so the curation period is a factor longer than Use Case 1.</td>
</tr>
</tbody>
</table>

BIT PRESERVATION AND DIGITAL LIBRARIES

Two of the key areas in a long-term sustainable data preservation strategy are the preservation of the bits themselves, plus a repository of all the associated scientific papers, write-ups, documents, descriptions and so forth. (A service to store the data behind the figures in these papers has been provided for many decades via the HEPData project. These data are also actively to validate simulation packages of physics processes by the corresponding authors). The LHC program requires us to have a plan for digital preservation that is valid for several decades: this means that the typical operations of media scrubbing, faulty media replacement, media replacement with newer, higher capacity volumes and so forth must all be implemented in a cost effective fashion as part of routine operations. Furthermore, a budget plan must and does exist for this “service”. The “archive” of the LHC data is based on “enterprise” tapes with a re-placement cycle of about 3 years. Thus, a clear plan can be established with a horizon of about one decade: this is then periodically updated so that we always maintain a fairly accurate budget for both personnel and materials (i.e. tape drives, media, infrastructure such as servers and net-works). This experience shows that the typical cost estimate, assuming constant “media” costs but halving in price with each generation, is in fact a lower limit: one must include the on-going operations that are required to ensure that the data remains intact and is migrated forward to newer devices. However, the manpower involved is relatively modest and roughly independent of the total data volume–about one full time equivalent per archive instance. (That is, maintaining two data copies at two independently managed sites would have twice the manpower costs, but with significantly higher protection against data loss, than just one).
Digital library tools and solutions are relatively well established. The HEP community is now largely based on the INSPIRE service, built across four institutes: CERN and DESY in Europe together with FNAL and SLAC in the US. The software behind this service is available under an Open Source license agreement and a “freemium” service is also offered.

Thus, we consider that two significant pillars of long-term data preservation are in hand, in the sense that there are clearly identified service offerings with well understood costs and a predictable lifetime measured in decades. Changes are bound to occur but there will be migration tools and assistance, as is today offered between INSPIRE and the previous SPIRES solution, as well as coordinated support to migrate to a different archive solution and/or to changes in access protocols and methods.

E-INFRASTRUCTURES, METADATA AND SOFTWARE

As is well understood, preserving the digital information and/or documentation and scientific papers is only part of a much more complicated story. In general, two approaches are considered for handling “the rest”. These are typically referred to as emulation and migration. Many people feel that emulation is only viable over relatively short periods not for the decades that we consider for HEP data and certainly not on “astronomical” timescales. On the other hand, migration can be very labour intensive, particularly when disruptive changes are faced, and in any case each migration requires in-depth verification.

To address the latter point, many of the HEP experiments have adopted semi-automated validation systems. This is a clear target for optimization, both within our community, where a common tool-kit or even service could be offered to multiple experiments, backed up by a coordinated support team, as well as beyond. The general concepts are discipline independent and this is felt to be a potentially interesting area for inter-disciplinary common projects, leading again to services or at least tool-kits.

As far as emulation is concerned—at least if one interprets it as virtualization most people agree that it is at least part of the solution, particularly in today’s cloud-oriented world. Whether it can be more is subject to debate: recent success in applying virtualization to data and the full software environment of a LEP experiment is encouraging and at least one of the LHC experiments is considering it as a key part of their overall data preservation strategy. These activities are more then proof-of-concept/principle: they can be evaluated “in the wild” over increasing periods of time in parallel to “migration” strategies and the pros and cons evaluated. The long lifetime of the LHC experiments will give significant experience into these perhaps complementary and provide input to future work.

A recent paper by members of the CMS collaboration one of the two large collaborations working at the LHC estimates the offline software to consist of several million lines of code with around 400 active developers. Integrating over all 4 main LHC experiments and including “general purpose” simulation and analysis software, the scale is over 1000 developers and at least 10M lines of code that is highly complex both algorithmically and in terms of dependencies. It is perhaps not unreasonable to claim that even the original authors struggle to understand the code after a period of a few years! (This is certainly true for this author, even in the much less complex times of LEP).
COSTS OF CURATION

Another equally crucial aspect of long-term preservation is a good understanding of the costs that can be used, together with the Business Cases, to build a sustainable long-term budget plan. At the time of writing, the HEP community has not completed a full analysis of all aspects of data preservation: this has been done for “bit preservation” and will be extended in 2014, via a series of workshops [2] and associated publications, to cover all areas of long-term data preservation.

Preliminary calculations suggest, however, that the annual costs are rather modest in terms of the Use and Business Cases outlined above and that they are manageable at least for one decade, if not two, three or possibly more. It has also been demonstrated within our community that should a strong scientific, i.e. business, case exist then even “resurrection” of old data can be achieved with affordable effort.

Concrete numbers to back the cost-related claims in this paper will be available in early 2014.

Experience papers on the more complex aspects of “knowledge preservation” will be produced throughout the active lifetime of LHC data taking and analysis, with significant advances expected during the remainder of this decade.

CONCLUSION

Bit preservation at the scale of 100PB–1EB, together with the associated documentation and other “library” information, is both affordable and in-hand. Now is the time to start to offer discipline-agnostic services, e.g. on a cost recovery basis.

“Knowledge preservation” is possible over a period of several decades with costs that are “only” a factor—or, in the worst case an order of magnitude more than that of the bits themselves (at the above mentioned scale and in the case of HEP).

Over longer terms, the cost of this knowledge preservation may not be justified as an ongoing task. However, all the necessary steps should be taken to facilitate “a resurrection”, including copious and redundant validation data, should such a requirement arise, even in the far distant future.

REFERENCES

[2] The first such workshop is scheduled for January 2014. More information can be found at https://indico.cern.ch/conferenceDisplay.py?confId=276820
Digital Preservation of Remote Sensing Data

V.V.S. Nageswara Rao¹, K.V. Rathna Kumar², Vinod Bothale³, M.V. Ravi Kumar⁴ and S. Muralikrishnan⁵
¹,²,³,⁴,⁵National Remote Sensing Centre, Dept. of Space, Balanagar, Hyderabad–500037, India
E-mail: ¹nageswararao_vvs@nrsc.gov.in

Abstract—National Remote Sensing Centre (NRSC) plays a major role in acquiring, preserving, processing and dissemination of Satellite and Aerial Remote sensing data over the past three decades. NRSC has been acquiring valuable remote sensing data since 1978 on various media namely High Density Digital tapes (HDDT), Digital linear tapes (DLT) and at present on Hierarchical Storage Management (HSM) which involves SAN storage. There is also a need to archive processed data information on natural resources to study the long term impact of changes on the earth surface, climate change etc; Web Portals like SIS DP, WRIS and NDEM will cater to such needs. Bhuvan is a visualization portal which showcases several such information on a canvas of Indian Remote Sensing satellites data acquired across the globe. The paper describes technological issues in preservation of remote sensing data, archival policy and current practices at NRSC.

INTRODUCTION

Earth Observation (EO) missions enable global coverage on a repetitive temporal and spatial interval with a variety of sensors. This data is vital for study of earth as a system. To study the long term impacts of changes that are taking place on the earth surface by man made and natural phenomena, it is essential to preserve the valuable data that is acquired by these Earth Observation (EO) missions and by Aerial Remote sensing. Hence data archival assumes a very important role in the operations of every space agency.

NRSC has been acquiring and preserving remote sensing data over the past three decades from several Earth Observation missions starting from IRS-1A and up to recently launched Radar Imaging satellite. A part from the need to archive raw data from these EO missions, there is also a need to archive processed data information on natural resources etc; to study the long term impact of changes on the earth surface, climate change etc; NRSC has developed application specific web Portals like Space based Information system for decentralized planning (SIS DP), Water Resources Information System (WRIS) and National Disaster Emergency Management (NDEM) will cater to such needs. Bhuvan–Gateway to
Indian Earth Observation, explores, experience, visualize and analyze Indian Remote Sensing (IRS) images. It is an NRSC open data archive (NOEDA) to browse and download satellite data.

**ARCHIVAL & STORAGE OF REMOTE SENSING DATA**

NRSC is responsible for data archival, storage and data base maintenance of data archived from all the Indian and foreign Remote Sensing satellites in addition to Aerial Remote sensing. NRSC has been acquiring valuable remote sensing data since 1978 on various media namely High Density Digital tapes (HDDT), Digital Linear Tape (DLT) and at present on SAN storage. In addition to regular archival and retrieval, the data is being transcribed on to new media periodically, for long term preservation.

Data from the scientific missions like Chandrayaan-1, MeghaTrophiques was archived and stored in centralized SAN storage at Indian Space Science Data Center (ISSDC).

**OPERATIONAL USE OF REMOTE SENSING DATA**

The operational use of remote sensing data span wide spectrum of themes which include water resources, agriculture, soil and land degradation, mineral exploration, ground water targeting, coastal and ocean resources monitoring, environment, ecology and forest and urban areas mapping. The satellite data processing facility is equipped with necessary infrastructure supported by trained manpower to catalog, process, qualify, archive and disseminate the satellite data products based on user needs globally.

**ARCHIVAL SIZE**

The Remote sensing satellite data which is archived is in digital format. At present the archival facility at NRSC contains 1.5 Peta Bytes of data of all the missions on 3-tier SAN storage. The data volumes increase as ISRO launches 2 remote sensing satellites every year. In the coming five years the volume of data expected for archival will touch 2.0 Peta Bytes. Apart from this the Aerial Remote sensing data volume is 100 Terra Bytes.

In addition to raw data archival, 150 Terra Bytes of processed Remote sensing data is available on Application specific user registered web portals like SIS DP, WRIS and NDEM are used to study long time impact of changes on earth surfaces. Bhuvan is a Geo portal of ISRO hosted in 2009, to evince the Indian Earth Observation capabilities from the Indian Remote Sensing Satellites. It is an interactive versatile Earth-Browser which showcases multi sensor, multi platform and multi temporal images has 20 TB of data on SAN storage. Web portal: http://www.bhuvan.nrsc.gov.in.

The Indian Space Science Data Center (ISSDC) at Bangalore has a separate SAN Archival facility for acquiring data from scientific missions like Chandrayaan-1 and MeghaTrophiques etc; the 20TB scientific data is globally accessible by several international registered users including NASA for scientific studies.
DATA FORMAT & STORAGE MEDIA

The digital raw data after reception is archived in SAN storage in FRED format (ISROs proprietary format). The data from the SAN is further used for higher levels of processing before dissemination to user community.

STORAGE REQUIREMENTS, MANAGEMENT & BACKUP ARCHIVAL

Raw data acquired in SAN Archival facility is maintained in a controlled environment ie. 18 to 22 deg Centigrade temperature and RH 45%. To safeguard the archives from external factors (floods, fire and disasters) a mirror site was established by keeping a second copy of all the archived data at a distant location.

The data is managed using an oracle data base and centrally controlled by Integrated Information Management System (IIMS) for utilizing of archived ancillary data for products generation and dissemination to user community.

DATA SECURITY

Safeguarding the archives from the external factors (flood, fire, disasters in general) is ensured through local risk mitigation infrastructure. Access to raw data for data product generation is handled with a request duly signed from the system manager. Online accessing and delivery services are in place for archived data/ products to the registered users. Biometric access to the archival facility to avoid unauthorized intrusion.

ARCHIVAL POLICY

Dept. Of Space (DOS) has evolved a comprehensive data archival policy which governs long term preservation of earth observation data. The entire raw data from all the operational remote sensing missions to be acquired and preserved irrespective of cloud and sea area. The meta information of the entire data archived is made available on the NRSC web site in sub sampled browse data to enable the user community for browsing and selection for ordering.

QUALITY ASSURANCE PROCESS

At NRSC Archival processes like data screening, data base updating, codification, data preservation is maintained in a controlled environment and user requirement is governed by ISO 9001:2008 standards on data/ information security. Internal and external audit is conducted regularly to monitor the performance. With these processes Earth Observation data product quality will be ensured for the product life time.

TECHNOLOGICAL ISSUES IN PRESERVATION OF DATA

Since 1978 NRSC has been archiving data on High Density Digital Tapes (HDDT). These HDDTs had a shelf life of 10 years under ideal conditions. But due to wear and tear of tapes and recorder heads the retrieval of data has become difficult after few years. Data losses were noticed while playing back the data, after five years of shelf life. This is primarily due to tape
stickiness and oxide shedding which was observed on the tapes in spite of keeping in a controlled environment. Specially fabricated ovens were used to bake the sticky HDDTs for data retrieval.

One needs to explore more effective way of preserving the archived data in media with better capacity which occupies less physical space and provides scope for archiving data from future missions. In 1997, NRSC migrated to new technology /media Digital linear Tape (DLT), which has got 30 years of shelf life. The entire archived data was transcribed on to DLTs to over come the problems of HDDT.

CURRENT PRACTICES AT NRSC

The amount of data acquired daily is increasing with the increasing number of Earth Observation missions and higher data rates because of higher spatial and spectral resolutions. The latency between data acquisition and dissemination to users is also decreasing because of use of data for dynamic modeling and other requirements. Hence at present NRSC under IMGEOS project (Integrated Multi Mission Ground Segment for Earth Observation Systems) the data is archived in a Hierarchical Storage Management (HSM) which involves a SAN (Storage Area Network) that consists of high performance FC disks, medium performance SATA storage and a third tier of tape storage using LTO-4 (Linear Tape Open) tapes was operational since 2012. The total storage includes tape library is 3.0 Peta bytes scalable to 10 Peta bytes. The data acquired on DLT media is transcribed on to SAN storage. Tape library will also generate backup tapes for online redundancy and data availability as well as vaulted tapes for meeting disaster recovery requirements.

ACKNOWLEDGEMENT

We are grateful to Director, NRSC for his continued support during migration and transcribing of the data to SAN storage and kindly agreed for writing a paper in this international conference on Digital preservation. We also thank Deputy Director (Training) and Head, Training Division for his constant encouragement and support. We convey our sincere thanks to all the Area Group In-charges in Data Processing, Bhuvan and Aerial Remote Sensing for timely providing inputs for this paper.

CONCLUSION

NRSC has experienced several technological issues in Remote Sensing data archival and preservation in different kinds of media namely HDDTs, DLTs and recently migrated to Hierarchical Storage Management (HSM) which involves SAN storage. The storage capacity of the LTO tape library is added @ 1.5 Peta bytes per year to cater future archival needs. The Archival of processed data in the application web portals SIS DP, WRIS and Bhuvan looking to the next decade as the demands placed to study the long term impact of changes on the earth surface, climate and atmospheric change and to better meet the needs of users.
REFERENCES

[1] Satellite Data Acquisition and Archival policy Document No. NRSA/ DPG/Archival/ Ver 4.0/ May, 98
The Purpose Behind Repurposing [HISTORIC] Data: A Case Study in Ozone

Elizabeth Griffin
Dominion Astrophysical Observatory (DAO), Victoria, Canada
E-mail: elizabeth.Griffin@nrc-cnrc.gc.ca

Abstract—Today’s plans for digital archives and repositories take for granted the need to preserve everything that is available electronically. Within those activities, objectives and goals are needed to ensure adherence to appropriate standards, and I argue that prioritizing the tasks must consider applications of the products as well, especially as concerns Planet Earth. Some data of acute value to the natural sciences are not even in digital form, and the case study described here identifies a scientific driver that demands attention to the whole life-history of a discipline’s data, not just its more recent electronic ones.

The natural world can undergo significant changes with time-scales of the order of decades or longer. Modeling them correctly requires data from throughout those spans, and therefore involves mining data from the pre-electronic era. Unfortunately, many of those early analogue data have never been converted into digital formats, and in their original states they are inaccessible to the very research programmes that urgently need them. Nevertheless, occasionally data suitable to meet some of those needs can be found in unanticipated places, perhaps within the realms of a quite different science or activity. The study of stratospheric ozone summarized below provides an excellent example of all of these features:

a. Locating relevant historical scientific data.
b. Converting quasi-inaccessible photographic [analogue] observations into electronic format, (c) a transdisciplinary exchange of data, and
c. Repurposing scientific data for an application that was not even dreamed of at the time the original observations were made.

Keywords: Digital Preservation, Trustworthy Digital Repository, Audit, Certification

The Earth's Ozone: Facts and Possible Fallacies

The presence of ozone (O₃) in the Earth’s atmosphere affects us all. The low stratosphere contains about 90% of the atmospheric O₃ concentration; the rest occurs in the low troposphere. The presence of O₃ in the stratosphere is vital to every biological organism through its effectiveness in blocking the harmful UV-C and UV-B radiation from the Sun, and its reduction would have such dire consequences that the threatened possibility energizes
international campaigns, influences the thinking of governments, and activates substantial alterations in the accustomed habits of millions throughout the world. However, the properties of \( O_3 \) are not generally that well known, while the actual facts concerning its concentration (and in particular its recent reduction) require sophisticated scientific measurements that are only accessible to the few and are limited by data that have not always been as plentiful as they are nowadays.

Concentrations of \( O_3 \) at a given location can today be measured both from the ground and from space, using techniques that determine respectively the absorption or the backscattering caused by the presence of the molecule. Measurements of its height-profile, carried out by sondings aboard a weather balloon and including the air temperature, pressure and relative humidity, date back to the early 1970s, while satellite records date back nearly as long. The longest time-base of ground-based measurements actually spans nearly 90 years, but uncertainties in the early data, coupled with the high sensitivity of \( O_3 \) to local meteorological conditions, have driven the need to seek alternative data-sets, especially ones employing independent technologies.

Tropospheric ozone is mostly produced by the action of sunlight on CO, oxides of nitrogen and volatile organic compounds in car exhaust and industrial chimneys, by waves pounding on the shore, or by lightning, and can be a harmful irritant if inhaled.

The evidence before us is that the benign stratospheric ozone has been in decline. A fall of only a few percent allows more, and shorter-wavelength, solar UV radiation to penetrate to Earth, affecting the living world in numerous ways. If the decline is prolonged, plants and ecosystems will modify their habits and their habitats (with obvious consequences for dependent fauna and our food chains), plant characteristics will alter, species of all kinds will suffer deleterious effects (skin cancers are only one such), while plastics in building materials will degrade and require expensive replacements. If human interference, by releasing into the atmosphere CFCs and nitrous oxides which diffuse upwards and catalytically destroy \( O_3 \), is the sole culprit for the decline then it is within human power to reverse the situation. However, while that explanation is the sort of cause-and-effect announcement which is popular with the media, the evidence is not quite as well established as we have been led to believe. That is not to say that the simplistic conclusion is wrong, merely that the evidence is not yet incontrovertibly proven.

**History and Development**

Ozone was studied in the laboratory before it was identified in the atmosphere. The strongest, most absorbent contribution to the Earth’s ozone shield is named after W.N. Hartley, who first analysed the gas in 1871. From an observation made in 1890 Professor and Mrs. Huggins drew attention to diffuse absorption bands in the UV spectrum of Sirius, a star too hot for molecules to exist in its atmosphere, and by about 1920 those same bands—since named the Huggins bands—were clearly detected in the solar spectrum, where they intensify considerably as the sun sets and then to be identiﬁed with weak features on the redward ﬂank of the Hartley band. Figure 1, a laboratory spectrum of ozone, shows the relationship between the two sets of bands. Figure 2 illustrates how the Huggins bands appear in the spectrum of a hot star such
as Sirius, which is always viewed low in the sky from northern sites. Absorption by ozone constitutes a background envelope for the stellar spectrum; in this plot the absorption by the Hartley band has been removed and the Huggins continuum has been normalized. Ozone research commenced in the 1920s, in Oxford (UK), under the leadership of Professor G.M.B. Dobson. Strongly motivated by curiosity, Dobson developed equipment and techniques under conditions that would be unthinkable nowadays. His ‘laboratory’—a mere hut on a small hill on the outskirts of Oxford—was innocent of electricity, so his group carted accumulators there on their bicycles. In order to make photographic observations he was obliged to develop photographic methods. Yet in that laboratory Dobson built the first spectrometer to measure the intensity of the ozone via the Huggins bands as seen in the solar spectrum; it observed the Sun at different altitudes, and compared the intensity of sunlight within a spectrum aperture that included an O₃ band, to its intensity within a reference aperture placed well away from the band.

![Image](image1.png)

**Fig. 1:** Laboratory Emission Profile of Ozone, Showing the Relatively Weak Huggins Bands between ~\(\lambda\) 3140-3400Å, on the Red Flank of the Dominant Hartley Band

![Image](image2.png)

**Fig. 2:** The Huggins Ozone Bands, Adding their Absorption to the Spectrum of Sirius (T \~\ 10,000 K), are much Broader than the Sharp Features which Originate in the Atmosphere of the Star. The Smooth Line Represents the Laboratory Spectrum of Ozone; the Hartley Absorption has been Removed, so the Huggins Bands Effectively form the Continuum against which the Stellar Spectrum has been Recorded. Note the Expanded Vertical Scale

The method was not without its difficulties, the chief being contamination of the readings by scattered sunlight and wavelength instabilities. The UV solar spectrum is very rich in Fraunhofer lines, and a wavelength drift in the equipment will unintentionally admit or exclude features at the aperture margins, vitiating the measurements by causing (respectively)
less or more light to be passed. Nevertheless, Dobson’s early work sparked considerable interest, and since it was already shown that measurements were substantially affected by local weather conditions he built replicas of his ‘master’ instrument and set them up at different sites in Europe. One site, at Arosa (Switzerland), at an altitude of 1850 m, commenced measurements in 1926 and is still operational today, constituting easily the longest continuous set on record. About 250-300 measurements have usually been made annually. O₃ concentrations are expressed in Dobson units (DU), or the equivalent thickness of a layer of pure O₃; a characteristic value of 300 DU would occupy a layer 3 mm thick.

The spectrometer at Arosa was upgraded from time to time and the photographic camera was eventually replaced with a photomultiplier. However, the need to allow adequate overlap between an instrument and its replacement was perhaps not always adequate, and because the early measurements were also more prone to noise and to lower accuracy than later ones the older values seem less trustworthy than modern ones. Each overseas spectrometer was routinely returned to Oxford for re-calibration against the ‘master’ in order to maintain uniform quality across all the European sites, particularly since those other measurements were aligned to the Arosa ones and used to complement them. Very few measurements were available in those decades from sites further afield, least of all in the southern hemisphere.

THE PROBLEM, AND ASTRONOMY’S UNIQUE CONTRIBUTION

The dominant argument today concerning the decline of ozone concentrations, particularly over Antarctic regions, is summarized in Figure 3, which shows annual Arosa means since 1926. The downward trend since the mid-1970s is undisputed, and a possible upward trend since the late 1990s may be real. What is unclear from that plot (and it is the only such plot that we have) is what was happening in those early days: was the concentration of O₃ effectively constant (as seems to be unquestioningly accepted) until acted upon by anthropogenic interference, or do the data suggest a component of slow natural variability? The lower quality of those early measurements calls for complementary databases, preferably using independent technologies and certainly at different geographic locations, to augment our evidence concerning the evolution of this vital component of our atmosphere.

Fig. 3: Ozone Concentrations, Shown as Annual Means from the Arosa Database as Archived by the World Ozone and UV Data Centre in Canada. The Downwards Trend Commencing in the 1970s is Unambiguous, and a Recent Recovery may be in Evidence, but it is Less Clear.
One independent resource that has been almost completely neglected to date is astronomy’s substantial heritage of archived stellar spectra. Every celestial observation made from the ground unavoidably records signatures of the atmosphere’s constituents, and it may seem surprising that this unique and valuable resource has not previously been scoured for that purpose until one realizes that those early records are photographic, and are therefore inherently non-digital. During the 1980s observational astronomy switched to the CCD as its workhorse detector, and very few observatories retained the tools and the expertise to access the scientific information in their more historic photographic spectra, only a very small percentage of which had ever been digitized and archived. Those historic data cannot therefore be accessed in a form that is usable by modern analysis without specialist handling. The investigation of photographic astronomical spectra for ozone research which I commenced in 2003 was the first of its kind, and required some essential track-laying before I could proceed towards determining actual concentrations of ozone.

**REPURPOSING ASTRONOMY’S HERITAGE**

My first challenge was to locate suitable observations. Regular global monitoring of ozone did not burgeon until the 1960s, but some 25–30 stellar observatories were operating worldwide well before that time, each carefully archiving its photographic plates in-house. However, whether those plates show meaningful and useable signatures of ozone depends upon the original purpose(s) of the observation (i.e. on the wavelength region recorded), on the properties of the spectrograph (transmitting optics, if any, needed to be of quartz since glass is opaque below ~\(\lambda 3500 \text{Å}\)), and on the type of star observed: the Huggins bands are best measured on spectra of warm or hot stars like Sirius, whose spectral lines are relatively few (Fig. 2).

However, no observatory has an on-line inventory of its photographic contents, and I had no alternative but to visit each one in person and search manually through shelves and cabinets. The most appropriate digitizer for photographic spectra is the ‘PDS’, a purpose-built instrument which a number of observatories once owned but then discarded when digital detectors replaced photography. In 2003 there was only one fully operational PDS in Canada (at the DAO in Victoria), so I carried there my borrowed spectra to digitize and analyse.

The next challenge was to devise a method of analysis that would be reliable, flexible, test-able and reasonably rapid, and would succeed with data that were never intended or even envisaged for such work. As a ‘proof-of-concept’ I selected some high-quality photographic spectra of Vega and Sirius that were exposed for the UV recently enough (1978–1992) to overlap the Total Ozone Mapping Satellite, TOMS, since I then had an opportunity to compare my results with ones from a different, purpose-built technology. In stellar spectroscopy all measurements are relative, and require some laboratory quantities to render them absolute. Those I therefore created by measuring each Huggins-band strength on a high-quality laboratory spectrum. Measurements of individual \(O_3\) bands in the stellar spectra could then be converted into the ‘column height’ of \(O_3\), the total amount of \(O_3\) along a sight-line of unit width, normalized to the zenith. Calculating the mean angle between the star and the zenith required knowing the date, time and duration of each stellar exposure-information that entailed more digging into hand-written log-books.
The results were then compared with measurements by TOMS, which maintained a geostationary orbit and recorded data in a raster of 1° latitude x 1°.25 longitude at local noon. In order to accommodate the time differences between day and night, a mean of 4 adjacent TOMS values was adopted for each stellar observation. In Fig. 4 the ozone abundances derived from the stellar spectra are compared to the TOMS measurements interpolated at the same locations as the respective astronomical observatories; the dotted line is a least-squares fit. The result illustrates the success of the proof-of-concept, and justifies applying the technique to much older stellar data. The values derived via Vega (observed near the zenith) and via Sirius (observed low in the sky) are distinguished in that plot by filled and open circles respectively, in order to demonstrate the lack of any bias despite the considerably different path-lengths of light from those two stars through the Earth's atmosphere.

**Built-in Uncertainties**

Some uncertainties are inevitable, and one of the worst in this work stems from the nature of the Huggins bands, which are ~20 Å wide and only about 10% deep; errors caused by placing the stellar continuum at the wrong height, or through blending with the star's own absorption features, could thus have an unfortunately large effect on measurements of the amount of O₃ absorption. Another uncertainty was due to the decreasing stellar signal towards shorter wavelengths, where the reduced sensitivity of the spectrograph plus detector was compounded by the increasing absorption by the Earth's atmosphere. Consequently, unless expressly exposed for features that occur within the Huggins-band region, most stellar spectra tended to be rather underexposed at those wavelengths.
Archival work is eclectic, and the researcher has no control over the quality and quantity of suitable data, though may beat down the errors by increasing the sample size. However, there are limits; in view of the high sensitivity of \( \text{O}_3 \) columns to meteorological conditions, it may not be prudent to average values over more than one night. Furthermore, since the ozone bands can only be measured reliably in stars that are at least as hot as Vega or Sirius, the sample is likely to be restricted on that account too.

**APPLYING THE CONCEPT**

One can now apply the technique outlined above to any spectrum which shows the Huggins bands without serious interference from stellar features. The case of Sirius has the added advantage that one can remove the actual stellar features by dividing by a spectrum of that star observed from space, since the latter contains no \( \text{O}_3 \) signatures; that was in fact done for the four Sirius spectra used in the analyses outlined above. Few (if any) other ozone-candidate stars seem to have been observed from space in the relevant wavelength region, so for those other cases one has to interpolate where an ozone profile is masked by stellar features. Tests carried out with the spectra of Sirius were invaluable in demonstrating whether, and where, significant errors could be introduced on that account. In the case of Vega I used a high-quality synthetic spectrum, and it worked almost as well as a space observation might have done; but there are very few equally reliable models for other stars.

An analysis of 16 spectra of various stars, observed from Mount Wilson between 1935-1941, revealed other factors that need to be taken into consideration. A substantial limitation was not so much the measurements, or the data quality, but actual variations in diurnal concentrations of local \( \text{O}_3 \). Deriving results that are meaningful to a study of ozone evolution therefore requires making as many measurements as possible per night, and (where data exist) by making repeated measurements on the same star on successive days. Although there are far fewer such spectra in any astronomical archive compared to the abundance of solar observations at an ozone observatory, there should be enough around the world to support a very worthwhile attempt to reconstruct \( \text{O}_3 \) concentrations in the early decades of the 20th Century. A few stellar spectra recorded at Potsdam from 1903-5 (well before most observers had heard of ozone!) were available, though of small format and my results had correspondingly low precision.

**THE BROADER VISION**

This example of recovering historic observations, rescuing their information by converting their analogue formats to digital ones, and finally translating the astronomy into atmospheric science, illustrates well the many challenges that one may meet when wishing to ‘repurpose’ data. Depending on the complexity of the observations, it can require specific and expert handling to perform the data conversions correctly, and—in the case of astronomical spectra—that will certainly involve the use of a purpose-built scanner and not a commercial flat-bed machine, whose propensity for positional errors towards the edges and for scattering light will quickly vitiate the results and devalue the effectiveness of this essential step of ‘repurposing’. Different types of historic data will present quite a different spectrum of challenges. Hand-written paper records may be illegible or too faded to read, damaged (not just crumpled), disordered, and easily prone to loss or physical damage until they are archived in a protective...
Even running them to ground, overseas or at home, can be a literal paper-chase. On the other hand, observations that have been ‘digitized’ and archived to magnetic tapes in now-forgotten formats require expertise of a quite different nature. Moreover, many of these efforts will probably not come within the purview of routine research, so finding both resources and people to undertake these data-rescue tasks is another challenge, and is where archivists, librarians and even the general public (‘Citizen Science’) may be of substantial assistance.

Some data-rescue missions have already been carried out very successfully, and more than justify the resources that they required. Digitizing historic hand-written oceanographic data by GODAR, a NOAA project involving worldwide networks, has transformed understanding of subtle and not-so-subtle changes in the temperature, plankton and chlorophyll profiles in a way that no amount of modern data could. Old log-books from sea-voyages have turned up a treasure-house of weather records that are now painstakingly being digitized (Old Weather project) in Europe, and are being used to firm up the statistics of changes in past weather patterns. A small group of hydrologists in South Africa, studying flow rates in the mountain streams that feed Cape Town’s reservoirs, digitized 74 years’ worth of hand-written records and discovered that mysterious drops in reservoir levels could be attributed to a misguided decision to reafforest the mountain slopes with non-native trees. Public health records, when combined with demographic, medical and census data, yield essential food for studies of diseases and how epidemics occur. Statistics of bird sightings, plant locations, migrating animals and modulations in species all provide the hard evidence that is infinitely more useful than mere impressions. We cannot foresee how science itself will evolve; yesterday’s compilation of patient statistics for a cholesterol study—or indeed observations of stellar spectra for highly-focussed astrophysics research—may well take on new significance as thinking in the field matures or new ideas need to be tried out. We now have the tools to tackle research of much more complexity and magnitude than ever before. In almost any field of science or social history fascinating new knowledge will emerge from repurposed old data once adequate resources and support are made available.

Considering that critical knowledge relating to variations in the Earth, the atmosphere and the sky is at stake, not only scientists but all citizens need to work together to ensure that these challenges be tackled competently, thoroughly, and soon. This is more than merely a splendid exercise in repurposing data; the future of our planet may depend on how well we respond collectively to the need.

**The Take-Home Message**

Individuals, institutions, organizations and nations are taking for granted the modern requirement to preserve everything that is available electronically. By building on those assumptions, the digital industry is burgeoning into a major employer and producer; we are at the mercy of dictates to upgrade constantly, while any deliberate destruction of electronic information has become a sin of the first magnitude. The situation requires the exercise of careful control.
An essential first step of that control is to look carefully at what is being done within the context of each discipline, and to identify the relevance and value of a comprehensive digital repository. To that end we must not forget that neither civilization nor the natural world—the Arts or the Sciences—commenced with the digital age. The date was incidental; there was no sudden jog in human or in Nature’s evolution, so pre-digital data have just the same weight as born-digital ones. The extent to which older analogue data from whatever field and discipline are being converted in order to match newer electronic sets is a measure of our grasp of that central principle.

The further principle—that archiving data is to a means to an end and not an end in itself—looks for its evidence to burning current issues that involve the well-being and perhaps the very future of the Earth and its inhabitants. The attempt outlined above, to undertake the transference of pre-digital ozone signatures into a product that has high potential for helping to understand a natural feature which is vital to everyone and everything, was carried out by an isolated individual for a limited period while grant funding was available. When placed in context against the relentless drives to digitize and save everything we possess or create because that is how we work nowadays, the need to prioritize our collective digitizing pursuits is clear indeed.
Towards a Virtual Centre of Excellence Supporting Digital Preservation of the Records of Science

Matthias Hemmje¹, Ruben Riestra², Simon Lambert³ and David Giaretta⁴

¹University of Hagen, Faculty of Mathematics and Computer Science, Chair for Multimedia and Internet Applications, Universitätstraße 1, D-58097 Hagen, Germany
²INMARK, Innovation Management, Av. Llano Castellanos, 4328034 Madrid, Spain
³Science and Technology Facilities Council, Scientific Computing Department, STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot Oxfordshire, OX11 0QX, United Kingdom
⁴Alliance for Permanent Access, 2 High Street, Yetminster Dorset DT9 6LF, United Kingdom

E-mail: ¹matthias.Hemmje@fernuni-hagen.de, ²rar@inmark.es, ³simon.Lambert@stfc.ac.uk, ⁴director@alliancepermanentaccess.org

Abstract—This paper provides overview of the work in various European research and development projects towards building a Digital Preservation e-Infrastructure for the Records of Science as well as a Virtual Centre of Excellence supporting its initial deployment, further development and sustainable operation in the future.

Introduction and Motivation

Digital Preservation (DP) offers the economic and social benefits associated with the long-term preservation of information, knowledge and various forms of know-how for re-use by later generations. However, DP initially had a great problem, namely that preservation support structures were quite often built on projects which were short lived and their results were fragmented.

Therefore, the unique feature of the currently ongoing project Alliance Permanent Access to the Records of Science in Europe Network (APARSEN) [1] is that it is building on the already established Alliance for Permanent Access (APA) [2], a membership organization of major European stakeholders in DP that was founded in order to overcome fragmentation and integrate results of a DP infrastructure as they are already available from the research and development activities of its different stakeholders.
Towards a Virtual Centre of Excellence Supporting Digital Preservation of the Records of Science

Therefore, these stakeholders have now come together within APARSEN to create a shared vision and establish a framework for a sustainable DP infrastructure providing DP support for permanent access to digitally encoded scientific information. To this self-sustaining grouping APARSEN is bringing a wide range of other experts in DP including academic and industrial researchers, as well as researchers in other cross-European organizations.

The members of the APA and other members of the APARSEN consortium already have undertaken and still undertake research and development in DP individually but even here the effort is fragmented despite smaller groupings of these organisations working together in specific international, European, national and even regional projects.

In summary, APARSEN currently helps to combine and integrate the results of these programs into a shared European framework, thereby starting to create a so-called Virtual Centre of Excellence (VCoE) in support of DP of the Records of Science in Europe which is based on a conceptual vision that is currently under development.

Therefore, APARSEN currently and after the end of the project APA itself through the then established APARSEN VCoE will provide a natural basis for a longer term consolidated provision of DP offerings and infrastructures for securing sustainable DP practices guaranteeing the access to scientific resources. These offerings will include the necessary software, services, and consultancy as well as in general the necessary expertise and pragmatic know-how as a basis for further transversal innovation and applied research in the area of DP infrastructures for science.

In consequence, the APARSEN VCoE’s portfolio of offerings will cover technical methods for preservation, access and most importantly re-use of data holdings over the whole Archive-centric Information Lifecycle in the context of DP [3]; legal and economic models including costs and governance models including digital rights management; as well as finally education and training within and outside the consortium to help to create a discipline of data curators with appropriate qualifications.

In the remainder of this paper we will first of all revisit a conceptual model for the emerging DP market that is followed by a description of the DP context of a scientific information life-cycle that can be considered the target model for any kind of valorization within DP value chains for scientific resources. Furthermore, the paper revisits our understanding of value adding within DP preparation and execution processes along such a scientific information life-cycle. Supporting this adding of value within the services of a Virtual Centre of Excellence is then outlined along some exemplar service that will soon become available in a growing European e-Infrastructure for DP of scientific resources.

THE EMERGING DP MARKET

Building on one of the specific results that have already earlier been presented in the DP Impact study [4], this chapter is dedicated to depict DP as an emerging market where market analyses are instrumental for a better understanding of which parties are directly
involved in providing and/or using DP products and services, which might may their likely future economic behaviours and to identify those other players who may be benefit from the deployment of DP policies and processes.

Taking into account conclusions outlined in DPImpact about further growth of DP demand allows to state that current situation of DP related activities just represent quasi the ‘tip of the iceberg’ and in order to follow-up in understanding this emerging market, market dynamics and market lifecycle models are especially helpful.

Within this context, DPImpact first of all presented the basic modelling adopted to describe the emerging DP Solutions & Services Market. In order to facilitate the analysis of the DP solutions and services market, the model depicted in Fig. 1 from DPImpact:

![Fig. 1: Model of the DP Solutions and Services Market in the DP Impact Study [4]](image)

The basic description of the components of the model is described in the following.

**DP macro-environment**: The setting of Political, Economic, Socio-cultural, and Technological external factors influencing the creation and development of the DP solutions & services market.

**Born-digital content**: The ‘input or raw material’ of DP processes. This element of the market comprises all born-digital (as well as the previously digitized) content worth being preserved, to guarantee its long term accessibility and usability.

**Preserved born-digital content**: The ‘final product’ preserved asset of DP activities, to be used in the future by ‘end users’ (e.g. the Designated Community (DC) in the OAIS model [5]) either by their own creators, i.e. in the case of organizations acceding to their own memory or by third parties, configuring the ‘end users’, e.g. researchers, general public, law-enforcing agencies or companies acquiring preserved digital resources for its repurpose and further commercial exploitation.

**DP solutions & services market**: This comprises the players involved in endeavors aimed at providing (Supply) or using (Demand) DP solutions and services rendering preserved born-digital content. This central component of the model includes: the Supply of DP
Towards a Virtual Centre of Excellence Supporting Digital Preservation of the Records of Science

solutions and services, i.e. suppliers of software applications and tools for performing DP as well as commercial and non-commercial providers of services covering parts of DP preservation process (e.g. appraisal, ingest, etc.) as well as the Demand players, i.e. organizations–mainly Memory Institutions and Scientific and Research Institutions–responsible for providing future access to preserved assets to ‘end users’, as well as those using DP for covering their internal needs of preserving digital assets, such as Businesses, Companies and Enterprises who must keep records for long periods of time, too (e.g. in the area of their own applied research or even more market oriented innovation processes).

One fundamental feature of this market is defined by the time required for completing the economic cycle of preserved content, i.e. from its production–converting a born digital resource into a Preserved asset–until the moment of its consumption–accessing and re-using a preserved asset.

This key trait shows critical implications for DP development and its long term economic sustainability, e.g. the time periods necessary for recovering the preservation costs, in terms of explicit or implicit prices paid for using preserved assets, might be measured in years or decades.

This situation so far has generated low attractiveness around DP, due to the lack of a short term Return on Investment (ROI) for most funding schemes acting on opportunity costs conditions. For example, investing in road safety leads to a lower number of accidents almost immediately, thus providing positive results to those responsible for the investment or its funding. Nevertheless corresponding business-models that can be applied in such a DP context over a long period of time are fundamental for the sustainability of the overall service of a DP infrastructure.

Once a collection of digital content is subject to a preservation process, such process should not be interrupted (e.g. ‘skipping’ migrations), to avoid obsolescence problems. This feature makes DP ‘production processes’ to be endless processes, exposed to multiple threats, such as funding disappearance once the initial phases have been surpassed. As some interviewees of DPlmpact from the public sector mentioned: ‘once the original players have moved to other positions, reasons for starting funding tend to vanish if there are no short term results to be shown’.

To conclude, DP Impact has presented the dynamics of the DP market in terms of a model, and has provided an early overview of its stage of matureness, instantiating the Crossing the Chasm market lifecycle model widely used for the analysis of ICT based markets.

Now it is one of the challenges of APARSEN in the final year of its operation to establish a much more concrete understanding of the current state of the emerging DP market as an opportunity to further develop and sustain the work that has already started on the APARSEN VCoE definition and its consecutive creation and sustainability.
THE VALORIZATION CONTEXT OF DP IN THE INFORMATION LIFECYCLE

As the OAIS model is very well known to many stakeholders in traditional DP application domains, this paper does not need to go into further details of its further refinement. Its main properties as already outlined are that it has a producer side where content is produced and that it has a consumer side where the intention is to retrieve and re-use the content after a certain period of time.

From a Computer Science point of view, this means that the overall challenge consists of a transport problem between these two producing and consuming communities that may exist in different times. But it is not like in classical Computer Science that transport means to get the information from one geographical place in the world where the producer has produced it, into another place in the world, bridging the geometric distance between these two places; but the really important challenges are related to the transport of digital content over a time distance.

Another challenge inherent to the DP problem is the fact that the complexity and therefore the cost of the transport problem does not increase with the distance but with the probability, and amount, of change taking place within this time distance on the level of the preserved objects as well as within their environment that consists of a preserved and a non-preserved part.

As users want to deliver content at one time and retrieve, understand, adopt and re-use it at another time the preservation has to bridge the temporal distance between the sender and the receiver and at the same time the amount of change that has taken place to the environment in which the content needs to be applied. Of course this conceptual model of DP is totally different from classical communication models in Computer Science and also the valorisation is not only taking place during the preservation time, but already in the preparation activities providing preservation readiness before ingest as well in the adoption activities providing reusability after access from the DP archive.

As already outlined earlier, the core DP functionality takes place in the centre of such an archival system. Taking content into such a DP collection is the ingest process, and providing access to it and distributing it again is located on the consumer side with a special interest in adaptation and preparation of the content in such a way that it is immediately fit for future re-use and especially re-use support by means of further automated processing.

Within the intake and core preservation processes, the really difficult issues are related to the fact that the core DP process as well as the supporting ingest and access functions depend on a lot of information from the production context that does not exist in the archive today, although many of them are already produced and exist in some form, e.g., as meta-data during the information lifecycle of digital-born objects in DP application domains of DCs.

This is why our conceptualisation of the valorisation taking place in the context of DP understood as being embedded into an information lifecycle also needs to take into account pre-ingest and post-access economic effects of such a life-cycles. This means, one can only properly archive, preserve, retrieve and re-use information if a lot of additional information about its production process, and about the original purpose that it was useful for, is captured and preserved as well.
Towards a Virtual Centre of Excellence Supporting Digital Preservation of the Records of Science

Therefore any kind of ingest process also needs to capture conceptual representations about the original purpose of this content and it needs to somehow reflect future re-use of the preserved content, because if the archive is not capable of reflecting the purpose therefore also the value of the archived information that is stored, it will not be able to optimally preserve and manage its changing forms of existence and technical representation over time.

As a bottom line, this means that we have arrived at the general observation that in the very abstract OAIS model; so far only the core functionality of planning and performing DP is explicitly conceptualized.

However, preparatory information-context-capturing steps that need to be taken to be able to perform an optimal preservation process in the core preservation functionality was outside the scope of OAIS itself and the OAIS roadmap lists potential standard(s) for the submission of digital metadata. Therefore, by design, OAIS by itself does define everything required for the purpose of archival and future re-use aspects that need to be taken into account to preserve the content in such a way that the potential re-use value can be achieved in an optimal way.

To overcome this deficiency, research and development in the SHAMAN project [6] has elaborated the model of an Archive-centric Information Life Cycle [3] as depicted in Figure 2 from the perspective of DP. Following this model, Digital Objects undergo during their lifespan the phases Creation, Assembling, Archiving, Adoption, and Reuse which are described in the following:

**Creation** is the initial phase during which new information comes into existence. More often than not, Digital Objects are not created for the purpose of archiving. Their shape aims at use within regular business models of the creator. Creation processes can be rather complex and involve a multitude of stakeholders until chunks of information result which are...
worth considering for archival. Use means the exploitation of information according to the original purpose the object was created for. Traditionally, objects are archived right after Creation. From the perspective of the Archive, Use is a concurrent thread in the life of a Digital Object that also starts with completion of Creation.

**Assembly** denotes the appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future, meeting the presumed needs of the DC. Assembly requires in-depth knowledge about the DC in order to determine objects relevant for long-term preservation together with information about the object required for identification of relevant archive objects and their Reuse some time later in the future. Assembly generates SIPs for ingest.

**Archival** addresses the life-time of the object inside the archive. During Archival,

Digital Objects are managed in the form of AIPs. In most archives, policies prohibit irreversible deletion of content. Hence, preserving is a perpetual activity. The Archiving phase is open-ended-unless Digital Objects are to be irrevocably removed from the archive. As a matter of principal, information disseminated by the archive must enable the Designated Community to use that information.

**Adoption** encompasses all processes by which accessed Digital Objects, received as DIPs from the archive, are examined, adapted, and integrated to enable understanding and Reuse. It is not the mandate of the archive to accommodate arbitrary consumer scenarios and the corresponding requirements of their respective DCs. Adoption makes efforts to re-contextualize Digital Objects and accompanying information for the respective Consumer Environment for effective exploitation within regular operational infrastructure and working processes. Altogether, the Adoption phase might be regarded as a mediation phase, comprising transformations, aggregations, contextualization, and other processes required for re-purposing data. Additional information beyond that provided by the DIP could be drawn on.

**Reuse** means the exploitation of information by the Consumer. In particular, reuse may be for purposes other than those for which the Digital Object was originally created. Reuse of Digital Objects can lead to the Creation of other, novel Digital Objects. Reuse also may add or update metadata about the base Digital Object held in the archive. For example, annotations change information content and associations change relationships with other Digital Objects. In collaborative working environments, there is a continuous flow between access and ingest, in that retrieved digital objects are reused and/or modified, yielding new revisions and additional (composite) digital objects which have to be preserved, along with their provenance information.

Therefore, in the remainder of this paper we intend to talk about *digital asset value generation and securing by means of DP* in the context of this Archive-centric Information Lifecycle of SHAMAN.
SUPPORTING THE BUILDING OF PRESERVATION VALUE CHAINS

One of the core objectives of APARSEN is the creation and continuous refinement of a Common Vision of the DP landscape which will guide the development of the VCoE on the one hand and can further be refined by the research activities APARSEN during the project’s lifetime as well as beyond during the lifetime of the VCoE that needs to be sustained.

Much of the work on building the common vision of course happens in very close collaboration within the research and development streams of APARSEN. However, it was felt important to dedicate a specific effort to the establishment this Common Vision because it forms the basis for the sustainability of the de-fragmentation and infrastructure-deployment activity as a whole, i.e. from technical, organizational and economic points of view.

The initial vision of understanding the digital preservation landscape has been based on the work and results of research projects such as CASPAR [7], SHAMAN as well as the PARSE.insight [8] roadmap and a number of other research and development, road-mapping and foresight activities undertaken by the APARSEN partners.

Since digital resources such as, e.g., scientific raw data and related cross-citation enhanced scientific publications are becoming increasingly complex, their preservation is a shared responsibility involving various stakeholders from publishers, libraries, data centres and research institutions.

In summary, the APARSEN project is now developing a first version of a common vision from all stakeholders in digital preservation of science and can now bring a different, transversal, insight to the wider community which should also promote insight towards understanding DP as an emerging service infrastructure creating/satisfying the demands its very own market and therefore as an emerging economic opportunity for its sustainability.

Value generation and preservation are closely linked. As Giaretta [9] has argued, the techniques that are needed for preservation, namely adding Representation Information (RepInfo), allow one to add value to data. This can help to satisfy the demands for value from data as well as justify the resources needed for preservation while one strives to leave options for future stakeholders. As noted from McKinsey [10], integration of data from many sources and the exploitation of it from more sophisticated analytics creates value from big data. Therefore, one can increase the value of data by ensuring it can be used more widely, and one can do this analogously to analyzing it by adding RepInfo. Furthermore, Giaretta [9] has drawn a diagram which summarizes a first economic meta-model on understanding DP activities as a set of economically driven activity types that need to be instantiated to implement a sustainable business process which is centered around the DP of a digital object as well as incorporated in the overall business cycle of an organization that is responsible for securing the future usage of such assets. The component types of this first model can become a basis for instantiating economic DP activities and for constructing DP value chains along the Archive-centric Information Lifecycle to sustain DP.
However, as these value chains must be sustainable on the long term, at the same time a coherent and cost-effective approach must be taken to DP activities themselves. APARSEN and later the APA VCOE is designed to provide a coherent set of tools, services, advice and training which can help repositories to improve the way in which they preserve their digital holdings, helping them to become a more trustworthy repository.

**Exemplar Tools for Innovative Services in a DP e-Infrastructure**

The currently ongoing project SClence Data Infrastructure for Preservation–Earth Science (SCIDIP-ES) [11] supports this vision of constructing innovative and cost-effective DP e-infrastructures supporting DP value chains by offering generic, sustainable services and toolkits for efficient DP preparation, planning, and management along with access and usability, i.e. adoption and reuse needs. As described in [12] SCIDIP-ES can help data archives to become preservation-aware through adopting its OAIS-conformant services and toolkits and makes them available all through an Archive-centric Information Lifecycle.

For example SCIDIP-ES provides toolkits to create Archive Information Packages (AIPs) using functionalities provided by its Packaging Toolkit and Storage Service. For an existing archive, elements in the AIP would reference locations within the archive, and also link to additional RepInfo in this way supporting an increase of data value.

The amount of RepInfo required is measured against the perceived skills, resources and knowledge base available to the target Designated Community (DC). An archive may wish to broaden exploitation of its data holding by providing additional RepInfo for a wider group of users with different knowledge and resource bases. Given the potential diversity and quantity of RepInfo involved, an archive cannot by itself be expected to capture and manage all the RepInfo that it might require.

To support these key requirements and help share the burden and efforts for DP on the long term, SCIDIP-ES provides the RepInfo Registry Service which would be used to query, retrieve and manage RepInfo required by a group of preservation archives. Note that SCIDIP-ES fully expects that repositories will locally cache the RepInfo they need. To facilitate the use of Rep-Info, the Registry will contain a special type of RepInfo called RepInfo Labels (RILs) which are pointers to multiple RepInfo objects. SCIDIP-ES notes that a RepInfo Registry must itself be an archive, and so, logically, the actual RepInfo data objects would be preserved as OAIS AIPs which include Persistent Digital Identifiers and PDI and its own RepInfo objects to facilitate interpretation. The latter construct gives rise to RepInfo Networks (RINs) and the Registry would also enable users to navigate a network of RepInfo objects to explore the knowledge represented. In SCIDIP-ES, a RIN represents the chosen solution for fulfilling a specific preservation objective.

The Gap Identification Service (GIS) is defined to help assess if a data consumer with a particular knowledge profile can ‘understand’ the preserved digital objects by identifying “gaps” in the corresponding RIN stored in the Registry [13].
Towards a Virtual Centre of Excellence Supporting Digital Preservation of the Records of Science

To assess if the defined archive is OAIS conformant and to identify potential areas for improvement, archivists may use the Certification Toolkit which implements the ISO16363 standards for Trustworthy Digital Repositories Audit and Certification [14] to perform a self-audit.

ACKNOWLEDGEMENT

The work reported here is result of the efforts of the large number of organisations and people involved in APARSEN and SCIDIP-ES.

APARSEN and SCIDIP-ES are funded by the Seventh Framework Program (FP7) of the European Commission (EC) under the Grant Agreements 269977 and 283401 respectively.

REFERENCES

[7] CASPAR project website http://www.casparpreserves.eu, and documents are also available on the APA web site.
[8] PARSE.Insight project http://www.parse-insight.eu, and documents are also available on the APA web site.
Digital Preservation of Cultural Heritage
Securing Scotland’s 3D Digital Heritage: Finding a Practicable Middle Ground

Emily Nimmo
Royal Commission on the Ancient and Historical Monuments of Scotland, Edinburgh, UK
E-mail: emily.nimmo@rcahms.gov.uk

Abstract—As the national collection for Scotland’s built environment the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) has a commitment to ensuring that the widespread capture of 3D laser scan data of Scotland’s archaeological and built heritage remains accessible to future generations. The fast pace of technological change in this area and the absence of an existing mature and workable approach to its preservation places this important record at risk.

Current guidance for the long term preservation of this data requires essential metadata which is often found to be too burdensome by the data creators. This paper will provide a case study of the domain specific challenges of preserving cultural heritage geospatial data through the activities undertaken in partnership to seek a balance between the ease of use required by data creators and the essential documentation sought by archives for preservation purposes.

Keywords: Metadata, Digital Preservation, 3D objects, Archaeology, Cultural Heritage, Laser Scanning, E57, Archaeology

INTRODUCTION

This paper will present the work underway in partnership between RCAHMS, Historic Scotland and English Heritage to produce guidance on the capture and preparation for archival storage of archaeology and cultural heritage 3D laser scan data. Through the partnership this guidance will then be fed back to the software and hardware producers to pursue the goal of ensuring that the defined metadata is captured by the hardware and is conveniently exportable from the software. This will achieve the necessary virtuous circle across the data lifecycle from capture to archive that will allow us to maintain access to this highly complex and valuable data.

The paper will begin by introducing the partners and describing the origins of the project and how this presents the potential to have lasting and tangible impact on the long term accessibility and usability of 3D laser scan data. We summarize the key approaches already identified in the preservation community and highlighting what we consider to be significant obstacles to these approaches offering a practicable system that could be implemented today.
by a national collection such as RCAHMS. Next we will describe the approach of the project and work already underway. Finally it will set out the next steps for this ongoing work and present what we have identified as a key solution to this problem.

**BACKGROUND**

**Origins of the Activity**

RCAHMS was established by Royal Warrant in 1908 in response to widespread concern about the destruction of historical monuments. For over 100 years RCAHMS has been surveying, recording and collecting information on the constantly evolving landscape of Scotland. RCAHMS curates both a traditional physical archive constituting many millions of photographs, drawings and documents about the buildings and landscapes of Scotland from the earliest times to the present day and a digital archive. RCAHMS, has been actively collecting digital archive since the late 1990s. It has a duty to ensure the long term preservation and accessibility of the archive generated as part of archaeological and architectural field work in Scotland and Scottish territorial waters, and set guidelines of the formats used and metadata created used in this process. As part of the Scottish research data infrastructure it is incumbent on RCAHMS to ensure that policy expectations regarding access to research data can be fully delivered.

Scotland is a world-leader in the deployment of 3D laser scanning technologies for the historic environment through the work of Historic Scotland’s Scottish Ten¹ project. This project sets out to scan the five Scottish UNESCO World heritage sites plus an additional five international sites. Historic Scotland has joined with The Glasgow School of Art’s Digital Design Studio, their shared umbrella organization The Centre for Digital Documentation and Visualization and CyArk² to achieve this goal. Beginning with the survey of Mount Rushmore in May 2010 the project³ has now scanned many more than the planned 10 sites and generated many Terabytes of data.

With objectives to digitally preserve important historical sites for the benefit of future generations, provide 3D digital models and data to site staff to better care for their heritage assets and create accurate 3D surveys of the sites for future development of innovative world class research, the project team are concerned that they conform to archival best practice and prepare their datasets in a way that facilitates its long term use and its deposit in a trusted digital repository.

English Heritage’s Geospatial Imaging team carries out metric surveys of the English Heritage estate, using laser scanning, photogrammetry and multi-image based survey approaches. The team also takes the lead in providing advice to the external sector on the heritage application of unmanned aerial vehicle platforms and Building Information Modeling (BIM). The team has extensive experience of image based survey approaches and produced

---

¹[www.scottishten.org](http://www.scottishten.org) [accessed 30 September 2013]
²[www.scottishten.org/index/partners/cyark.htm](http://www.scottishten.org/index/partners/cyark.htm) [accessed 30 September 2013]
³[www.scottishten.org/index/scotten/property1/mountrushmoreproject.htm](http://www.scottishten.org/index/scotten/property1/mountrushmoreproject.htm) [accessed 30 September 2013]
the ‘Metric Survey Specifications for Cultural Heritage’ which sets the standard for metric surveys across the heritage sector. They also represent the UK at CIPA (the International Scientific Committee for Documentation of Cultural Heritage) and co-chair the ‘Cultural heritage data acquisition and processing’ WGV/2 working group within ISPRS (the International Society for Photogrammetry and Remote Sensing).

Consequently the partners found themselves at a significant intersection of a series of research processes. Each partner faces different aspects of the challenges presented by our shared commitment to ensuring that the important heritage data that we collect, create or commission remains meaningfully accessible for future generations. It was therefore natural that we should come together informally to try to find practical solutions. The challenge we seek to address therefore is general to all digital resources; to guard against hardware and software obsolescence by identifying surrogate formats which do not result in significant loss and by adequately describing the objects and the processes that resulted in their creation ensure that they might be used and understood in the future.

Domain Specific Challenges of Preserving and Documenting Cultural Heritage Spatial Data

From anecdotal reports made to RCAHMS and the practical experience of working to secure the deposit of this data to our archive it is evident that current best practice guidance for the long term preservation of 3D laser scan data, in particular the required metadata has been found to be too burdensome to produce by data creators. We have found the impact of this to be extremely negative; deposits of laser scan data come without any metadata from depositors, or the data is not deposited to the archive at all. However unlike the Higher Education sector the challenge we face here is not one of advocacy. Archaeologists hold central the tenet of preservation through archive due to the often destructive nature of their research process.

The transient nature of the built heritage environment—subject to continuous change and modification and subject to complete loss as a result of human or natural disaster—requires that maximum re-use of data be our goal. It may not be possible in the future to return to a site to re-survey due to the disappearance of the original. In this way the documentary archive may become the primary resource.

The Institute for Field Archaeology, the main professional body in the UK states clearly that:

“All archaeological projects that include the recovery or generation of data and/or archaeological materials (finds) will result in a stable, ordered, accessible archive. All archaeologists are responsible for ensuring that the archive is created and compiled to recognized standards, using consistent methods, and is not subject to unnecessary risk of damage or loss.”


6www.archaeologists.net/sites/default/files/node-files/Archives2009-2.pdf [accessed 30 September 2013]
The challenge we face then lies in the cost to companies to achieve this. The commercial archaeological sector in Scotland is based on a developer pays model. Through the planning system archaeological conditions are placed on any type of work proposed that might impact on the historic environment. This work is paid for by the developer and put to competitive tender. The current economic climate has resulted in great pressure to reduce costs in every aspect of the tender an archaeologist submits. Lack of standards relating to the costs of preparing archive for deposit has meant that these costs are subject to the competitive process. A tension is therefore created between accurately costing this work and winning the tender, and between thoroughly preparing and depositing archive and remaining in business.

**Key Approaches**

The potential cultural heritage application for 3D laser scanning was quickly realized and in the last decade it has assumed a great deal of importance both for visualization, virtualization of access and for monitoring and conservation issues. Issues around its long term curation and continued accessibility have been subject to investigation among the research community now for over a decade.

As with any new technology we have a seen a flurry of development of the capture devices across a range of different suppliers. The data generated is very much tied to these proprietary devices and software used to process the data. In practice this makes it very difficult for users of different brands to share and exchange data, or to compare similar datasets for research purposes.

The American National Information Standards Organization (NISO) defines metadata as the “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.” Metadata can cover a variety of different information types such as descriptive metadata which aids information retrieval, preservation metadata which describes the functions or processes of preservation and provenance metadata.

It is provenance metadata that is of specific concern here, as it is that which ensures the maximum re-use of data into the future. This draws heavily on the scientific principal of reproducibility of results. Doerr explains that scientific data is often based on measurements taken on devices with digital outputs. Here provenance metadata allows the user to ‘assess meaning (the recorded things, experimental setup, instrument used), relevance (status, conditions of the recording and derived information), quality (calibration, tolerances, measurement errors, processing artifacts and error propagation) and possibilities of improvement and data reprocessing.”

A number of metadata schemes have been proposed over the last decade to address the need to adequately describe 3D laser scan data to facilitate its ongoing understandability and future reuse. These have addressed the particular challenges of this data type with varying

\[9\] Doerr, M. and Theodoridou (2011)
degrees of specificity. Doyle\textsuperscript{10} for example looked to the OAIS\textsuperscript{11} reference model, METS\textsuperscript{12} and CEDARS\textsuperscript{13} before concluding that they must develop their own bespoke scheme. Pitzalis\textsuperscript{14} points to LIDO, others such as Pittarello\textsuperscript{15} point to standards originating from the semantic web.

It is the extension to the CIDOC-CRM\textsuperscript{16} standard, CIDOC-CRM\textsubscript{dig} developed the by the European Commission funded project 3D-COFORM\textsuperscript{17} that offers the most promise. It provides for the description of provenance metadata. CIDOC-CRM itself describes the material heritage environment as observed by archaeologists and other heritage professionals. In his introduction to the extension Doerr describes how it allows for ‘digital objects participating in actions measuring physical properties and capturing sensor data in a material environment’.\textsuperscript{18} It also captures information relating to the devices that are used and makes it possible to track devices across their life time allowing recurring errors to be detected and corrected for. Through this standard much diligent and painstaking effort has been exerted towards exhaustively defining the information necessary to meet the requirements of understandability and reproducibility.

CIDOC-CRM\textsubscript{dig} appears to be the best effort to date to meet the needs of the archival community in documenting cultural heritage spatial data. However it is in a very real way unimplementable without considerable effort. Felicetti, admits that ‘the difficulties in using these schemas impose on the developers a high degree of automation and the creation of extremely user friendly interfaces’\textsuperscript{19}

As we have seen the data creators, despite their desire to do so, cannot afford the extensive time and effort needed to document their scans to these standards. Organizations such as RCAHMS do not have the resource to develop the extensive necessary user interfaces to make this a simpler process. Lighter weight standards have been developed such as Adobe’s 3D PDF (ISO 32000-1:2008), but what it gains in its light weight approach it loses in the scope for manipulation and reuse of the data and overall lack of flexibility.\textsuperscript{20}

**The Approach of the Project and Work Already Underway**

Initial requirements of the practitioner community have already been gathered at a workshop\textsuperscript{21} organized in partnership with the Digital Preservation Coalition\textsuperscript{22} in May 2013.
At this workshop we brought together representatives involved in archaeological archiving, digital preservation, national cultural heritage organizations, commercial archaeologists and companies producing software and hardware for 3D laser scanning to educate the data creators and software and hardware producers about the challenges we face and to seek their input and consensus on the best next steps. This event also sought to influence the software and hardware producers of laser scanning equipment to produce developments that will make the export of the necessary metadata much simpler and more convenient for users.

Throughout the workshop speakers and delegates alike demonstrated their commitment to the proper archiving of 3D laser scanning projects and ensuring its long term accessibility and usability. However it was clear from this exercise that as a community we must seek to establish new guidelines for the long term preservation of 3D laser scan data, balancing the operational ease of the data creators with capturing essential. We must also find a way to influence the software and hardware producers of laser scanning equipment to produce developments that will make the export of the necessary metadata much simpler and more convenient for users.

It is in this ground up approach, coming directly from the requests of the data creators, and also the enthusiasm with which they have engaged with our efforts to date that affords us a unique opportunity to achieve lasting impact. Despite the commercial pressures already discussed many delegates have offered their time and expertise to engage in our work without remuneration. Furthermore the influence of the partners in the project and the combined influence of the practitioners as the customers of the software and hardware suppliers have already secured their engagement in our efforts.

In light of the significant obstacles to the current approaches to documenting 3D laser scan data for cultural heritage offering a practicable system that could be implemented today outlined in Section 3, the partners identified E57 as a pragmatic answer. Faraz Ravi, Director of Product Management for Point Clouds at Bentley Systems and Chair of the ASTM E57 Standards was invited to provide the workshop with a better understanding of its use and limitations.

The E57 file format is a vendor-neutral format for storing point clouds, images, and metadata produced by 3D imaging systems, such as laser scanners, but can also encode data for LIDAR systems, structured light 3D scanners and others. The file format is specified by the ASTM, formerly the American Society for Testing and Material, an international standards organization, and it is documented in the ASTM E2807 standard.

Among its many benefits as an interchange format E57 supports a wide range of supported features such as gridded data, multiple coordinate systems, embedded images and built in error detection. Furthermore it has wide support across hardware and software developers with many of the latest versions of the leading software providing convenient export to E57.

24www.dpconline.org/ [accessed 30 September 2013]
25www.pointools.com/ [accessed 30 September 2013]
24Huber, D. (2011)
25www.astm.org/ [accessed 30 September 2013]
Nevertheless E57 was originally developed for data interchange, not as a working format and not for archival or preservation purposes. In its initial development Huber\(^{26}\) describes that in seeking to achieve a minimalist approach additional information of use to future users such as vendor-specific calibration parameters were to be excluded. It is this specific metadata however that we seek to record and therefore much work remains to be done if we are to achieve our goals.

**CONCLUSIONS AND NEXT STEPS FOR THIS ONGOING WORK**

We propose to develop an archival extension to the E57 format, building on the important work of 3D-COFORM and the CIDOC-CRM\(_{3D}\) metadata standard. To this end the E57 subcommittee on Data Interoperability has already been approached about developing such a preservation extension and this has been warmly received.

Our next steps will be to secure support for the design of the extension and its realization and continue to engage with the data creators to ensure this approach meets their needs and that they quickly adopt the resulting solution. The industry representation on E57 committee and the existing widespread take-up of the standard within the software today means that this approach significantly diminishes the barriers to success that we have seen in other approaches. We propose that this will allow us to balance the essential documentation sought by archives for preservation purposes with the need for an implementation that minimizes the effort required of data creators.

Alongside this effort we will develop guidance documentation and training materials to support the practitioner and archival community in taking this up. Finally we will undertake further requirements gathering from our range of user communities to establish the best way to make this important data available to our various audiences in a way that meets their needs and technical capabilities and maximize its value for the people of Scotland and the international research community.

**REFERENCES**


\(^{26}\)Huber, D. (2011)


Audiovisual Archives in India in the Digital Era and Initiative for Setting up a ‘National Cultural Audiovisual Archives’

Pratapanand Jha
Director (Cultural Informatics), Indira Gandhi National Centre for the Arts, New Delhi
E-mail: pjha@ignca.nic.in

Abstract—India has a huge repository of recorded cultural heritage, stored in audio-visual form, at different institutions. The content of these holdings enshrines the creativity of some of India’s greatest artistic talents. Such a large volume of cultural wealth, created in the last six decades is on the verge of being lost forever. The main reason behind this dangerous situation is the absence of systematic and modern preservation technologies, lack of awareness and proper upkeep, as well as the fragility of the medium they are stored in. Moreover, with frequent changes in hardware and advancements in technology, the playback of these audio-visual materials, which are mainly in analogue format and stored on different types of tapes, records, cylinders etc, has become extremely difficult. The machines to play these tapes etc are no longer being manufactured. In cognizance of this concern, the Ministry of Culture, Government of India has sought to take an initiative to avoid any further deterioration of such materials. This paper desires to give an overview of the issue of archiving the perishable and delicate audio-visual cultural heritage of India in the digital era and the Ministry of Culture’s concern in preserving them in the long run.

Keywords: Cultural Heritage, Audio-Visual Form, Preservation Technologies, Changes in Hardware, Analogue Format, Archiving, Government Initiative, Dissemination

INTRODUCTION

A large volume of India’s cultural wealth, created in the last six decades, is stored in audio-visual form with various Government, non-governmental institutions and private collections. The content of these holdings enshrines the creativity of some of India’s greatest artistic talents. This is an invaluable national heritage that needs to be preserved for all time and also made accessible to the citizens of the country.

In the absence of systematic and modern preservation technologies, lack of awareness and proper upkeep, as well as the fragility of the medium they are stored in, these materials are in imminent danger of being lost forever. Thus, preservation becomes critical. Moreover, with frequent changes in hardware and advancements in technology, the playback of these
audio visual materials, which are mainly in analogue format and stored on different types of tapes, records, cylinders etc, has become extremely difficult. The machines to play these tapes etc are no longer being manufactured. Therefore, even if such legacy material were preserved, playing and using the material due to hardware obsolescence has become a major challenge. This is a world-wide phenomenon.

**MINISTRY OF CULTURE, GOVERNMENT OF INDIA’S INITIATIVE**

Recognizing the concerns, the Government of India, Ministry of Culture approved for setting up of the “National Cultural Audiovisual Archives” to minimize further deterioration of such materials. The vision and objective of the project is to extend the awareness about the audiovisual materials, their proper upkeep and create digital archives of international standards through partner institutions identified for the purpose. To achieve this task, an appropriate technological and institutional framework will be setup. This project will result in creation of state-of-the-art audiovisual repositories and secured dissemination of the content as well as its long term preservation and accessibility. The process of digitisation will also entail identification and categorisation of all legacy material and creation of metadata for easy retrieval. Such an archive would be of great significance to lovers of the performing arts, historians, journalists, sociologists, scholars, researchers and aesthetes in general.

There is no doubt that the scope and magnitude of the project is very vast, given the rich cultural diversity of India and the huge volumes of such material available in different institutions. However, this is a task that has to be undertaken in the national interest before this cultural heritage is irretrievably lost. It has to be taken up in a mission mode with earmarked financial resources, clear time lines and specific milestones. At present, the scheme is envisaged to be completed in the 12th Plan period. However, once the extent of cultural audio-visual holdings across the country is identified and quantified, it is possible that this may have to be treated as Phase I of the project and the project may have to be continued beyond the 12th Plan in order to reach fruition.

In this scenario, archives and collections the world over have successfully adopted the process of digitisation to preserve and re-purpose their holdings. A similar exercise is imperative for the audio-visual content available in our cultural institutions. While most audio-visual content is generated digitally today, which makes preservation of current material relatively easy, the critical task is to digitise the legacy material that are stored on different types of non-digital analogue formats. Once digitised, the content will be stored on file formats in servers without danger of deterioration or loss. Digitisation also facilitates easier retrieval and dissemination of content, thus making it easier to use and to make it more widely accessible to the public and to future generations.

**THE ROLE OF IGNCA**

The IGNCA, which has the experience of successfully completing the UNDP project titled ‘Strengthening National Facility for Interactive Multimedia Documentation’ on behalf of the Ministry of Culture through its Cultural Informatics laboratory, is ideally suited for
undertaking this project. Apart from partner institutions, it will also use the services of archival experts who are well versed in archival administration, digital technology and establishment of archival libraries as members of the Committees that will be set up to steer and monitor the project. The IGNCA already initiated a pilot project with C-DAC, Pune, for setting up of the digital archives under National Digital Preservation Program (NDPP). C-DAC is a scientific body under the Department of Information Technology, MCIT who has been assigned to prepare a long term digital policy for the country.

In order to kick start the process, IGNCA convened a meeting of Experts on July 19, 2013 to generate informed opinion about the scope and contours of this project. Detailed discussions were held and the Experts were unanimous that such a project was absolutely essential to preserve our cultural wealth, which is in imminent danger of decay and destruction. The experts also gave valuable suggestions on the technical processes to implement the project and the proposed road map to make this project a reality at the earliest.

It may also be stated that in order to ensure long term preservation of audio-visual material, there is no alternative to digitisation. Legacy material like tapes, spools, records, films etc are subject to deterioration and decay even if stored under optimal conditions. Thus, the project design necessarily has to be based on adopting the digital process to preserve this material over the long term. This task is proposed to be undertaken by IGNCA along with 4-5 partner institutions to be identified for the purpose.

The proposed project will serve as the primary mechanism at the national level for the preservation and accessibility of cultural audio-visual materials and its permanent availability.

**Scope and Objective of the Project**

The first and foremost objective of the project is to identify and preserve the cultural heritage of India available in audio-visual form in different institutions across the country through a process of digitisation and making it accessible to the people. Secondly, instituting state-of-the-art digitization and storage systems through the aegis of IGNCA and 5-6 partner institutions to preserve these audio-visual resources is an important objective. Setting up a dedicated Website and a virtual network of these repositories and offering online access to their resources is the third objective. Standardization and periodic upgradation of the methods and technologies used in production, storage and retrieval of audio-visual resources is an important objective. The genres to be covered will include oral traditions, traditional crafts and textiles, dance, music and theatrical practices, cultural practices and traditional knowledge is. The fifth objective is Capacity Building in conservation, cataloguing, digitization and retrieval of audio-visual materials. Another very important objective of the project is, instituting Outreach and Awareness programmes.

**Nature of the Cultural Audio-Visual Materials and Targets of the Project**

The Table 1 below indicates the audiovisual holdings in some of the Governmental and Non–governmental institutions in India.
Table 1: Cultural Audio-Visual

<table>
<thead>
<tr>
<th>Institution</th>
<th>Audio Visual Resources (in Hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indira Gandhi National Centre for the Arts, New Delhi</td>
<td>17,100</td>
</tr>
<tr>
<td>Sangeet Natak Academy, New Delhi</td>
<td>16,500</td>
</tr>
<tr>
<td>The Indira Gandhi Rashtriya Manav Sangrahalya, Bhopal, M.P.</td>
<td>4,767</td>
</tr>
<tr>
<td>Archives and Research Centre for Ethnomusicology, AICS, Gurgaon, Haryana.</td>
<td>12,361</td>
</tr>
<tr>
<td>Kalakshetra Foundation, Chennai</td>
<td>1,650</td>
</tr>
<tr>
<td>Regional Resource Centre, Udupi, Karnataka</td>
<td>5,000</td>
</tr>
<tr>
<td>Rupayan Sansthan, Jodhpur</td>
<td>9,500</td>
</tr>
<tr>
<td>Madhya Pradesh Adivasi Kala Parishad</td>
<td>100</td>
</tr>
<tr>
<td>Shreyas Folk Museum of Gujarat, Ambawadi, Ahmadabad, Gujarat</td>
<td>248</td>
</tr>
</tbody>
</table>

In all, about 65,000 hours of audio-visual materials are in possession of selected institutions. Given the fact that such holdings are available with a very large number of organisations/collections across the country, providing measurable baseline data and specific targets at this stage is obviously not possible. However, the data and baselines will be established through a detailed survey that will be undertaken to assess the physical targets, at a later date. The task of digitisation will commence only after full details of the baselines, scope of work, measurable targets, milestones in physical and financial terms etc are available and approved by the Steering Committee.

THE MAIN TARGETS OF THE PROJECT

Firstly, Constituting a Project Unit under a Project Director along with a core team to run the project and preparing a detailed DPR to be considered and approved by a Steering Committee.

Secondly, identifying various types of legacy machines to play back the content is a big issue. The project scope will cover hiring, purchase or repair of such equipment, wherever necessary.

Thirdly, Preparation of a priority list of the content to be digitised, based on clear guidelines to be formulated for the purpose by the Steering Committee but ensuring that the most endangered material is digitised first.

Fourthly, it is important to establish digitisation centres having facilities/equipment This would be at par with international standards. The process can be done at IGNCA and at 5-6 zonal Centres to be run by partner institutions. The central body along with partnering institutes would complete the task of digitisation in congruence with each other. The costs of digitisation at the Zonal Centres (partner institutions) will be based on transparent and realistic yardsticks and computed on a ‘per job’ basis.

Further, ingestion of data from analogue to digital form as per international norms, along with relevant metadata and preserving the digitised material in file format on servers will be under taken. Again, providing international standard storage for archiving the physical legacy audio-visual materials in at least at five major repositories in the country, with prescribed standards of temperature and humidity will be taken care of. Moreover, developing an appropriate IPR Policy based on relevant laws in force for use of the material.
Establishing a dedicated website and publishing on-line catalogues of the cultural audio-visual materials available in cultural institutions/individual collections is also in the agenda.

Providing free accessibility of copyright free materials and payment based accessibility of copyrighted materials also forms part of the target.

The target includes capacity building in the area of audio-visual digitization, metadata creation, cultural dissemination, storage and conservation.

The last important aspect of the target is to implement Outreach/ Awareness programmes in the field of audio-visual preservation.

**The Design of the Project**

Schematically, the design of the project incorporates some important issues. These are, as follows:

Digitisation and ingestion of metadata under the project will be done in the digitisation centre at IGNCA and at selected major repositories of analogue/ digital audio-visual materials across the country. Such institutions will be identified based on their track record, expertise and technical capabilities. They will digitise their own materials and also extend their services to other institutions and collections. The digitised material will be kept in servers at IGNCA and in the partner institutions with permission of the owners. The original legacy materials will be maintained at IGNCA and the partner institutions in accordance with international standards of storage and conservation, unless the owners themselves have this capability.

The digitization lab at the IGNCA will take care of digitization of the most critical cultural audio-visual materials in the country. Most of the materials generated today are born digital and they do not require any such facility. A ‘Trusted Digital Archive’ will be set up at the IGNCA, as per the approved standards and in collaboration of the National Digital Preservation Program (NDPP), for long term availability of the digital materials. A dedicated website of NCAA will be created to showcase the catalogues of the material after digitisation. Searchable metadata of the collections in standard format will be made available online, in the public domain, through this website. Web based accessibility of this data, open or restricted, complete or partial, free or priced, based on the nature of collections, will be provided. However, these can be made accessible to the academic institutions, with the signing the MOU with IGNCA.

Capacity building workshops/ trainings will be conducted periodically by IGNCA. Collaboration with reputed institutions like CDAC, Digital Library of India, IIT’s, NIIT’s etc will be considered by the Steering Committee to ensure the best possible standards. Awareness/outreach programmes will be conducted as regular activity of the project. Final decisions of the project design will be detailed in the DPR and approved by the Steering Committee. As part of digital archives, and for long term preservation of data, disaster-safe copies of the content created after digitisation will be stored in neutral locations in accordance with disaster management procedures.
CONCLUSION

The project will be monitored by the National Steering Committee headed by the Secretary, Ministry of Culture assisted by Member-Secretary, IGNCA. For day-to-day project monitoring and implementation respectively, a Monitoring Committee of experts from various disciplines will be constituted.

This proposed Archive is going to be a place ‘where recordings are stored for the purpose of both preservation and use.’ (Chaudhuri 2004) Archives, by dint of having the aim of preservation always differ from libraries. The present project also seeks to build the National Cultural Audio-Visual Archive in this mode. Furthermore, this project will ‘...also place a stronger emphasis on preservation for the future than most of the libraries which [will] make the recordings more easily available to the public.’ (ibid) So the entire process envisages a thorough system of preserving our cultural heritage in audio-visual format, as well as, disseminating them far and wide to the people.

REFERENCE

Digital Preservation of Cultural Digital Archives—Work in Progress

Neeraj Mittal¹, Milind Kapre², Sandeep Gaikwad³, Dinesh Katre⁴ and O.N. Chaubey⁵
¹²³Project Engineer, Human-Centred Design & Computing Group, C-DAC, Pune, India
³Senior Technical Officer, Human-Centred Design & Computing Group, C-DAC, Pune, India
⁴Associate Director & HoD, Human-Centred Design & Computing Group, C-DAC, Pune, India
⁵Ass. Library & Information Officer, Indira Gandhi National Centre for Arts, New Delhi, India
E-mail: ¹milindk@cdac.in, ²neerajm@cdac.in, ³sandeepg@cdac.in, ⁴dinesh@cdac.in, ⁵onchaubey@gmail.com

Abstract—This paper presents the work in progress report on the digital preservation project undertaken with Indira Gandhi National Centre for Arts, New Delhi by C-DAC, Pune as part of the Centre of Excellence for Digital Preservation project. In this project, the digital preservation R&D team at C-DAC, Pune is developing DatāntarTM and Sanskriti DigitālayaTM systems for the preservation of cultural digital data, which is briefly discussed to provide an overview of the activities.

INTRODUCTION

Pilot Digital Repository of Cultural Digital Data

As part of the Centre of Excellence for Digital Preservation project sponsored by DeitY, the digital preservation technology development team at C-DAC Pune is working with Indira Gandhi National Centre for the Arts (IGNCA), New Delhi in order to develop the pilot digital repository of cultural digital data. IGNCA collects and manages the cultural heritage of India in terms of ancient manuscripts in Microfilm/ Microfiche form, rarebooks, slides, audio visual material of cultural significance. They have already digitized a significant portion of this material and maintained the digital copies in various kinds of digital storage media such as CDs, DVDs and hard disks. There are such thousands of CDs and DVDs which has turned into another library of its own kind. The digital preservation technology development team at C-DAC Pune is developing the DatāntarTM and Sanskriti DigitālayaTM systems for preserving this cultural digital data. Long term preservation of such data in terms of images, audio/ video clips and 3D
models is our main focus. It is extremely important as the cultural digital data is subjected to
digital obsolescence, media failure and it could be easily destroyed or corrupted by computer
-crash or virus.

This paper provides an overview of the Datāntar™ and Sanskriti Digitalaylak™ systems
which are being developed for the preservation of such cultural digital data.

Related Work
Therefore, in order to address the requirements of IGNCA with regard to controlled access,
archival formats, online availability and digital preservation, we studied similar international
initiatives. British Library [5] and National Library having over 280 terabytes of collection or
over 11,500,000 million items are planning to preserve to make the content accessible for
different institutes like Archives, Libraries and Museums. Digital Preservation of cultural tangible
heritage in Indonesia 'eCultural Heritage and Natural History (eCHNH) Framework’ [2]
facilitate multi-channel access.

Context capturing in terms of metadata is very important while preserving the cultural
digital objects [6]. As part of our project, our system is designed to support intranet based
collaborative framework for enrichment of the metadata by curators, indologists and scholars. It
also supports OCR and metadata description using large variety of parameters provided by
MARC21 in order to capture the context of the cultural digital object. There is also much
discussion on participatory model for heritage organizations wherein the expertise available in
different organizations can be used [9]. C-DAC is exploring the collaborative framework over
internet for incorporating the expertise available across India and abroad in virtual museums
project [8]; however, in the present scope of digital preservation project it is yet to be included.
A Digital Archives Framework for the Preservation of Artistic Works [1] provides coherent
framework approach that is quiet similar to our work.

MAIN CHALLENGES
Following are the challenges associated with cultural digital data

Challenges of Cultural Digital Heritage as Perceived by IGNCA
Over the years IGNCA has enriched its collections with help from various individuals and
institutions from India and Abroad in terms of the following broad categories:

• Rare Books.
• Personal Collection.
• Cultural Archives Collection.
• Manuscripts Collections in Microfilm/ Microfiche form.
• National Mission for Manuscripts, IGNCA.
• Slide Collection.
• National Audio Visual Archive (new initiative).
Most of the above collections available at IGNCA are already digitized or being digitized. The challenges faced in managing and accessing these digital collections is as under:

- Open access versus restricted access is a debate going on at IGNCA since long. Still no decision has been taken about access to cultural heritage at IGNCA under open access environment.
- More than 60% of non-print material at IGNCA has been digitized. However, there is no centralized archiving system or online digital library in plan. Many of these materials are available in either CD/ DVDs or digital in form of TIFF and JPEG files. Millions of pages of digital material have been stored in form of different files. There is no PDF or searchable PDF file to read particular manuscript or other multiple page documents.
- Lack of comprehensive Metadata for all digitized material is also a big hurdle in access to digitized material.
- IGNCA also not having any well-drafted digital preservation policy for long term preservation of digitized material.
- IGNCA is leader of digitization of cultural heritage, however it does not have all well placed IT infrastructure in place. Lack of trained IT manpower also affects the various digital Library initiatives.

The digital preservation R & D team at C-DAC Pune has analyzed the challenges as under.

**Handling CD/ DVDs**

We have observed that IGNCA or other cultural heritage organizations and museums are producing enormous amounts of digital data that they are primarily storing in CDs and DVDs. As we know, the average capacity of CD and DVD for storing data is approx. 700 MB and 4.7 GB respectively, which results in hundreds and thousands of CD/ DVDs to store such data. Regular audit to ensure that the CDs or DVDs are in working condition, maintaining duplicate copies, refreshing, migrating, cataloging tasks are unmanageable due to shortage of human resource and lack of technical expertise. In this approach it is not possible to store the metadata of the digital objects and facilitate search and retrieval due to offline storage media. Therefore, in spite of the digitization efforts, accessibility of this data is questionable. The external storage media is also subjected to physical decay or degradation, damages caused due to mishandling and variations in temperature and weather conditions.

**Obsolete file Formats**

Proprietary, copyrighted, patented and closed source file formats and obsolete formats pose major threats for the cultural digital data e.g. we have come across lot of data stored in Kodak PCD format which is not supported by Kodak anymore. Large volume of cultural digital content in many organizations is like to be stored in such unsupported or obsolete formats which must be migrated quickly otherwise such data will be lost permanently.

**Proprietary Software Solutions**

We observed that organizations have relied upon proprietary software solutions for maintaining the catalogues. Such software solutions often store the information in their own proprietary format and do not permit exporting or migrating the cataloging metadata into open format. It becomes a major handicap for the data owners as they have to always depend on the vendor. The data owners can lose entire data if the software license is not upgraded or the software is
discontinued in future. In such situation, C-DAC’s digital preservation technology development team has rescued the data by migrating it into open source and standardized formats.

**Lack of Manpower and Infrastructure**

Most of the cultural heritage institutions and museums do not have adequate computing infrastructure and the manpower with required technical skills to manage the activities like digital data, preservation, backup, etc. C-DAC is presently supporting the pilot digital repository at IGNCA within the scope of this project. However, long term support systems and human resource provisions are necessary for sustaining these efforts.

In order to address the digital preservation requirements at IGNCA, the team at C-DAC Pune is developing the following digital archival solutions as under.

**DIGITAL ARCHIVAL SYSTEM**

Therefore, the digital preservation technology development team at C-DAC Pune has initiated the development of Datāntar™ and Sanskriti Digitālaya™ systems based on the study of workflow, metadata requirements, file formats, current digital storage practices available in the cultural heritage organizations. The functional prototypes of these systems are already deployed at Indira Gandhi National Centre for the Arts, New Delhi for testing, feedback and continuous development as part of the pilot project. The solution will be productized and given a generic form to make it suitable for all the cultural heritage organizations. Datāntar™ and Sanskriti Digitālaya™ systems are developed using open source technologies and allow collaborative participation of the staff through local network.

**Datāntar™ System**

Datāntar™ word is a fusion of two words from English and Hindi languages. ‘Data’ means digital object and in Hindi language ‘antar’ means transformation. Therefore, ‘Datāntar™’ means to convey transforming the data in order to make it preservable. Datāntar™ software provides a set of best practices such as content organization, watermarking, OCR, naming conventions, tagging, structuring, production of low resolution copy for dissemination, etc for image, audio and video formats. It delivers the valid Submission Information Package (SIP) as per the requirements of OAIS.
Sanskriti Digitālaya™ System

Before processing digital data through Sanskriti Digitālaya™ Archival System, data is required to be organized in the form of Submission Information Package (SIP) using Datāntar™ system. These SIPs are generated by Producers and sent to the Archival system. Producers are participating institutes having cultural digital data which deliver SIPs to the Archival system.

Sanskriti Digitālaya™ Archival System fully complies with ISO standard on Open Archival Information System (OAIS) reference model. If any digital content is need to be preserved then it should follow the guidelines of OAIS model. Under the OAIS model, digital content is transmitted to the Archival system in a form called a Submission Information Package (SIP). The system’s ingest process validate, accepts and analyze the SIP and archives its contents according to a specific preservation plan for long-term preservation. Our system produces the Archival Information Package (AIP) and manages to ensure its integrity, security and future accessibility. When a user requests for a particular digital object or a group of objects from the Sanskriti Digitālaya™ Repository Portal, then Dissemination Information Package (DIP) is configured comprising the object and relevant metadata, as per the assigned rights and privileges of the user. Sanskriti Digitālaya™ system is aimed at supporting all the file formats used in the cultural heritage domain such as images, audio, video, slides and manuscripts and metadata schemas like MARC21, MODS, Dublin Core, METS, etc.

When SIPs are coming to the Archival system for processing, they are to be validated first because some SIPs will have insufficient representation information or preservation description information to meet stringent AIP requirements. Validation is a process that ensures that the files which are being accepted into the Archival system comply with the standards. After successful validation of SIP metadata enrichment stage comes, so that SIP Creator can enrich descriptive metadata of SIP for better access. Collaborative framework is adopted in the system so that Archivist can check and approve the SIP sent by SIP Creator and after approval, SIP Creator can ingest that SIP. The ingest process of SIP includes various steps such as XML parsing, technical metadata extraction, fixity calculation, XML generation, indexing, packaging, etc. The system also allows defining preservation policies based on various technical parameters, administering the user accounts and accessing control for the material published on the repository portal.
Access through Repository Portal

Sanskriti Digitālaya™ Repository Portal is for accessing records present in repository through internet. In this system user can search records through basic search and advanced search. Fuzzy logic is used for searching, so that it find matches even when users misspell words or enter in only partial words for the search. If user wants to perform specific searching within required fields then Advanced Search and Command Search features can be used. In these all sorts of combinations of fields through logical operators like AND, OR & NOT can be made and appropriate results can be fetched. Search results can be filtered with fields like author, publisher etc. and appropriate result list will be displayed. User can select any record from result list and view its content with metadata. The repository portal facilitates the DIP with proper access controls.

Present Status

Presently the working prototypes of Datāntar™ and Sanskriti Digitālaya™ are deployed at IGNCA for trial ingesting. Current design of these systems is made to work specifically for the requirements at IGNCA. We will required to provide a generic set of features and productize both the systems in order to make them suitable for any cultural heritage organization in India and the neighboring Asian countries.
CONCLUSION

We confront a common question whenever we interact with the experts from cultural heritage organizations that why to go for digital preservation if it is so fragile and unreliable? Therefore, in conclusion, we would like to summarize as under:

- The digital surrogates or replicas or copies of the physical objects of cultural significance are the best digital captures of those physical objects which are already deteriorating due to handling, progress of time, changing weather conditions. The digital copies can help in reducing the handling original physical objects.
- The digital copies can be easily made available to the scholars and researchers for their study across the globe with adequate access controls.
- Digital preservation of cultural digital data can certainly help towards keeping the cultures alive at least in the virtual sense for posterity, and Sanskriti Digitalayaa™ is aimed at contributing towards this objective.

ACKNOWLEDGEMENT

The authors of this paper acknowledge the support received from the digital preservation technology development team at C-DAC Pune namely Ritesh Malviya, Shrikant Salgar, Doppa Srinivas, Jaywant Karale, Sudhir Patel, Lakshmi Prasanna and Anshuma Acharya. We would also like to acknowledge the cooperation received from P. R. Goswami, Pratapanand Jha, and many other colleagues from IGNCA. We also acknowledge the support received from Dr. Ramesh Gaur when he was working at IGNCA. C-DAC acknowledges the funding support and encouragement received from Department of Electronics and Information Technology (DeitY), Government of India for the Centre of Excellence for Digital Preservation project, which is entrusted with Human-Centred Design & Computing Group at C-DAC, Pune.

REFERENCES

Collaborative Services through Websites of Institutional Repositories in India: An Evaluation

Vaishali B. Wadnerkar¹ and Minakshi R. Sontakke²
¹Librarian, Dhanwate National College, Nagpur
²Librarian, Institute of Design Education & Architectural Studies, Nagpur
E-mail: ¹vaishu.wadnerkar@rediffmail.com, ²minu.sontakke@gmail.com

Abstract—The technological advances today make it possible to think in terms of storing all the knowledge of the human race in digital form and several organizations worldwide are experimenting with less-expensive ways to create Institutional Repositories for long-term preservation of our knowledgebase and cultures. At present many institutions have repositories and maintain much of its own websites and provide all kinds of information about the functioning. This paper is focused on “An Evaluation of Websites of Institutional Repositories in India” This paper aims to study and analyze the Websites of Institutional Repositories in India. These websites of Institutional Repositories reveal various facets such as services, and facilities available on Institutional Repositories in India i.e.54. The evaluation process is achieved by grading and ranking. Out of these, some Institutional Repositories websites are ranked and finally a model of Institutional Repositories is given.

Keywords: Institutional Repositories, Digital Repositories, E-Archives, DSpace

INTRODUCTION

Digital technology can make all the significant literary, artistic, and scientific works of mankind permanently digitally preserved and made accessible to the billions of people all over the world. The technological advances today make it possible to think in terms of storing all the knowledge of the human race in digital form and several organizations worldwide are experimenting with less-expensive ways to archive and disseminate scholarly information and in-house knowledge as Institutional Repositories. For example, the California Digital Library’s e-Scholarship program at the University of California, ArXiv.org at Cornell University etc support open-access distribution of scholarship in cross-disciplinary through subject-based approach. U.S. Library of Congress’ US $100 million National Digital Information Infrastructure and Preservation Program (NDIIPP), which is developing a standard way for institutions to preserve their digital archives. Librarians and computer scientists are working together at MIT for Dspace digital asset management system, which aims to create an institutional repository that will include digitized versions of lecture notes, videos, papers, and data sets—in short, everything produced by faculty and staff.
Institutional Repository

An Institutional repository is a digital archive of the intellectual output of an Institution. It makes the quality and breath of scholarship produced at that Institution accessible to others worldwide over the Internet. It is a set of services that a University/Organization offers to the members of its community for the management and dissemination of digital material created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of the digital materials including long term preservation. An effective Institutional repository of necessity represents collaboration among librarians, Information technologies, archives and record managers, faculty and University administrators and policy makers. It is a new channel for structuring the organization’s contribution to the border world and as such invites policy and cultural reassessment of this relationship. It also form part of a larger global system of repositories which indexed in standardizes way.

Institutional Repositories in India

Nowadays, it is a trend of Institutional repositories all around. Many of the institutions have their repositories which they have built on various open source software. The saddest part of this is that various institutes had created these digital repositories for testing or trial purpose only and could not maintain the pace to streamline those. Many of them are closed and are not being updated regularly. Following is the list of institutional repositories from India which are currently active on the Internet.

Need of the Study

In this information revolution age, online information services provided every day, every moment in all field of human activity. This study attempts to explore coverage of Institutional Repositories on internet. Taking in to consideration the scope and utility of this topic, undoubtedly, this is surely beneficial to all libraries and I.T professionals in this study.

Objectives of the Study

1. To study the Websites of Institutional Repositories available on Internet in India.
2. To study data available on Homepage of Institutional Repositories in India.
3. To evaluate the Websites of Institutional Repositories in India.
4. To evaluate the ranking of top of Institutional Repositories in India.

Scope and Limitations of the Study

An Evaluation of Institutional Repositories in India includes a study based on the information available on website of repositories. In India, there are 54 Institutional Repositories available on Internet, out of these; the study is restricted to 45 Institutional Repositories. The study is all about the data available on Homepage of Institutional Repositories in India. The study is limited to 45 Institutional Repositories in India.
Methodology

Data is collected by using the websites of Institutional Repositories available on Internet. It was designed to find out the data available on the Institutional Repositories homepage. The data has been analyzed and interpreted.

**FINDINGS OF THE STUDY**

The table 1 shows that, State wise distribution of Institutional repositories in India out of these 54 institutional repositories in India, 9 Institutional repositories are not active properly. So, 45 institutional repositories are active properly on the Internet and the maximum no. of repositories in India are in Delhi-8, Karnataka-8 and Maharashtra-7.

<table>
<thead>
<tr>
<th>Name of the State</th>
<th>No. of IR</th>
<th>Not Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>08</td>
<td>01</td>
</tr>
<tr>
<td>Goa</td>
<td>02</td>
<td>02</td>
</tr>
<tr>
<td>Gujrat</td>
<td>04</td>
<td></td>
</tr>
<tr>
<td>Jammu &amp; Kashmir</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>Karnataka</td>
<td>10</td>
<td>02</td>
</tr>
<tr>
<td>Kerala</td>
<td>07</td>
<td>01</td>
</tr>
<tr>
<td>Karnataka</td>
<td>10</td>
<td>02</td>
</tr>
<tr>
<td>Maharashtra</td>
<td>07</td>
<td>---</td>
</tr>
<tr>
<td>Meghalaya</td>
<td>01</td>
<td>---</td>
</tr>
<tr>
<td>Orisa</td>
<td>02</td>
<td>---</td>
</tr>
<tr>
<td>Pondicherry</td>
<td>01</td>
<td>--</td>
</tr>
<tr>
<td>Punjab</td>
<td>01</td>
<td>--</td>
</tr>
<tr>
<td>Tamilnadu</td>
<td>02</td>
<td>--</td>
</tr>
<tr>
<td>Uttarakhand</td>
<td>02</td>
<td>01</td>
</tr>
<tr>
<td>Uttar Pradesh</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>West Bengal</td>
<td>05</td>
<td>--</td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>09</td>
</tr>
</tbody>
</table>

Table 1: State Wise Distribution of Institutional Repositories in India

Graph 1: Statewise Distribution of Institutional Repositories in India
It has been observed that out of 54, 11(20.37%) IR provides complete contact details on Homepage. Phone No. by 6(11.11%), Fax 4(7.40%), Email Address 19(35.18%), Separate Contact 12(22.22%), Last Update Date is given on Homepage 6(11.11%) and 45(87.03%) Institutional Repositories have website with their links.

Table 2: Contact Details on Homepage of Institutional Repository in India

<table>
<thead>
<tr>
<th>Criteria</th>
<th>No. of Institutional Repository Having Facility</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address on HP</td>
<td>11</td>
<td>20.37%</td>
</tr>
<tr>
<td>Phone No.</td>
<td>6</td>
<td>11.11%</td>
</tr>
<tr>
<td>Fax</td>
<td>4</td>
<td>7.40%</td>
</tr>
<tr>
<td>Email Address</td>
<td>19</td>
<td>35.18%</td>
</tr>
<tr>
<td>Separate Contact</td>
<td>12</td>
<td>22.22%</td>
</tr>
<tr>
<td>Last Update Date</td>
<td>6</td>
<td>11.11%</td>
</tr>
<tr>
<td>Total No. of Link</td>
<td>45</td>
<td>87.03%</td>
</tr>
</tbody>
</table>

Graph 2: Contact Details on Homepage of Institutional Repository in India

It is observed that out of 54 Institutional Repositories, 37(68.51%) provide home link, 11(20.37%) Login link, 8(14.81%) Register link, 1(1.85%) forgot password link, 1(1.85%) year link, 11(20.37%) contact us link, 33(61.11%) feedback link, 29(53.70%) RSS Feeds link, 1(1.85%) university and Department link, 1(1.85%) Dictionary link, 1(1.85%) distance Learning Centre link, 1(1.85%) Electronic Document Delivery Centre link, 1(1.85%) Rajyasabha Official Debates link, 1(1.85%) Debates Title Wise link, 1(1.85%) Member Participated Wise link, 1(1.85%) Debates Date Wise link, 1(1.85%) Debates title subject wise link, 7(12.96%) Browse link, 23(42.59%) title link, 39(72.22 %) Author link, 9(16.66) By Date link, 2(3.70%) Keywords link, 11(20.37%) publication link, 3(5.55%) have other links, 1 (1.85%) have related link, 2(3.70%) Multimedia link, 3(5.55%) Select Language link, 42(77.77%) Search link, 38(70.37%) Advanced Search link, 2(3.70%) recent Event link,
1(1.85%) Welcome link, 1(1.85%) Sitemap link, 31(57.40%) Subject Search link, 1(1.85%) thesis Guide link, 12(22.22%) Thesis submission date, 23(42.59%) Communication and Collection link, 35(64.81%) Issue Date link, 10(18.81%) Type link, 1(1.85%)ISI link, 1(1.85%)Gallery link, 1(1.85%)Subject link, 4(7.40%) Subject Link, 1(1.85%)Event link, 1(1.85%)LDL link, 1(1.85%)Statistics link, 1(1.85%)Sign on link, 30(55.55%)Receive Email Update link, 33(61.11%)MY Dspace>Edit Profile, 5(9.25%) Research link, 1(1.85%) Calendar link, 1(1.85%) Education Link, 1(1.85%) People Link, 32(59.25%) Help link, 32(59.25%) About link, 1(1.85%) Vision link, 1(1.85%) E-Resources link, 1(1.85%) My Account Link, 1(1.85%) CC License link, 3(5.55%) Recent Submission link, 1(1.85%) Last Update link, 1(1.85%) Webmaster link, 2(3.70%) Homepage Visit link, 2(3.70%) Research Journals link, 1(1.85%) NISCAIR Online Repository link, 1(1.85%) User Guide link, 1(1.85%) Disclaimer link, 4(7.40%) Conference proceeding link, 1(1.85%) Press Media link, 2(3.70%) Copyright link, 1(1.85%) Subscription link, 1(1.85%) Dictionary link, 1(1.85%) Recent Addition link, 1(1.85%) Administrator link, 1(1.85%) Top Contributing Universities link, 1(1.85%) Recognition link, 1(1.85%) UGC Notification link, 1(1.85%) Guest book link.

Table 3: Analysis of Content Coverage on Homepage of Institutional Repository in India

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Criteria</th>
<th>No. of Institutional Repository having Facility</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Home</td>
<td>37</td>
<td>68.51%</td>
</tr>
<tr>
<td>2</td>
<td>Login</td>
<td>11</td>
<td>20.37%</td>
</tr>
<tr>
<td>3</td>
<td>Register</td>
<td>8</td>
<td>14.81%</td>
</tr>
<tr>
<td>4</td>
<td>Forgot Password</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>5</td>
<td>Year</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>6</td>
<td>Contact Us</td>
<td>11</td>
<td>20.37%</td>
</tr>
<tr>
<td>7</td>
<td>Feedback</td>
<td>33</td>
<td>61.11%</td>
</tr>
<tr>
<td>8</td>
<td>RSS Feeds</td>
<td>29</td>
<td>53.70%</td>
</tr>
<tr>
<td>9</td>
<td>Univ &amp; Department</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>10</td>
<td>Dictionary</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>11</td>
<td>Distance Learning Centre</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>12</td>
<td>Electronic Document Delivery Centres</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>13</td>
<td>Rajyasabha Official Debates</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>14</td>
<td>Debates Title Wise</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>15</td>
<td>Member Participated Wise</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>16</td>
<td>Debates Date Wise</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>17</td>
<td>Debates Title Subject Wise</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>18</td>
<td>Browse</td>
<td>7</td>
<td>12.96%</td>
</tr>
<tr>
<td>19</td>
<td>Title</td>
<td>23</td>
<td>42.59%</td>
</tr>
<tr>
<td>20</td>
<td>Authors</td>
<td>39</td>
<td>72.22%</td>
</tr>
<tr>
<td>21</td>
<td>By Date</td>
<td>9</td>
<td>16.66%</td>
</tr>
<tr>
<td>22</td>
<td>Keywords</td>
<td>2</td>
<td>3.70%</td>
</tr>
<tr>
<td>23</td>
<td>Publications</td>
<td>11</td>
<td>20.37%</td>
</tr>
<tr>
<td>24</td>
<td>Other Links</td>
<td>3</td>
<td>5.55%</td>
</tr>
<tr>
<td>25</td>
<td>Related Links</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>26</td>
<td>Multimedia</td>
<td>2</td>
<td>3.70%</td>
</tr>
<tr>
<td>27</td>
<td>Select Language</td>
<td>3</td>
<td>5.55%</td>
</tr>
<tr>
<td>28</td>
<td>Search</td>
<td>42</td>
<td>77.77%</td>
</tr>
<tr>
<td>29</td>
<td>Advanced Search</td>
<td>38</td>
<td>70.37%</td>
</tr>
</tbody>
</table>

Table 3 (Contd.)...
### Table 3 (Contd.)

| 30 | Recent Event | 2 | 3.70% |
| 31 | Welcome       | 1 | 1.85% |
| 32 | Sitemap       | 1 | 1.85% |
| 33 | Subject Search| 31| 57.40%|
| 34 | Thesis Guide  | 1 | 1.85% |
| 35 | Thesis Submission Date | 12 | 22.22%|
| 36 | Communications & Collections | 23 | 42.59%|
| 37 | Issue Date    | 35 | 64.81%|
| 38 | Type          | 10 | 18.51%|
| 39 | Who Doc.      | 1 | 1.85% |
| 40 | IS            | 1 | 1.85% |
| 41 | Gallery       | 1 | 1.85% |
| 42 | Subject       | 1 | 1.85% |
| 43 | Project       | 4 | 7.40% |
| 44 | Event         | 1 | 1.85% |
| 45 | LDL           | 1 | 1.85% |
| 46 | Statics       | 1 | 1.85% |
| 47 | Sign On       | 1 | 1.85% |
| 48 | Receive Email Updates | 30 | 55.55%|
| 49 | My Dspace Edit Profile | 33 | 61.11%|
| 50 | Research      | 5 | 9.25% |
| 51 | Calendar      | 1 | 1.85% |
| 52 | Education     | 1 | 1.85% |
| 53 | People        | 1 | 1.85% |
| 54 | Help          | 32| 1.85% |
| 55 | About         | 32| 59.25%|
| 56 | Vision        | 1 | 1.85% |
| 57 | E-Resources   | 1 | 1.85% |
| 58 | My Account    | 1 | 1.85% |
| 59 | CC Licence    | 1 | 1.85% |
| 60 | Recent Submission | 3 | 5.55% |
| 61 | Last Update   | 1 | 1.85% |
| 62 | Webmaster     | 1 | 1.85% |
| 63 | Homepage VISIT | 2 | 3.70% |
| 64 | Research Journals | 2 | 3.70%|
| 65 | NISCAIR Online Repository | 1 | 1.85%|
| 66 | User Guide    | 1 | 1.85% |
| 67 | Disclaimer    | 1 | 1.85% |
| 68 | Conference Proceeding | 4 | 7.40% |
| 69 | Press Media   | 1 | 1.85% |
| 70 | Copyright     | 2 | 3.70% |
| 71 | Subscription  | 1 | 1.85% |
| 72 | Dictionary    | 1 | 1.85% |
| 73 | Recent Addition | 1 | 1.85%|
| 74 | Administrator | 1 | 1.85% |
| 75 | Top Contributing Universities | 1 | 1.85%|
| 76 | Recognition   | 1 | 1.85% |
| 77 | UGC Notification | 1 | 1.85%|
| 78 | Guestbook     | 1 | 1.85% |
It is observed that out of 54 institutional repositories, 45 (83.33%) websites have included the information of the resolution of the screen and the browser in which the site can be best viewed on their Home Pages. 44 (81.48%) websites used images on Home Page. 44 (81.48%) Homepages links are clearly labeled. 45 (83.33%) websites have printed pages separately. 45 (83.33%) have scrolling web-pages. Out of 54, 37 (68.51%) show Homepage link on each page. 2 (3.70%) websites have Sitemap, 1 (1.85%) websites have keyword searching. All websites have scrolling pages.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Criteria</th>
<th>No. of Institutional Repository having Facility</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Best Viewed</td>
<td>45</td>
<td>83.33%</td>
</tr>
<tr>
<td>2</td>
<td>Use of Images</td>
<td>47</td>
<td>87.03%</td>
</tr>
<tr>
<td>3</td>
<td>Link Clearly labeled</td>
<td>46</td>
<td>85.18%</td>
</tr>
<tr>
<td>4</td>
<td>Pages Printed Separately</td>
<td>47</td>
<td>87.03%</td>
</tr>
<tr>
<td>5</td>
<td>Home link on each page</td>
<td>37</td>
<td>68.51%</td>
</tr>
<tr>
<td>6</td>
<td>sitemap</td>
<td>1</td>
<td>1.85%</td>
</tr>
<tr>
<td>7</td>
<td>Keyword Searching</td>
<td>2</td>
<td>3.70%</td>
</tr>
<tr>
<td>8</td>
<td>Scrolling Page</td>
<td>45</td>
<td>83.33%</td>
</tr>
</tbody>
</table>
Table 5: Institutional Repositories in India

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Name of Institutional Repository Criteria</th>
<th>Place</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IGNOU</td>
<td>Delhi</td>
<td><a href="http://www.egyankosh.ac.in/">http://www.egyankosh.ac.in/</a></td>
</tr>
<tr>
<td>2</td>
<td>Indian Institute of Technology, Delhi</td>
<td>Delhi</td>
<td>eprint.iitd.ac.in</td>
</tr>
<tr>
<td>3</td>
<td>Indraprastha Institute of Information Technology</td>
<td>Delhi</td>
<td><a href="https://repository.iitd.edu.in/jsui/">https://repository.iitd.edu.in/jsui/</a></td>
</tr>
<tr>
<td>4</td>
<td>NISCAIR-National Science Digital Library</td>
<td>Delhi</td>
<td><a href="http://nsdl.niscair.res.in/handle/123456789/939">http://nsdl.niscair.res.in/handle/123456789/939</a></td>
</tr>
<tr>
<td>6</td>
<td>Parliament of India, Official debates of Rajyasabha</td>
<td>Delhi</td>
<td><a href="http://rsdebate.nic.in/">http://rsdebate.nic.in/</a></td>
</tr>
<tr>
<td>7</td>
<td>World Health Organisation South East Asian Region</td>
<td>Delhi</td>
<td><a href="http://repository.searo.who.int/">http://repository.searo.who.int/</a></td>
</tr>
<tr>
<td>8</td>
<td>Inflibnet- Institutional Repository</td>
<td>Gujrat</td>
<td><a href="http://ir.inflibnet.ac.in/">http://ir.inflibnet.ac.in/</a></td>
</tr>
<tr>
<td>9</td>
<td>Inflibnet- Institutional –Shodhganga- Indian ETD</td>
<td>Gujrat</td>
<td><a href="http://shodhganga.inflibnet.ac.in/">http://shodhganga.inflibnet.ac.in/</a></td>
</tr>
<tr>
<td>10</td>
<td>M.S University of Baroda</td>
<td>Gujrat</td>
<td><a href="http://14.139.121.106:8080/jsui/">http://14.139.121.106:8080/jsui/</a></td>
</tr>
<tr>
<td>11</td>
<td>Pandit Deendayal Petroleum University, Ganghinagar</td>
<td>Gujrat</td>
<td><a href="http://203.77.192.116:8080/xmlui/">http://203.77.192.116:8080/xmlui/</a></td>
</tr>
<tr>
<td>12</td>
<td>Indian Institute of Astrophysics, Bangalore</td>
<td>Karnataka</td>
<td><a href="http://prints.iiaap.res.in/">http://prints.iiaap.res.in/</a></td>
</tr>
<tr>
<td>13</td>
<td>Indian Institute of Horticultural Research</td>
<td>Karnataka</td>
<td><a href="http://www.erepo.iihr.ernet.in/">http://www.erepo.iihr.ernet.in/</a></td>
</tr>
<tr>
<td>14</td>
<td>Indian Institute of Science, Bangalore</td>
<td>Karnataka</td>
<td><a href="http://etd.ncsi.iisc.ernet.in/">http://etd.ncsi.iisc.ernet.in/</a></td>
</tr>
<tr>
<td>15</td>
<td>Indian Statistical Institute, Bangalore</td>
<td>Karnataka</td>
<td><a href="https://dspace.gipe.ac.in/">https://dspace.gipe.ac.in/</a></td>
</tr>
<tr>
<td>17</td>
<td>KLE University</td>
<td>Karnataka</td>
<td><a href="http://182.18.8080/jsui/Karnataka">http://182.18.8080/jsui/Karnataka</a></td>
</tr>
<tr>
<td>18</td>
<td>Raman Research Institute, Bangalore</td>
<td>Karnataka</td>
<td><a href="http://dspace.rrri.res.in/">http://dspace.rrri.res.in/</a></td>
</tr>
<tr>
<td>19</td>
<td>University of Mysore –Vidyamithi Project</td>
<td>Karnataka</td>
<td><a href="http://dspace.vidyamithi.org.in:8080/dspace/">http://dspace.vidyamithi.org.in:8080/dspace/</a></td>
</tr>
<tr>
<td>20</td>
<td>Cochin University of Science and Technology-Digital</td>
<td>Kerala</td>
<td><a href="http://dspace.cusat.ac.in/jsui/">http://dspace.cusat.ac.in/jsui/</a></td>
</tr>
<tr>
<td>21</td>
<td>Cochin University of Science and Technology-Dyuthi</td>
<td>Kerala</td>
<td><a href="http://dysuthi.cusat.ac.in/xmlui/">http://dysuthi.cusat.ac.in/xmlui/</a></td>
</tr>
<tr>
<td>22</td>
<td>Indian Institute of Management, Kozhikode</td>
<td>Kerala</td>
<td><a href="http://dspace.imrk.ac.in/">http://dspace.imrk.ac.in/</a></td>
</tr>
<tr>
<td>23</td>
<td>Indian Institute of Spices Research</td>
<td>Kerala</td>
<td><a href="http://dspace.imrk.ac.in/">http://dspace.imrk.ac.in/</a></td>
</tr>
<tr>
<td>24</td>
<td>School of Communication and Management Studies (SRM Group)</td>
<td>Kerala</td>
<td><a href="http://dspace.sctimst.ac.in/jsui/">http://dspace.sctimst.ac.in/jsui/</a></td>
</tr>
<tr>
<td>25</td>
<td>Sree Narayana Gurukulam College of Engineering</td>
<td>Kerala</td>
<td><a href="http://117.239.73.132/">http://117.239.73.132/</a></td>
</tr>
<tr>
<td>26</td>
<td>Godhule Institute of Politics and Economics</td>
<td>Maharashtra</td>
<td><a href="http://dspace.gipe.ac.in/jsui/">http://dspace.gipe.ac.in/jsui/</a></td>
</tr>
<tr>
<td>27</td>
<td>Indian Institute of Technology, Bombay</td>
<td>Maharashtra</td>
<td><a href="http://dspace.library.iitb.ac.in/jsui/">http://dspace.library.iitb.ac.in/jsui/</a></td>
</tr>
<tr>
<td>28</td>
<td>Indira Gandhi Institute of Development Research</td>
<td>Maharashtra</td>
<td><a href="http://dspace.iitd.ac.in:8080/jsui/">http://dspace.iitd.ac.in:8080/jsui/</a></td>
</tr>
<tr>
<td>29</td>
<td>Inter-University Centre for Astronomy and Astrophysics</td>
<td>Maharashtra</td>
<td><a href="http://www.iucaa.ernet.in:8080/jsui/">http://www.iucaa.ernet.in:8080/jsui/</a></td>
</tr>
<tr>
<td>30</td>
<td>National Centre for Radio Astrophysics</td>
<td>Maharashtra</td>
<td><a href="http://ncraib1.ncra.tifr.res.in:8080/jsui/">http://ncraib1.ncra.tifr.res.in:8080/jsui/</a></td>
</tr>
<tr>
<td>31</td>
<td>Urban Aspirations in Global Cities</td>
<td>Maharashtra</td>
<td><a href="http://uagc.tiss.edu/jsui/">http://uagc.tiss.edu/jsui/</a></td>
</tr>
<tr>
<td>32</td>
<td>Vidya Prasarak Mandal</td>
<td>Maharashtra</td>
<td><a href="http://dspace.vpmthane.org:8080/jsui/index.jsp">http://dspace.vpmthane.org:8080/jsui/index.jsp</a></td>
</tr>
<tr>
<td>33</td>
<td>North-Eastern Hill University</td>
<td>Meghalaya</td>
<td><a href="http://dspace.nehu.ac.in/">http://dspace.nehu.ac.in/</a></td>
</tr>
<tr>
<td>34</td>
<td>Indian Institute of Technology, Roorkee</td>
<td>Orissa</td>
<td><a href="http://bhagirathi.iitr.ac.in/dspace/">http://bhagirathi.iitr.ac.in/dspace/</a></td>
</tr>
<tr>
<td>35</td>
<td>National Institute of technology, Rourkela (NITR)</td>
<td>Orissa</td>
<td><a href="http://dspace.nitkri.ac.in/dspace/">http://dspace.nitkri.ac.in/dspace/</a></td>
</tr>
<tr>
<td>36</td>
<td>Pondicherry University</td>
<td>Pondicherry</td>
<td><a href="http://dspace.pondiuni.edu.in/jsui/">http://dspace.pondiuni.edu.in/jsui/</a></td>
</tr>
<tr>
<td>37</td>
<td>Thapar University, Patiala</td>
<td>Punjab</td>
<td><a href="http://dspace.thapar.edu:8080/dspace/">http://dspace.thapar.edu:8080/dspace/</a></td>
</tr>
</tbody>
</table>

Table 5 (Contd.)...
...Table 5 (Contd.)

<table>
<thead>
<tr>
<th>No.</th>
<th>Institute Name</th>
<th>State</th>
<th>Repository URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>Institute of Mathematical Sciences</td>
<td>Tamil Nadu</td>
<td><a href="http://www.imsc.res.in/xmlui">http://www.imsc.res.in/xmlui</a></td>
</tr>
<tr>
<td>39</td>
<td>SRM University</td>
<td>Tamil Nadu</td>
<td><a href="http://digitallibrary.srmuniv.ac.in/dspace/">http://digitallibrary.srmuniv.ac.in/dspace/</a></td>
</tr>
<tr>
<td>40</td>
<td>Aryabhatta Research Institute of Observational Sciences</td>
<td>Uttarkhand</td>
<td><a href="http://210.212.91.105:8080/ispui/handle/1234567891">http://210.212.91.105:8080/ispui/handle/1234567891</a></td>
</tr>
<tr>
<td>41</td>
<td>Anwesan @ SNLTR by CSE, IIT Kharagpur</td>
<td>West Bengal</td>
<td><a href="http://anwesan.iitkgp.ernet.in:8080/ispui/">http://anwesan.iitkgp.ernet.in:8080/ispui/</a></td>
</tr>
<tr>
<td>42</td>
<td>Indian Association for the Cultivation of Sciences</td>
<td>West Bengal</td>
<td><a href="http://arxiv.iacs.res.in:8080/ispui/">http://arxiv.iacs.res.in:8080/ispui/</a></td>
</tr>
<tr>
<td>43</td>
<td>Indian Statistical Institute Kolkata</td>
<td>West Bengal</td>
<td><a href="http://library.isical.ac.in/ispui/">http://library.isical.ac.in/ispui/</a></td>
</tr>
<tr>
<td>44</td>
<td>Jadavpur University</td>
<td>West Bengal</td>
<td><a href="http://dspace.jdvu.ac.in/">http://dspace.jdvu.ac.in/</a></td>
</tr>
<tr>
<td>45</td>
<td>West Bengal Public Library Network</td>
<td>West Bengal</td>
<td><a href="http://dspace.wbpublibnet.gov.in:8080/ispui/">http://dspace.wbpublibnet.gov.in:8080/ispui/</a></td>
</tr>
</tbody>
</table>

REFERENCES

[3] http://dspace.inflibnet.ac.in
Development of Digital Archives
Digital Preservation and Permanent Access to Print Media Resources at the TAKC, India: A Case Study

R. Venkata Kesavan
DGM & Head, Times Archives & Knowledge Centre (TAKC),
Dr. D.N. Road, Mumbai-400001, India
E-mail: venkata.kesavan@timesgroup.com

Abstract—Digital preservation and digital archiving of print media resources in a newspaper organization is a challenging task as well as a rewarding opportunity. This paper makes an attempt to present the case study of the various digital preservation initiatives taken up and implemented by the Times Archives and Knowledge Centre (TAKC) of The Times of India Group. Apart from briefly profiling the TAKC and the features of its archival collection, this study lucidly presents the details of its various digital preservation activities, by highlighting the print media resources digitized; features of the digital archival contents; digital archiving processes; access facilities; usage of digital archives; challenges in digital archiving; opportunities and benefits of digitization, etc. Further, this paper, based on the successful digitization experiences at the TAKC, highlights a number of best practices, which could be useful for similar kind of digitization projects in any newspaper archives and libraries.

INTRODUCTION

The worldwide, over the last two decades, several archives, libraries and museums have undertaken numerous digital preservation projects of their collection of various archival information resources and artifacts [1,2]. If cultural institutions are to extend their missions of preserving society’s cultural heritage and intellectual capital into the digital age they need more reliable and more affordable digital archiving methods, systems, and technologies [3]. It has been observed that the basic objectives of such digitization projects are to ensure long-term preservation of these heritage archival resources for ensuring permanent access and enhanced use. On a more positive note, digital archives can be far larger, more diverse, and more inclusive than traditional archives [4]. Thus, digitization technologies, as long-term preservation tools of print media archival resources, have become an integral part of modern day newspaper libraries and archives. In a newspaper organization, digital preservation technologies are being used, not only to preserve, archive and retrieve the current born-digital contents of newspapers, magazines, photographs, cartoons, etc., but also being used for digitization of archival newspapers collection from their original hard copies as well as microfilm formats. Digitization raises the prospect of a far greater and more sophisticated
engagement with newspaper content [5] and digital archive enables in providing online access to multiple users from anywhere in the world at any time. As stated in an Oracle Corporation whitepaper a digital archive 'enables longer-term storage, maintaining content for decades and provides very high data integrity, helping eliminate loss of content [6]. A number of digital preservation projects, funded by national libraries, archives and independent media organizations, covering major historical newspapers in USA, UK, Australia and India have already become successful and been adding value to their user community. The digital archives of newspaper organizations, apart from adding value to their respective newspaper journalists are also available for the general public to access them, as a valuable historical research resource. On the one hand, the digital preservation has opened up a number of new opportunities for the print media libraries and archives to exploit their archival resources for the benefit of their user groups and on the other hand, such digital preservation projects also present quite a few challenges. In a digital archive environment there are concerns about content accessibility and service sustainability, ease of use, interoperability, usefulness, efficiency, etc. [7]. This paper, by keeping the various digital preservation initiatives of print media resources taken up by the Times Archives and Knowledge Centre (TAKC) of The Times of India Group, as the organizational reference point, makes an attempt to present a case study, covering the details of the various digital preservation activities handled by TAKC, by highlighting the print media resources digitized, features of the digital archival contents, digital archiving processes and access facilities, usage of digital archives, challenges in digital archiving; opportunities and benefits of digitization, etc. Further, this paper, based on the successful digitization experiences at the TAKC, highlights a number of best practices, which could be useful for similar newspaper archives digitization projects.

About the Times of India Group and TAKC

Bennett, Coleman & Co. Ltd. popularly known as The Times of India Group (http://www.timesgroup.com/) is the largest media house in India that has revolutionized the world of information and entertainment through innovative thinking, winning strategies and global management practices. The 175-year-old Times Group is a market driver across a number of media platforms and it is one of the most respected business houses in India. The Group’s leading print media brands include: newspapers such as The Times of India, The Economic Times, Navbharat Times, Maharashtra Times, Mumbai Mirror and magazines like Femina and Filmfare. The Times Group having a pan India presence employs over 10,000 people and reaches to millions of consumers through its various media initiatives. In supporting the above business activities of the Group, the Times Archives and Knowledge Centre (TAKC) plays a very vital role using its vast collection of current as well as archival print media information resources. TAKC, as a newspaper reference dept., was set up over a century ago in Mumbai (then Bombay), and has over the years gradually expanded to Delhi, Ahmedabad, Bangalore, Pune and other cities in India. The TAKC - Mumbai houses, one of the largest newspaper archives in Asia, acts as the central knowledge centre of the Group with a collection of over sixty current and erstwhile publications, the earliest being the flagship newspaper brand of the Group, The Times of India published since November 3,
1838. The TAKC in its collection has over 5 million unique print media resources, covering newspapers, press clippings, photographs, cartoons, etc. TAKC also houses a well-equipped reference library, which provides access to a special collection of over 10000 books, 100s of current newspapers, magazines and 1000s of CDs, DVDs and other audio visual resources. TAKC with its vast collection of specialized print media information resources, offers a range of value-added information services primarily to the in-house users comprising the Editorial staff and various business information users ranging from marketing, advertising, sales professionals, etc.. Often, TAKC based on special information needs, caters to external users like students, researchers, legal professionals, govt. departments and other general public. For commercial use of TAKC resources, the external customers are directed to the Times Syndication Service, a division of The Times Group, which looks after its media contents marketing and syndication arrangements, located at New Delhi.

**ARCHIVAL RESOURCES AND COLLECTION MANAGEMENT AT TAKC**

TAKC selects, collects, processes and preserves a wide range of media resources in the form of multimedia contents. Some of these resources are dating as far as back as 1830s. The archival collection include The Times of India Group’s print publications (preserved in the form of bound volumes, about 5 million pages), microfilm collection of publications (about 2.5 million pages), photographs (over 2 million items), clippings of magazine articles and news items (over 1 million items, covering 6000 subject areas and 16,000 personalities from all walks of life) and a rich collection of original cartoons of eminent cartoonists, particularly of the legendary R.K. Laxman. As Brown and Fenton recommended microfilm is the prudent medium for long-term preservation and digitization the medium for access [8] and the TAKC, probably, is the only print media archive in this part of the world, that continues to produce microfilm collection of all of its major publications, in-house, despite setting up digital archives and e-paper archives. The reputed international libraries such as, The Library of Congress, USA, The British Library, UK and The National Library of Australia are some of the regular subscribers of TAKC’s microfilm collection. Collectively all these information resources form as the core base for the various information and knowledge sharing activities within the Group.

**DIGITAL PRESERVATION AND ARCHIVING AT THE TAKC**

Hedstrom defines ‘digital preservation as the planning, resource allocation, and application of methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable’ [9]. In this context, recent advances in digital technology have created the potential to make the vast stock of recorded knowledge searchable using sophisticated tools by anyone with an Internet connection [10]. Moreover, a digital archive or a digital library allows a user to preserve, search, retrieve, print, save, share and re-use the archived information, instantly [11]. Therefore, digital preservation requires new workflows, new skills and close co-operation across different professions ranging from traditional preservation management skills to computing science [12]. TAKC, over the years, has become a much sought after newspaper archives at national level to cater to the information...
needs of various external users on various subject domains and facets, covering the Country’s social, cultural, political, sporting, entertainment and economic history. Therefore, in order to preserve the vital archival resources for posterity, ensure permanent access and enable faster information storage and retrieval, the TAKC has begun digitization projects around the year 2000, in a small-scale, and since then, a number of projects have been initiated covering archival collection of newspapers, photographs, press clippings, cartoons, etc. The details of these projects are enumerated in the following sub-sections.

**Digitization of Archival Newspapers**

In a newspaper organization, when it comes to launching digitization projects, among the various print media resources, the priority is normally given to its archival newspaper collection. For the purpose of digitization of archival collection of newspapers, most organizations prefer to use their microfilm collection. Because, it is possible now to plan to have microfilms scanned digitally, with the digital images capable of further manipulation by software, to permit a good degree of readability, with the texts also being searchable [13]. At the TAKC, since 2000, so far, over one million pages from newspapers collections (including the complete collection of *The Times of India*) have been digitised and preserved in a customized digital archive. *The Times of India* digital archive, covering (since its inception) from November 3, 1838 to December 31, 2003 have been digitised from its source microfilm collection and the project was handled between 2009 and 2011. A complete set of high quality microfilm positives were produced from master negatives and exclusively used for the purpose of digitization. While these microfilm positive copies were produced in-house, the entire process of digitization was outsourced to a commercial agency. *The Times of India* digital archive, covering the period 1838-2003 are available for The Times Group internal users through a customized digital archive management system, strategically integrated with its corporate Intranet called Timescape and the same collection is made available to the external users through ProQuest Historical Newspapers ™ collection [14] on the Internet.

**Born-digital E-paper Archive**

In a latest development, the print and online editions have been joined by a third version of the daily newspaper, namely the so-called E-Paper version, also referred to as the “digital edition” [15]. E-paper helps a newspaper organization to expand its readership at the global level in reaching out to new markets and news consumers. With these kinds of objectives, from 2004 the entire collection (born-digital) of the two main publications of the Times Group, namely, *The Times of India* and *The Economic Times* (including their supplements) are available in the e-paper archives with an Intranet based access for the Times Group internal users and the Internet-based public access for external users (free in India and fee based abroad). This e-paper archive facility can be accessed online at http://epaper.timesofindia.com. The e-paper archive provides access to complete newspapers, as they have been published in hardcopies (with text matters, images as well as ads) and it is being updated on a daily basis. The Times Group e-paper archives uses Olive Software (http://www.olivesoftware.com) and its runs on its Active Paper Archive™
components and entire work flow of e-paper archive is managed by a well-trained in-house team and updated on a daily basis. As during the last few years, the mobile phones and other handheld devices have become useful tools for accessing e-resources by users, spotting such trends, starting year 2008, the Times Group has launched its m-paper (mobile e-paper), which can be accessed at http://m.timesofindia.com, by using any kind of TABs and high-end cellular phones with Internet data connectivity.

Digital Preservation of Photographs Collection

Apart from digitizing archival newspaper collection, TAKC has also started digitizing and preserving its various precious and high-value print photographs from its hard copy collection of over one million photographs dating back to year 1900. Here for the digital preservation process, exclusively trained in-house human resources as well as custom set-up ICT facilities are being used. Occasionally, based on the special digitization needs, some amounts of outsourced resources were also used in the past. For the purpose of digital photo archiving, from year 2004, TAKC uses a proprietary web-based photographs content management tool called COMYAN Media System (http://www.comyan.com), which facilitates faster image processing (for editorial, production and printing work-flow processes) as well as archival storage and retrieval of photographs including the new ones (born-digital images) that are shot on a day-to-day basis by in-house journalists as well as images received thorough various agencies through syndication services. In the COMYAN photo archive, the internationally accepted IPTC (International Press Telecommunications Council- http://www.iptc.org) ‘Photo Metadata’ headers are used as the metadata standards. The digital photo archive currently has over one million digital images, a majority of them born-digital (JPEG 300 DPI) with relevant metadata and is growing consistently by every passing day. The COMYAN digital photo archive is currently accessible by the Times Group employees through restricted access over Intranet through individual Login IDs. In order to cater to the external users, in the year 2009, a fee-based website called http://www.timescontents.com, has been launched with selected high quality images sourced from TAKC photo digital archives. Timescontents.com is also managed by the Times Syndication Service.

Digital Preservation of Press Clippings Collection

Within the print media industry, it is a universally known fact that the press clippings (current as well as archival), regularly tracked from the competitive publications are one of the vital form of information sources for the journalists. While the older press clippings in the TAKC collection (more than 500,000 items) are maintained in hard copy format in subject specific folders (on 6000 subjects and 15000 personalities), selected clippings (about 800000 items) published since 1996, on various subject areas and leading personalities (on the later aspect the coverage is from 1950 onwards in some cases) have been digitised. The digitised images are saved as single bit B & W BMP format and they are accessible on a proprietary intranet based content management tool called ‘DataScan Online’, which uses MySQL server for the back-end database and ASP.NET for front-end applications. As the full-fledged content
management tool, DataScan package is used for digitization of clippings as well as archival storage and retrieval of digitized contents. Metadata elements in this database are customized according to our local requirements and the key fields are record Id, classification No., clippings type, title, byline, source, date, pagination, subject, and keywords etc. The information resources archived in DataScan Online are only meant for internal users for reference and research purposes and strictly adhering to copyright restrictions, no external users are given access to this resource. In the electronic environment the need for access tools will be more evident, and users will expect appropriate and standard software to be readily available [16] and in this context DataScan overall fulfills, such expectations as the digital press clippings content management tool.

Digital Preservation of Cartoons Collection

Since 2006, yet another digitization project exclusively focused on the Cartoons Collection of TAKC has been started. All the original cartoons, about 20000, created by R.K. Laxman (leading Indian cartoonist, who has been associated with The Times of India, since 1954) and Mario Miranda (a popular cartoonist was engaged with The Economic Times during 1990s) have been digitised (150 DPI TIFF format greyscale images) and preserved in our local archives. All relevant metadata pertaining to digitised cartoons are being captured, based on a customized data structure in the lines of Dublin Core Metadata Standard. Currently, TAKC is in the process of developing a full-fledged web enabled, XML based digital archive content management tool, by benchmarking its features with BCAD (The British Cartoon Archive Digitization Project-http://www.cartoons.ac.uk), which is one of the best online examples of cartoons digitization projects, located at the University of Kent at Canterbury.

Usage of Digital Archives at the TAKC

A newspaper library and archival facility basically exists to meet the existing as well as potential information needs of its user groups. This principle is true with TAKC, as well. The journalists of the Times Group very frequently use TAKC information resources, and over the last couple of years, more often its digital archives of print publications, photographs, press clipping, etc. At the TAKC, it has been observed (through its quarterly MIS reports covering last seven years) that ever since the introduction of digital archival resources in the Organization, particularly since 2004, after launching the e-paper archive, the information seeking patterns of journalists have gradually changed. Mainly, the uses of in-house archival resources in digital form have increased substantially. To cite a few examples journalists have been frequently exploiting the digital archives for compiling background information on various ongoing issues and current affairs, covering in a range of subject areas such as politics, business, economy, social issues, crime, cultural activities, entertainment, sports, etc. The journalists also have been using the TAKC digital archives extensively for the other purposes such as, preparing special reports, historical anniversary publications, researching for feature articles; researching for current news items and follow up of stories, authentication and verification of factual information, etc. The ease of access to newspaper information resources through the digital archives have not only been benefiting the journalists of the
Times Group, but also the other internal business information users such as, advertising, marketing and sales professionals. This group of internal user have been using the TAKC digital resources, particularly, the e-paper archives for keeping track of clients related ads and media coverage. Besides, the TAKC team, using its digital archives resources also offer a number of pro-active business information services such as, compiling information on companies, brands, products, services, industry sectors, markets, business personalities, etc. and share such vital business information resources with the internal users, as value-added information services.

**CHALLENGES IN DIGITAL ARCHIVING OF PRINT MEDIA RESOURCES**

One of the biggest challenges in maintaining preservation information for a long period of time is the continuous change in storage and computing technologies [17]. In the context of ‘challenges of long-term digital archiving’ a technical report on the subject has identified failure modes of digital archiving, such as, - natural disaster; external attacks; internal attacks, components failure; storage media faults; human errors; media/hardware obsolesce; software/format obsolesce; organization fault; economical fault, etc. [18]. Moreover, the responsibility and cost for digital preservation is too great for any one person, institution, or marketplace to bear [19]. On the contents side, in terms of the sources used for digitization, early microfilming techniques sometimes created film with poorly focused images, uneven lighting, and dark spots. Poor film, in turn, produced poor digital images and usually decreased OCR accuracy [20] and some of these problems cited above are also encountered during the various digitization projects at the TAKC. On the one hand, the digital preservation initiatives implemented by TAKC have been beneficial to the Times Group and on the other hand, during the course of carrying out to the digital preservation projects and while continuing the maintenance of the digital archives, a number challenges and problems were encountered and handled by the TACK team. The following are the major challenges, which TAKC had to deal with:

- **Quality of the original archival resources**: Some of the original old newspaper collection archived in TAKC had physical damages, which resulted in poor quality microfilm collection. When some of these poor quality source microfilms were used for digitization, which again resulted in poor quality digital images, incomplete issues in the digital archives, inferior quality OCR text output etc. On the other hand, even some of the old photographs and press clippings were difficult to scan due to their poor conditions, such as wear and tear, incomplete metadata availability, etc.

- **Budgetary resources**: Like any other media archive, getting adequate funds for digitization has been a big challenge for TAKC. Since the ‘return-on-investment’ (ROI) takes such a long time to realize, this will be a critical factor in future digitization projects of TAKC.

- **ICT tools and facilities**: Identifying Suitable ICT facilities for digitization as well as hosting the digital archives on the Internet/Intranet continuous to be a challenge in terms of their cost as well as availability of robust technological solutions.
• **Metadata schema**: In the initial stage of digitization project life cycle, standardizing metadata schema for specific type of print media resources was an issue (e.g. specific metadata schema for archival newspapers, clippings, photographs, cartoons, etc.)

• **Manpower**: Since TAKC has been handling a majority of its digitization projects using in-house manpower, acquiring quality manpower, training and retaining them in specific digitization projects are some of the ongoing issues in this front.

• **Content management tools**: There is no single best digital archival system available in the market that would suit all kinds of print media resources. Therefore, TAKC has ended up using some of expansive proprietary content management tools for each type of print media resource and now integrating these tools on a common interface is not possible due to their inherent complexities.

• **End-users concerns**: Some of the TAKC internal users’ resistance for using digital resources continues to be an issue, which calls for motivation and training at the end-user level.

• **Legal issues**: Protection of copyrights of digital resources, preventing any kind of misuse of digitized archival resources, promoting fair use of resources etc. are some other challenges TAKC is trying to address.

**Benefits of Digitization of Print Media Resources**

The TAKC currently operates in a hybrid library and archive environment, where, it continues facilitating access to its archival resources in physical form and it has been increasingly facilitating access to digital form of print media archival resources (both digitized from physical collection as well as born-digital). This kind of hybrid newspaper library and archive environment offers a number of opportunities and benefits to the TAKC users as well as its staff. Digital collections are better because they are dynamic, changeable, constantly updatable and available any time, any place, anywhere [21]. Moreover, in a newspaper archive, digitization provides new opportunities to extend their utilization and thus avoid damage to these valuable historical archives [22]. Similar to these benefits, based on over a decade long experience in digital preservation of archival resources, TAKC is able to:

• Preserve its archival collection for posterity and ensure permanent access.

• Save the time of its staff as well as users in the processes of archival information storage, search and retrieval, etc.

• Offer simultaneous access to multiple users anywhere any time.

• Add value to improved editorial research work, re-purposed publications and new content generation process.

• Effectively market its archival information resources in content monetization and revenue generation activities by reaching out to external users.
• Spare the physical storage space in the library/ archive area, by shifting and storing a huge volume of physical form of original resources (e.g. bound volumes of back issues of newspapers, hard copy of photographs and folders of clippings collection, etc.) to a suitable less expensive, remote location.

BEST PRACTICES FOR NEWSPAPER ARCHIVES DIGITIZATION PROJECTS

In the context of digital preservation projects, as a management philosophy ‘best practices’ could cover areas such as, successful business strategies adopted, systematic digitization processes followed, appropriate metadata formats used, suitable digital archiving methods adopted, robust ICT tools deployed, user-friendly search and retrieval interfaces developed, etc. In the similar aspects, a number of already published scholarly literature recommend best practices such as, selecting of high quality archival resources for digitization [23]; scanning of newspapers from microfilm collection for ensuring inexpensive scanning and low conservation costs [24]; giving importance to metadata descriptions, thereby enabling useful future access [25]; capacity building of internal manpower for their contribution in digital preservation projects, as part of their daily activities [26]. With regard to actual scanning of newspaper collection, the ‘Australian Newspaper Digitization Program’ followed 13 potential methods for improving OCR accuracy [27] and from these to name a few methods, which TAKC newspaper digitization project followed are-scan at 300 dpi resolution or above; use TIFF files only for OCR; De-skew every page on a vertical and horizontal grid, etc. Furthermore, as Gatos, et al. recommended [28], in TAKC’s newspaper digitization project, the outsourced agency has used, an automated process for image enhancement, segmentation, article tracking, and OCR process and thereby it has ensured saving in cost as well as time. Another area in terms of best practices could be while deploying the best possible digitization tools and technologies, it is always necessary to keep the end-users in mind. To improve the end-user information search and retrieval capabilities TAKC has offered training and demo session in the usage of its digital archives for its internal users, which, over a period of time, has resulted in increased usage of digital archival resources. Moreover, creators of digital libraries and its contents, namely, born digital, turned digital, and gained digital should keep copyright issues in the scheme of digital library creation and management so that copyright issues are addressed within a legal framework and the rights of the creators and owners of the content is protected [29] and in this aspect TAKC has ensured copyright protection compliances of its entire digital archive. Based on the reasonable successes already achieved in its various digital preservation projects, TAKC has followed a number of these best practices and found them very useful and helpful in accomplishing its goals.

CONCLUSION

The preservation and archiving process is made more efficient when attention is paid to issues of consistency, format, standardization and metadata description in the very beginning of the information life cycle [30] and TAKC has already successfully demonstrated these principles in its digital preservation projects. As the way forward, TAKC surely will continue to
engage itself in various digitization projects and thereby it will ensure enhanced permanent access to its rich collection of archival resources. In the process, it is hoped that the TAKC’s role, as India’s largest print media archive, will be further strengthened in efficiently serving the dynamic information needs of the employees of The Times Group as well as the genuine external users from India and abroad.

REFERENCES

Diplomats in the Context of National Archives of Zimbabwe: A New Frontier for Extending Records Management Services

D. Maboreke¹ and L. Muchefa²
¹,²National Archives of Zimbabwe, P. Bag 7729 Causeway, Harare, Zimbabwe
E-mail: ¹muchefa@gmail.com, ²dmaboreke@archives.gov.zw

Abstract—The National Archives of Zimbabwe (NAZ) is the storehouse of the national memory and the only legal repository for all public records in Zimbabwe whether print or electronic. NAZ mission is to acquire, preserve and provide access to these records or archives in whatever format to the public. There is so much trust amongst NAZ clients that its collection is authentic and accurate premised on the notion that the records administered by NAZ are found nowhere else. With diplomatics, records and archives are not supposed to be merely trusted based on legal grounds but real scientific proof that they are authentic like the most influential records (legal or judicial) such as Acts and treaties, their genesis form, contents and what they communicate should be validated. The new frontier for the NAZ is that such records those has been preserved as “Special collections,” meaning that they are kept for posterity and are not easily accessible by the public. The trend towards digital technology, digital preservation and digital archiving challenge this access model. As it stands it is only NAZ and Parliament that houses originals of these Acts. Under such circumstances, how do historians, scholars and other researchers make a critic of these records and not their content as may be given to them when they are treated as “sacred”? There is therefore urgent need to explore the intellectual, scientific and technological intricacies that determine the management of these legal records. The paper will seek to find out the drawbacks of managing colonial treaties in the context of digital diplomatics against a background of mainstream politics and economics. NAZ therefore has a lot to put in place to meet the expectations of the digital age and the demands of the digital natives. The paper will explore the challenges and opportunities of using diplomatics to build virtual archives in Zimbabwe considering that the digital wave is engulfing the globe. It will also seek to find out lessons that Zimbabwe’s National Archives can draw from others in the developed world. The paper will also highlight the benefits that users can get through networking with others. It will highlight the challenges and opportunities of reproduction of legal documents using principals and techniques of diplomatic.

Keywords: Diplomats, Digital Age, Frontier, Access, Preservation, Special Collection, Authenticity
INTRODUCTION

The digital age has brought about complex phenomena in determining authenticity of records and archives. Diplomatics is basically based on the notion that whatever the record claims to be, in regard to its genesis, content and meaning must be verified and approved by both users and custodians of the record. This period of history has been called the Information Age because it makes available instant access to knowledge that would have been difficult or impossible to find previously. The term information age has been a key 21st century feature which best describes the widespread proliferation of emerging information and communication technologies as main drivers of societies and economies. Information age is a component of globalization that has promoted learning culture thereby enhancing mutuality between and amongst different societies. In this information age globalization has seen the rise of virtual memory, digital or electronic records. These have necessitated easy information sharing. Some people would understand globalization in the economic sense in that it is characterized by the transition of companies or firms from being localized to multinational and global. Alberts and Papp [1] explains that technology helps influence changes in society in a way that will help it. Diplomatics in informatics therefore, helps to achieve the objectives of information age that is sharing information which is authentic and trustworthy. When a record is said to be authentic it means it is real and genuine [3].

Katuu [3] traces the history and development of authenticating a record from Middle Ages when a vital record was only considered authentic by the use of seals. Seals were ways of determining that a record was genuine and there was nowhere else such a seal would be obtained as counterfeit. The rise of globalization alongside information technology (IT) has seen the emerging virtual archives and digitization of records which brought about the issues of digital signatures and metadata in determining authenticity of electronic records. Unlike the Middle Ages seals which were physical features, electronic records require more scientific mechanisms in determining originality.

Determining authenticity of a record involves not just the scientific evidence of its originality but continuous preservation of the record by legitimate custodians from the creator to other successors [3]. Duranti [2] underscores the fact that authenticity of records is based on things such as that, the writings of the record testify the practices of the time indicated by the time and place indicated by the record. With such a backdrop NAZ faces the daunting task if the users are to one day try and authenticate its collection. Is there anything in place that will help convince generations to come that what we have in our special collection is genuine record of the colonial period? What shall NAZ put in place to authenticate that the Thomas Baines it has in collection as “invaluable oil paintings” are genuine?

The majority of what NAZ has as its collection is unique in the sense that there are no duplicates to such records. Public records are unique in the sense that it is only at NAZ where access to them is a requirement. Acts and other legal documents may be found elsewhere at places such as Parliament but public access to these is restricted. What therefore follows is the fact that most Zimbabweans are not familiar with such kind of documents despite being familiar with their contents. In light of the aforementioned scenario, it is therefore a
prerequisite for NAZ to establish mechanisms that will start giving the scientific descriptions of the components of these crucial records in relation to the times they have been generated. These will help act as a security measure to NAZ collection in the sense that, selling or auctioneering of some of NAZ holdings such as stamps may require only NAZ staff expertise to justify that it is genuine and not forged.

**INFRASTRUCTURE AND EQUIPMENT**

One of the most fundamental things in establishing mechanisms that ensure accurate metadata in the context of information diplomatics is to put in place the right infrastructure. Infrastructure entails that there has to be the right physical structures and the right equipment in place. If it means there has to be a science laboratory NAZ has to invest into that. Such arrangements will enable NAZ to capture the necessary scientific data that should be captured for all records format ranging from paper records, audio-visuals and electronic records.

Trends in information management and sharing are scaling up daily in maximizing digital concepts and science technology. The challenge remains that being a 3rd world country what then should be done to be abreast of these changing times. Are we then to remain lagging due to economic and financial challenges we face as a nation?

**STAFF TRAINING**

Having the right infrastructure and equipment in place is not enough without qualified personal in the name of archivists, records managers and librarians. Zimbabwe in general is not yet advanced as a nation in utilizing Information Communication Technologies (ICTs). E-governance was introduced in the 2009 and was supposed to be embraced by all government departments; however, implementation is at a slow pace. The crucial reason being neither that it is facing resistance nor being misunderstood, but the fact that training among the civil service was not a priority. Expecting implementation of e-governance by mere placing of computers among other infrastructural equipment without training the worker is putting the cart before the horse. What is essential after putting in place the required resources is to train the staff. NAZ exemplifies how such a program would require the staff to have technical training in the new era of professional duties to be adopted.

It is essential to note that 30% of NAZ staff members have qualifications that had ICT modules or a certificate in ICT. 70% have got degrees in general records management, history and archaeology; this is not to suggest that they are computer illiterate but a reflection of the bias of their strength in executing their duties. Information diplomatics is a complex and technical area of specialty that requires further training and exposure to other institutions that have already embarked on such projects through collaborations.

NAZ offers research services to the public, records management services to government departments and parastatals and creation of records such as oral histories and acquiring manuscripts. These services in the context of information diplomacy should not remain mundane, traditional and static, they should move with times and information practitioners should be well placed in their qualities to execute such mandate. Therefore, training of staff
has to be all-round, from the services that we offer to the clients, disaster plan (preparedness) to the new establishments of the institutions. This is not an easy task for our dear institution which was established 1935 and all these years’ records and archives have been gathering, being preserved, accessed and repaired in their original formats.

Migrating information from one media format to the other through digitization is one task that is mammoth to the institution in view of the resources available, amount of information to be digitized and nature of such records and archives. This ranges from audio-visuals, newspapers, public archives and manuscripts. This however, cannot be said and done without putting the wider context which NAZ falls into; Zimbabwe. Zimbabwe for the last two decades has been facing serious economic challenges and such professions as archiving were not part of the top priority list of government expenditure. It is naturally unfortunate that under such circumstances it will be a miracle to have everything required from the state coffers except by other means as collaboration, co operations, donation of equipment and funds.

**WHOSE HISTORY?**

Inasmuch as the authenticity of a record/archive may be tested and proved to be beyond any controversy, history is dynamic and what that record may have testified may be refuted generations to come. Archives are centred around memory, power and politics, these aspects are not static or constant, the ruler and the ruled change, power shifts, archives controlled by different governments and capturing all records of the society may seem a nightmare but essential to the future generation. Ranger [4] in discussing these dynamics of history understood that patriotic history was much narrowed down version of nationalist history which focused on the three ‘revolutions’ - 1896, the guerrilla war and the ‘third chimurenga’ of land redistribution. This divided the nation into ‘patriots’ and ‘sell-outs’ and the other effect was that there was another version of our past which becomes unacceptable to some circles of the society and to the readership. The issue at stake here is that history; diplomatics and informatics are inseparable in archival business and requires professionals, technocrats and impartial archivists. Archives should be eventually be some form of arbitration for future furore or conflict among the citizens of a nation. Such arbitration and impartiality can only be possible if NAZ considers modern information diplomatics.

**REFERENCES**


Ongoing Development of Digital Preservation and Repository Infrastructure for National Archives of India

Dinesh Katre\textsuperscript{1}, Shashank Puntamkar\textsuperscript{2}, T. Hussain\textsuperscript{3} and Ankit Sharma\textsuperscript{4}

\textsuperscript{1}Associate Director, Human-Centred Design & Computing Group, C-DAC, Pune, India
\textsuperscript{2}Joint Director, Human-Centred Design & Computing Group, C-DAC, Pune, India
\textsuperscript{3}Assistant Director, National Archives of India, New Delhi, India
\textsuperscript{4}Project Engineer, Human-Centred Design & Computing Group, C-DAC, Pune, India

E-mail: \textsuperscript{1}dinesh@cdac.in, \textsuperscript{2}ssp@cdac.in, \textsuperscript{3}thussain2011@gmail.com, \textsuperscript{4}ankits@cdac.in

Abstract—This paper presents an overview of ongoing development of digital repository infrastructure (software, hardware, network) for National Archives of India, as part of the Centre of Excellence for Digital Preservation project entrusted with C-DAC, Pune. The main components of digital repository infrastructure being developed by C-DAC Pune includes software systems like Online Public Records Portal, Abhilekh Digitālaya\textsuperscript{TM}, Datātar\textsuperscript{TM}, Digital Preservation Planner\textsuperscript{TM} and repository hardware setup. The paper provides coverage to changing context during growing volumes of digital records across all ministries, departments and government organizations and new retention policies. It also touches upon the issues involved in transfer of digital records through network to National Archives India.

INTRODUCTION

Centre of Excellence for Digital Preservation at C-DAC Pune

The Centre of Excellence for Digital Preservation project is entrusted with Human-Centred Design & Computing Group of C-DAC Pune. It is a flagship project under the Indian National Digital Preservation Programme initiated by Department of Electronics and Information Technology (DeitY), Government of India. The Centre of Excellence for Digital Preservation has the mandate to develop various digital preservation tools, technologies, standards, best practices and pilot digital repositories in collaboration with different organizations. Therefore, C-DAC has entered into a MOU with the National Archives of India to be developed as a pilot digital repository.

As part of this project, it is proposed to help in evolving the digital preservation policy, repository best practices and training for NAI. The primary goal will be create a proper ecosystem for long term sustenance of digital preservation.
About National Archives of India

The National Archives of India is the nodal agency of the Government of India for the implementation of the Public Records Act, 1993 and the Public Records Rules, 1997. According to definition of public records in Section2(e) of the Act public records include files, documents, manuscripts, microfilms, microfiche or any other materials produced by computer or any other device of any records creating agency. Going by the definition, the digital records are also coming under the purview of the public records. There are approx. 85 Ministries/Departments of Government of India which are the records creating agencies covered under the Public Records Act. Besides, under each Ministry/Department there are large number attached and subordinate offices which are also included as the record creating agencies.

Changed Context and Emerging Challenges

New Retention Schedules for Electronic Records by DARPG

The Department of Administrative Reforms and Public Grievances of the Ministry of the Personnel, Public Grievances and Pensions, have brought out the revised edition of the Records Retention Schedule in Respect of Records Common to All Ministries/Department in 2012 in which the Retention period of electronic records have been defined. Viz. there will only be two categories of e-files as under:

Category I: The e-files which are to be preserved permanently or which are of historical importance. The e-files included under this category will be as follows:

For 10 years, it will be kept in the Departments server and thereafter transferred to the server of the National Archives of India (NAI).

Category II: This category will include e-files of secondary importance and having reference value for a limited period up to 10 years akin to the retention period of physical files/records on the Department server. In exceptional cases, if the record is required to be retained beyond 10 years it will be upgraded to Category I.

Further, as per the 1st edition of the Central Secretariat Manual of e-Office Procedure (CSMOeP) (February, 2012) the e-files/records have been categorized into two categories:

1. Category I (e-Files which qualify permanent preservation for administrative purpose or which are of historical importance).

2. Category II (e-Files of secondary importance and have a reference value for a limited period not exceeding 20 years. In exceptional cases, if the record is required to be retained beyond 20 years it will be upgraded to Category I.)
**e-office Implementation Across Ministries/ Departments of Government of India**

In order to access the status of e-office implementation across Ministries/ Departments the National Archives of India has issued a circular to all Ministries/ Departments to Government of India seeking information about the digital records holdings available in those Ministries/ Departments along with the format and also the medium used for their storage. Another survey form has also been developed by C-DAC to collect information on the digital records of various records creating agencies. The purpose behind these measures are not only to ensure the quantum of digital records that are being generated but also for the preparedness to receive these records in near future. The work in this is under progress.

Therefore, it has inevitable for National Archives of India to develop the technological infrastructure and competencies to be able accept the transfer of electronic records and its long term preservation. Presently NAI has the competencies, procedures and infrastructure to manage the physical records in paper and microfilm formats. Digitization of paper records and microfilms is ongoing. Organization wide networking and computerization of processes is ongoing.

**DIGITAL PRESERVATION AND REPOSITORY INFRASTRUCTURE**

In this project, the Centre of Excellence for Digital Preservation team at C-DAC, Pune has taken up the challenge of developing the digital preservation methods, planning and estimation tools, digital migration tools and digital archival and repository software tools for National Archives of India and the government institutions at large, which are explained as under.

**Online Public Records Portal (OPRP) for Record Officers**

C-DAC has developed an online portal for National Archives of India, wherein the Departmental Record Officers (DROs) can register and access the following forms to initiate the transfer of records. As per Public Records Act, 1993 certain reports are to be generated by the records creating agencies and submitted in specific forms prescribed in the Public Records Rules, 1997 as under:

- **Form 1:** Particular of records of permanent nature due for appraisal.
- **Form 2:** Transfer list.
- **Form 4:** Half yearly statement on periodical review of classified records.
- **Form 5:** Submission of the annual report.
- **Form 6:** Particulars of records destroyed during the year.
- **Form 7:** Half yearly report on recording, indexing, reviewing and weeding of records.
- Online survey of digital records.

An online survey form is also developed to know the present status of digital records in terms of images, audio, video, databases, e-mails, maps, etc. produced or stored with various
ministries, departments and organizations of government of India. The portal provides administrative controls to manage the user accounts, receive, process and respond to the submissions from DROs. NAI is yet to formally announce the availability of OPRP. It is currently in process of testing and further enhancements.

Introduction of OPRP will enable NAI in effective administration, synchronization, management of the record transfer activities (both digital and paper).

**Abhilekh Digitālaya™ an Open Archive Information System (OAIS)**

C-DAC is developing the Abhilekh Digitālaya™, which is a digital archival system specifically designed and developed to meet the requirements of government archives. It is fully compliant with the ISO 14721: 2012. Abhilekh Digitālaya works in the client/ server model over the LAN. It is developed using the latest and open source web technologies by following the user centred design approach. The functional prototypes are constantly kept in use on site and enhanced depending upon the user feedback. Presently this system is designed to ingest digital copies of government records and cartographic maps. Development for ingesting of other data types and formats is ongoing. The system supports different roles with rights and responsibilities such as SIP creator, Digital Archivist, Repository Administrator and Director. The repository administration includes preservation policies, statistics, graphs, integrity assurance, reports, user management, etc. The system allows Director to digitally sign the Archival Information Package (AIP) for authentication.

Abhilekh Digitālaya™ also provides a digital repository portal for giving public access to digital records with customization of Delivery Information Package (DIP). It allows search and retrieval and controlled access to users. The development of Abhilekh Digitālaya™ is presently ongoing.

Fig. 1: Digital Migration using Datantar™ from Microfilm Scanning to Abhilekh Digitālaya™
Datântar™ for Digital Migration

C-DAC has designed and developed Datântar™ for digital migration which facilitates the preparation of valid Submission Information Packages (SIPs) for reformatted digital images. Datântar™ is a desktop software which has to be installed on multiple computers which uses the centralized database through LAN. It provides features like defining and assign unique record identifiers, URI register with fixity for authentication, digital watermarking, automatic and manual OCR, tagging, access image conversion, open packaging, and user management etc.

Digital Preservation Planner

C-DAC has designed and developed the digital preservation planner software which can be accessed online. It provides a searchable and sortable database of over 1000 record types and corresponding retention rules and policies of different government organizations such as:

- Record retention rules for financial records (GFR).
- Record retention rules for banking records.
- Record retention rules by Central Vigilance Commission.
- Record retention schedule for all Ministries and Departments.
- Record retention rules for substantive functions/ employment.
- Record retention rules of UPSC institutions.
- DARPG e-office procedure.
- e-record retention rules by DARPG.

This software allows the record officers to select the record types based on the retention requirements and assign the digitization specifications (file formats) to estimate the digitization cost, storage space, storage media, etc. Currently the file formats and digitization specifications for audio and video are being integrated. The software also allows the users to define their own record retention policies.

Hardware Infrastructure for Testing Purpose

C-DAC has presently procured and established the hardware infrastructure in terms of Intel based Server and Network Attached Storage (NAS) at National Archives of India for testing of the various software systems as discussed earlier. NAI has established a dedicated server room with hosting infrastructure along with a lab for the SIP creators and consultants for using Abhilekh Digitâlaya™ and Datântar™. C-DAC Pune has also setup the Storage Area Network (SAN) to facilitate remote backup through NKN (National Knowledge Network) connectivity which is yet to be commissioned. The full-fledged hardware infrastructure will be established only after all systems and processes are stabilized and internalized by NAI.
TRANSFER OF DIGITAL RECORDS

The underlying network infrastructure has a major role in file transfer operation. As
government organizations are already equipped with internet connection, it is appropriate to
transfer data using network file transfer method to NAI. However the possibility of transferring
data and files using offline medium cannot be ruled out. In both cases, the transfer operation
will be governed by record retention and transfer policy set for the organization. Online and
offline mechanisms of data transfer must cater to the aspects such as secure transfer, fixity,
transfer metadata, control information, validation, etc. The file transfer will be initiated by
DROs after processing the request through Online Public Records Portal. The details of
transfer request will comprise of the metadata of file contents as well as file transfer
environment. The transfer process will require to ensure the validity, authenticity and
provenance of source in terms of the human user, organization, network and hardware.

CONCLUSION

Digital repository infrastructure development is a wide spread activity. It is necessary to get
active participation from the stakeholders for internalization of digital preservation processes.
The digital repository development and its sustenance requires long term support. The need
for defining a set of standards and guidelines for transfer of digital records is also evident.

REFERENCES

[2] Electronic Records (Section J), Record Retention Schedule in Respect of Records Common to All
Ministries/Departments, Published by Department of Administrative Reforms and Public Grievances
(DARPG), 2012.
Lifecycle Model, In Proceedings of Indo-US Workshop on International Trends in Digital Preservation,
(March 2009), Pune, India, pp. 137-141.
Digital Preservation of Information Resources in Academic Libraries in Nigeria

Attahiru Saminu
Department of Library and Information Science, College of Science and Technology, Hassan Usman Katsina Polytechnic P.M.B. 2052, Katsina State, Nigeria
E-mail: attahirusaminu@yahoo.com

Abstract—The world is witnessing a rapid and progressive transformation as a result of IT revolution, which has predominated all aspect of societal development. One of such transformation has manifested in the form of digital libraries. The evolution of digital libraries has ensured the emergence of global networked environment that has dramatically changed the face of libraries, their functions, services as well as their storage and delivery system. One of such services is digital preservation, which is the work needed to ensure that digital content is maintained and accessible into the future use. This paper intends to examine the status of digital preservation of information resources in Academic libraries, its benefit, challenges and prospects in Nigerian perspective.

Keywords: Digitization, Digital Preservation, Information Resources, Challenges, Academic Libraries

INTRODUCTION

Digital preservation is a major concern for digital libraries. Libraries all over world have always placed emphasis on the preservation of information resources and this has posed serious challenges to librarian and other information managers. Preservation no doubt has in increasingly continued to take centre stage information management and handling particularly with proliferation of publications and other information resources in different formats. The present global information environment propelled by information and communication technologies has made preservation of information resources, particularly in academic libraries more critical to information professionals than ever before. Research out puts of any academic institution are considered to play critical roles the assessment of intellectual and cultural growth of any society particularly in developing countries where knowledge production and dissemination have been considered very low. Aajibili [1] urges that, the major goal of any preservation technique is to hand over the society’s civilization to another generation and equally make them available and accessible to the international scholarly community. Therefore transmission of Institutional information recourses which are part of the cultural heritage of a people is far to library and information professionals.
Muhammed [9] maintains that the information resources contain active and inactive documents which accumulate overtime in the life of the Institution. Usually information resources are mainly a collection of grey literature and other publication from scholars.

The rationale for a digital preservation of information resources in academic environment has been underscored by Ahrams [2] when he explained that information resources creates enabling environment for scholarly publishing and make research productivity of a particular institution more visible globally. It therefore adds value to credibility of a university in terms of its intellectual. The objective of this paper is to examining the status of Digital preservation of Information Resources in Academic Libraries in Nigeria. To find out the benefits, challenges and way forward as Nigerian academic libraries adopt the need of globalization of information for their educational development and the fact that Nigeria cannot afford to stand a loaf.

THE CONCEPTS OF DIGITIZATION

The term digitization’ has been described by scholars and writers alike in different ways. In other words the term means many things to many experts. However, Fabunmi [4] has observed that digitization is the art of converting the contents of a document from hard copy into machine readable formats. Mutula [10] has noted that digitization makes information resources available electronically. He also noted that, in the African context, digitization refers to ‘conversion of non-digital materials to digital form’. Jones [6] asserted that digitization convert, material from formats that can be read by people (analog) to a format that can be read only by machines (digital) such as read–only scanner, digital cameras, planetary cameras and a number of other devices which can be used to digitize information materials. Lesk [8] observed that the primary and usually the most obvious, advantage of digitization is that it enables greater access to collections of all types. All manner of materials can be digitize and delivered in electronic formats and the focus of the contents that are selected for digitization varies across institutions.

In Nigeria, the state of a library and information environment is not very encouraging. This is because it is limited to the existence of wide spread poverty, high illiteracy rate, poor funding of educational libraries, lack of infrastructures, as well as large scale corruption that has eating deep into the fabrias of the Nigerian society. These and many more factors tend to create very hostile and unpleasant academic environment. The top priority of most Nigerians is on survival.

The above situation aptly presents a clear picture of the contemporary Nigerian society which is peculiar in nature, and within which libraries including digital libraries are expected to function. However, libraries in Nigeria are provide and organized access to information and knowledge and more importantly in form of a digital library due to the shift in paradigm for obvious reasons.

Digital libraries development in Nigeria is a quite recent phenomenon. Ogunsola [11] reported that the federal Ministry of education had embarked on the establishment of National Virtual (Digital) library project with the aim of sharing locally available resources
with libraries all over the world using digital technology. Gbaje [5] observed that the National Universities Commission virtual library project initiated in early 2002, the National Board for Colleges of Education virtual library, the open university library project and the UNESCO virtual library pilot project initiated in 2003 were some of various initiatives by the Nigerian Ministry of Education regarding Virtual (Digital) libraries for higher institutions in Nigeria; but yet none of the efforts had yielded any functional (visual) library at that time.

Besides, the rapid growth of the telecommunication industry in Nigeria, another recent success story in digital library development in the country was reported by Rosenbergn [12] which stated that the Mobile Telephone Networks (MTN) Nigeria had successfully deployed and commissioned digital libraries in three (3) universities in Nigeria. The three Universities are the Ahmadu Bello University Zaria, the University of Lagos and the University of Nigeria, Nsuka. The fourth one which has been billed for commissioning early 2010 established at university of Benin.

**DIGITAL PRESERVATION IN ACADEMIC LIBRARIES**

Digital Preservation in academic libraries raise challenges of a fundamentally different nature which are added to the problems of preserving of traditional format materials. The term ‘digital preservation’ means the planning, resources allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable.

As more information resources are digitized or born digital, the question of how to keep digital objects accessible for future generation becomes increasingly pressing. Digital preservation in academic libraries presents digital libraries with both technical and service challenges. The method required to preserve digital objects in a readable format test the technological capabilities of digital libraries in additions to presenting complex service challenges. Institutions must make decisions about which documents to preserve and if, or how to preserve the context of the document. The decisions made today will directly impact upon digital libraries ability to meet future user needs.

The major issues the digital library services that will satisfy user expectations and resolve their information needs for generations to come include how to determine the stake holders in, as well as the legal issues affecting digital preservation initiatives. As part of the effort to preserve institutional information resources and ensure global accessibility of these information resources, the University of Nigeria, Nsuka started digitization project in 2008. The ICT centre which was completed recently with modern information and communication facilities is used for digital preservation initiative. The main objective of the project is to digitize local information resources such as undergraduate research project, postgraduate theses and dissertations generated over the years by the university and publications authored by staff of the university.

The Nmandi Azikwe library is taking a centre stage in the digital preservation initiatives with most of librarians participating in the project. In addition to the librarians, technical staffs were recruited to handle such jobs like scanning of documents, conversion and migration.
Librarians are involved in supervision of the work to ensure that the work is properly done. Eakin [3] maintained that in Nigeria digital preservation in academic libraries in Nigeria depends upon: Copying, Transferring and other strategies.

- Copying also referred to as refreshing or ‘migration’ which is complex than simply transferring a stream of bits from old to new media or from generation of systems on the next.
- Transfer digital information from less stable magnetic and optical media by printing page images on paper or micro filming.
- Another strategy for digital preservation is to preserve digital information in the simplest possible digital format in order to minimize the requirements for sophisticated retrieval software.

Digital preservation is an essential element of digital library management due to the increasingly important role digital information resources play in our academic environment. The work of libraries and information workers will be central to the success of digital preservation activities in academic libraries in Nigeria. Therefore, librarians and information workers must develop skills to take the many challenges raised by digital Meredith ship.

**BENEFITS OF DIGITAL PRESERVATION**

Due to digitization all knowledge has become very accessible due to the free open access to information resources made available online. The benefits are:

1. No physical limit for storing.
2. Can be access via the internet.
4. Create saving space.
5. Preservation of old texts/ manuscripts.
6. Easy retrieval of information using key words.
7. It is cheaper to maintain digital library than book library.
8. Any number of times digital files can be duplicated with exactness.
9. Many can access a digital file at the same time.
10. Integrated online resources sharing.
11. Linking and networking possibilities.
CHALLENGES OF DIGITAL PRESERVATION IN ACADEMIC LIBRARIES

Digital preservation has so many advantages but it is worthy to note that there are some factors that hinder it. Kanyengo [7] identified some specificity of preservation issue in academic libraries of Nigeria:

1. Information Policies.
2. Infrastructures.
4. Technical knowledge.
5. Legal Barriers.

THE WAY FORWARD

In the light of the foregoing, the following are here by recommended as the way forward:

- Government at both the federal and state level in Nigeria must develop a more proactive and progressive attitude to implementation of the national policy for information infrastructure and facilities.
- The problem of epileptic power supply and poor telecommunications infrastructure should be doggedly addressed by the Nigerian government by injecting the necessary funds and technical expertise.
- Academic libraries should explore more alternative sources of funding as over reliance on the government on monies that are not fourth coming may not provide the desired solutions.
- Practicing librarian must be involved in training and retraining in the knowledge of ICT competencies required for them to effectively manage the resources in the Academic libraries towards digitization of library services.
- Finally, the need for the committees of University librarians and their counterparts in Polytechnics and colleges to sensitize their respective Intuition’s administrators on the central role of academic libraries in teaching, learning and research activities in academic communities.

CONCLUSION

Digital revolution has dramatically changed the face of libraries in the 21st century. This posed a challenge to academic libraries to digitize their services and resources through appropriate ICT application in order to remain relevant. However, academic libraries in Nigeria are faced with enormous problems notably that of lack of proper funding. Despite that, they must against all odds try to measure up with their contemporaries elsewhere. Their ability to overcome some of the major challenges undermining their effort in providing the desired services to their communities in this information age should therefore serve or one of the coping strategy methods for the 21st century academic library service in Nigeria.
REFERENCES


Legal Aspects of Digital Preservation
INTRODUCTION

Copyright is about protecting original expression. Copyright protects “original works of authorship” that are fixed in any tangible medium of expression from which they can be perceived, reproduced, or otherwise communicated either directly or with the aid of a machine or device. Copyright arises as soon as a ‘work’ is created (or fixed). It does not extend to any idea, procedure, process, system, method of operation, concept, principle or discovery, unless fixed in a tangible form.

Copyrightable works include the following categories: literary works; musical works, including any accompanying words; dramatic works, including any accompanying music; pantomimes and choreographic works; pictorial, graphic, and sculptural works; motion pictures and other audiovisual works; sound recordings, architectural works and computer programs. In the digital medium, every web page accessible or published in the World Wide Web is to be taken as a literary ‘copyrightable’ work. It protects all written text materials, graphic images/ designs, drawings, any linked sound, video files or films, whether part of a web page or a website. Copyright also protects the “look and feel” of a website.

COPYRIGHT IN THE DIGITAL MEDIUM

Technology is a double-edged sword. On one hand, it is creating new means to fix the original expression in a tangible form and on the other it is being exploited in infringing the copyrights with impunity. Even some of the Internet activities, like caching, browsing, mirroring, scanning, downloading, uploading, or file swapping are an anathema to a purist.

- **Caching**: It is used to improve response time for end users. It means copying of a web page/ site and storing that copy for the purpose of speeding up subsequent accesses.
- **Browsing**: A software driven process for searching the world wide web using a browser.
- **Mirroring**: It improves service for the users by replicating a web site across various servers all over the world and make available the critical information to all the users at all times.
• **Downloading**: It means to receive information, typically a file, from another computer via modem.

• **Uploading**: It means to send information, typically a file, from another computer via modem.

• **File-swapping**: A “peer-to-peer” transmission of digital files from one computer to another via the Internet.

The question is whether such activities really infringe the five exclusive statutory rights of a copyright owner, such as:

a. To fix (store) the information in a tangible form.

b. To reproduce the copyrighted work.

c. To sell, rent, lease, or otherwise distribute copies of the copyright work to the public;

d. To perform and display publicly the copyright work; and

e. To prepare derivative works based on the copyright work.

The answer is yes. Internet activities like caching, browsing, mirroring, scanning, uploading, downloading or file swapping may result in:

a. Transmission of information from one computer system or network to another, involving temporary storage (RAM) of that information.

b. An unauthorized storage of such information a violation of the copyright owner’s exclusive right to make copies, i.e. to reproduce the copyrighted work.

c. A violation of the copyright owner’s exclusive distribution right.

d. An appearance of a copyright image in a web browser infringing the copyright owner’s public display right; and

e. An infringement of the copyright owner’s exclusive right to prepare derivative works.

Prima facie, the nature and characteristic of Internet activities is such that there will certainly be infringements to the exclusive statutory rights of a copyright owner.

Lawmakers have refused to accept that a ‘paradigm shift’ has taken place and the law related to copyright can no longer revolve around these five statutory principles as enumerated above. Here, we are looking for a leap of creative imagination from the lawmaker, which is somehow missing to address the issue of technological advancements.

Advent of printing press and gramophone led to the expansion of copyright laws encompassing new technological developments. It went into further expansion mode in terms of “neighbouring rights” with the development of broadcasting and related technologies. In other words, each and every technological advancement led to an addition of ‘copyright layer’ in the existing legislation. Every ‘copyright layer’ further extended ‘fair use’ category.
Internet technologies did lead to creation of an additional copyright layer, but somehow the said layer failed to do justice as it ignored the basic philosophy of Internet, i.e., Open source. It is important to understand that ‘fair use’ cannot be equated with ‘open source’, as ‘fair use’ is still part and parcel of copyright legislation. Protecting copyright in a digital medium goes against the “open source” nature of the Internet.

**CODE IS LAW**

Code [software] is embedded in hardware and it is the code which would regulate the human behavior.

Digital content is the third layer; it rests on other two layers, namely hardware and software. Interestingly, this third layer is dependent upon both the layers in order to manifest itself. In this context, it is significant to note that the ‘hardware layer’ is protected under the patent law, ‘software layer’ protected under both patent & copyright laws, whereas the ‘content layer’ is protected under the copyright law only.

Hardware layer and software layer are the basic templates of Internet. For an open source to manifest its completely vis-à-vis Internet technologies, it is imperative both hardware and software layers must also come under open source. Although, software licenses are increasingly falling under open source, but the same cannot be said about hardware.

Hence one of the critical elements of the Code is hardware and unless the ‘hardware layer’ is not on open source, the content layer cannot be free from copyright regulations.

**CODE IS CODE**

It would be fallacious to assume that no copyright exists in the digital medium. In fact, there exists a strong copyright regime in the digital medium as it gets support from the legal regime that protects the hardware and software. Code protects as well regulates the additional copyright layer. In a way, Code is status quoist, it extends the existing copyright regime to cover newer technologies. Code does not support open source.
Digital Preservation of Court’s Disposed Case Records—A Case Study from Indian Judicial System’s Perspective

Payal Abichandani¹ and Rishi Prakash²
¹,²CDAC, C-56/1, Sec-62, Noida (UP), India
E-mail: ¹payalabichandani@cdac.in, ²prish@cdac.in

Abstract—In India there are about 24 High Courts with large numbers of district and subordinate courts. Every year thousands of cases get disposed and corresponding file and evidences dispatched to record rooms for archiving. Managing and accessing large amount of documents in the form of physical records is itself a challenge for judicial system. Therefore, proper infrastructure and space is required for storing and accessing large amount of disposed case records. Financial and physical resources, such as, manpower is required for handling enormous amount of records. To solve this issue Indian courts are moving towards the concept of digitization. On the other hand, only conventional and unstructured digitization does not solve the problem completely. The challenge is to preserve the disposed case records for long term in technologically independent environment and keep information and knowledge available for the judicial fraternity. The reference model ‘Open Archival Information System (OAIS ISO 14721)’ can be adapted to achieve technologically independent long term preservation. In order to preserve these records we need to extract the metadata of respective disposed case records in accordance with the OAIS standard. Each and every court in India is maintaining and storing disposed case records separately in their respective manner. There is a need for metadata standardization process for the ease of preservation and archiving of records. This paper is based on the initial attempt (Proof of Concept) for metadata standardization of judiciary records.

Keywords: Metadata Encoding and Transmission Standard, Metadata standardization of Judicial records, Representation Information, Digital Obsolescence, Submission Information Package and Archival Information Package.

INTRODUCTION

Legal system is one of the prominent pillar of any country as it nature society, politics and economics. It acts as a best mediator between people and institutions. In India every court is maintaining their records independently irrespective of any standardization and global access. Today, we are able to store the data, but, are faced with serious challenges when it comes to preserving the information that the data represents. Long Term Digital Preservation attempts to address these challenges. Long term preservation, basically, preserves the digitized information
along with its metadata, so that the valuable information can be used seamlessly and when required in the future in spite of obsolescence of hardware, software, processes, format, people, etc. To preserve valuable judiciary records for future access, there is a need to standardize metadata among all courts in India.

Metadata: is data about data which is required to understand and interpret the actual data that need to be archived. There is a conceptual framework developed by NASA’s Consultative Committee for Space Data Systems (CCSDS) called Open Archival Information System (OAIS- ISO 14721 standard). This paper will explain about the implementation of OAIS framework for long term technology independent preservation of Judiciary records. In India each court and its sub courts are maintaining their data independently irrespective of any standard policy. Before preservation of disposed case records, there is a need to systematize the records among all district courts in a standard manner as per OAIS standard.

CDAC: Noida developed a tool called Disposed Case portfolio Manager. This tool will collect and extract the metadata from all Delhi district courts. After, extraction this tool will standardize and encode the metadata of respective disposed case records. There is also an option to manually enter the metadata in case of missing metadata information. NIC is already doing the process of capturing the case records domain level information. To use the data generated by NIC, database administrator had written some procedures functions, views which need to run inside the oracle database, to extract and map the disposed case records data to the long term preservation application database.

Following are the concerns which can be solved by adopting this solution:

- Preservation of case records along with their metadata.
- Defined strategy is required for maintaining disposed case records.
- Centralized digital library for all decided cases at High Court is required for future reference of records.
- Standardized migration policy is required to prevent from the loss of data.
- Because of scattered information it is very difficult to search old case records.
- Online connectivity among District courts and High court for automatic case record transfer is required.

IMPLEMENTATION STRATEGY

CDAC NOIDA developed a Metadata Standards for Long Term Digital Preservation of Disposed Case Records. This document encloses the details of metadata required to preserve the electronic records of disposed cases in the judicial scenario. This metadata, acts as an input for the Open Archival Information System (OAIS- ISO standard 14721) for the digital preservation of Delhi District Court’s disposed cases records. After extraction of records from different sources, normalization process will start as per the metadata standard. There will be provision of entering missing domain level metadata.
Metadata Standardization Procedure

Metadata will be divided at following levels:

1. **Domain level metadata**: Comprise all the metadata collected by the domain experts. This metadata will cover descriptive, structural and administrative metadata. For judiciary domain, metadata is standardized into following section:
   - Case Information.
     - General case log: will contain all the information related to the disposed case record, such as, Unique Case Id, Case Type, language of the case record etc.
     - Petitioner/ complainant information: It will contain the information related to the petitions of the disposed case, such as, petitioner name, appearance date of petitioner etc.
     - District court information: This section will contain the information related district court in which the respective case was disposed.
     - Filing data: Date on which the case id filed in district Court.

2. **Technical level metadata**: contains information that is required to render the data. Representation Information -The information that helps to understand and interpret the data (e-record) is called Representation Information (RI). There are two form of RI i.e. Structural and Semantic information. Structural information interprets the digital object and semantic information adds additional meaning to i.e. Document either a .pdf or .doc file can be understood with the help of technical specification. Without technical metadata it is impossible to understand and interpret data, like in the above case you will not be able to judge whether it’s a .doc file or .pdf file.

   Following are the benefits of Metadata Categorization:
   - Metadata helps to preserve specifications about the digital object.
   - Metadata needed during migration from old technology to new technology.
   - Supporting and managing access, privacy and rights.
   - Support Authenticity and Understandability.

Metadata Validation and Verification

Without metadata every data will be in raw form and it will require some technology which can interpret the data in readable form. To get the accurate information about the data it is very essential that metadata should be as accurate as possible. After the metadata extraction and generation procedure, one cycle of review is required to check the correctness of the metadata generated. In case of some error there should be provision to correct the metadata. Metadata for disposed case records in Delhi District Courts will be collected from various sources. There are three scenarios for collection of Case metadata & e-Files from the existing system.
1. Scanned files from Delhi District Courts.
2. Case e-Files already exist in NIC System.
3. Case e-Files with Video recordings files from Karkardooma e-Court.

![Fig. 1: e-Court Recording](image)

After, the collection process metadata needs to be standardized based on preservation standard.

**Packet Generation**

Submission Information Package- is a package which will contain the data that need to be preserved and related information which will be required to preserve the data for long term period. This packet will be generated as per the OAIS framework.

- Procedure for creating Submission Information Package After the collection, extraction, standardization and verification of data and its metadata among all district courts, a package will be created by the validator known as Submission Information Package.

![Fig. 2](image)

- Packet Components
  - Data Section is the actual content (Disposed Case Records) that need to be preserved for long term.
  - Metadata Section will contain all the essential information required to interpret and understand the data, which will be beneficial during archival.
- Preservation Description Information file will contain all the essential information required to preserve the packet for long term.

- Manifest file will contain the transmission information about the SIP package.

- As per OAIS standard, SIP package will be created at Producer level and then transmitted to OAIS framework for long term storage.

- The solution developed by the CDAC Noida will be deployed in all Delhi district courts and data along with its metadata will be collected, generated and standardized in every district court. Simultaneously, SIP packet will be generated in every district court and transmitted to High Court of Delhi for object based storage.

- Disposed Case Portfolio Manager Application developed by CDAC- NOIDA assures this need in following manner:

  This is a role based web application which is strictly based on the access rights applicable to respective roles. There will operator- who will do the metadata entry for respective case records and will have the facility to upload the records. There is another user called validator- who will validate the disposed case records. In case of major error he/she will revert the record to the respective operator to do the correction suggested by the validator. Validator will also have the right to do the edition in case of minor changes and can also view the records uploaded by the operator. If everything is ok, then validator will accept the record and start the activity for packet generation. The Output of a Disposed Case Portfolio Manager is a SIP packet generated as per OAIS Standard.

Following architecture adopted in Disposed Case Portfolio Manager:

![Diagram of Disposed Case Portfolio Manager](image-url)

Fig. 3: Architecture Adopted in Disposed Case Portfolio Manager

There is another application named as e-Goshwara which is under development in CDAC Noida. This application will be deployed in High Court of Delhi which will receive the SIP

---

1Goshwara-This word is derived from the Urdu language, that means record and it is commonly used in northern parts of Indian courts as a nomenclature for record rooms.
packet from various Delhi district courts. As per OAIS standard, this application will verify the quality and authenticities of the SIP packet received from various sources and then reframe it as an Archival Information Package which will be send to the Object based storage system.

- Quality and Verification Procedure.

After, the SIP is transferred through transmission session between district court and high court of Delhi.

Following sequence of test will be applied to the packet:

- Antivirus Check.
- Hash Value Check.
- Format Check.
- Path Check.
- Number of files in SIP.

In case of any error, packet will be rejected and proper acknowledgment will be send to the sender. After taking necessary actions, sender will retransmit the SIP packet. If all parameters are ok then respective acknowledgement will be send to the sender and e-Goshwara application will start the process of SIP to AIP conversion.

Storage device will store the AIP packet in the form of object with unique identifier. Object based Storage will store the packets in the form of object which can be accessed using unique identifier called OID. Object metadata will contain the information about the physical location of the object. For accessing the object, as per the user request open source protocols can be used. Usage of the following protocols will add the advantage of interoperability. HTTPS, FTP, RESTFUL are some of the protocols which can help to read and write the packet from the storage device.

Packet Rendering

CDAC NOIDA is also developing Access portal which will acts as an interface between user and e-Goshwara application. User can simply search for the particular case records based on their need. There are some technical background activities which will run based on the user input to get AIP object. Based on the requested parameter access portal will hit the centralized database and fetch the AIP packet id. Based on the unique id of AIP packet, one of the process of access portal will fetch the object of respective AIP. As per OAIS standard AIP packet will be reframed as Dissemination Information Package (DIP). With the help of preserved metadata and technology individual content in the DIP packet will be displayed to the end user. With the help of object based storage we can store any type of data. In case of judiciary records, data can be a simple case file or there can be any type of data as an evidence e.g. audio, e-mails etc. which need to store in their respective manner.
RECOMMENDATIONS

- e-Record Retention, Disposition & Access Policy for digital or electronic records produced by e-courts needs to be clearly defined.
- The domain specific metadata for judicial records requires to be standardized.
- The requirements of standards with special significance for e-courts in terms of preservation of digital evidence need to be investigated and defined by involving the domain experts.
- The e-court record creation processes need to be standardized by incorporating the e-record management practices as integral part of CIS system.
- Record validation processes needs to be drafted & documented for archival purposes.
- Audit procedures are required to be framed for process validation.
- Reliable Broadband connectivity in district courts and in High Courts is required to be made available; currently NIC-net connectivity is not sufficient for the project. Availability of National Knowledge Network would be the ideal solution for courts.
- Manpower/ staff cadre to manage the digital preservation is required to be defined. Currently the available manpower and their skillset are not at all sufficient to manage digital preservation activities.
- Means of economic & financial sustainability of Trusted Digital Repository are also required to be visualized & channelized so that it could itself run for longer period of time.
- National Judicial Referral System must be built on established Trusted Digital Repository using BIG DATA, semantic web concepts and by developing domain ontologies.

CONCLUSION

The solution developed by CDAC Noida will help the Indian Judiciary System to preserve their records for longer period of time. These records are assets of the Indian Judicial fertility. Current and future Law students, lawyers and judges can refer these old records for various reasons. Even after, decades these records can be rendered because this solution is storing the records in the object form along with their metadata. This is a solution to the threat of technology obsolescence.

Besides long term storage, this solution will also gather and standardize the scattered records and its associated metadata among all district court in centralized location. The records in the form of packets can be rendered for various statistical analysis also. Standardization of metadata among all Delhi district courts will solve the following concerns:

- Maintenance of case records will become easy.
- Fast and efficient searching.
• Reduction in redundancy.
• No scattered information.

REFERENCES
Digital Preservation for Legal Compliances: Where does Swaziland Stand?

Thiyam Satyabati Devi¹ and Ntombikayise Nomsha Mathabela²
¹,²University of Swaziland, Kwaluseni Campus, Swaziland, Southern Africa

Abstract—Legal information is at the core of democracy. Legal information includes primary records and information created by the government as well as secondary writings and compilations that organize, explain and evaluate those primary records. Today, legal resources come in a variety of formats – from leather bound books to microfilm to computer systems. No matter how the information is recorded, all are vital resources for our democratic system and need to be preserved in order to ensure ongoing and sustainable access. These articles will describe how the legal materials are preserved in Swaziland in the digital age. The study is based on the survey of a structured questionnaire related to the current digital environment. The organization/institute included in the survey is the University of Swaziland Library, The SwaziLII (Swazi legal information institute) and the National library of Swaziland.

Keywords: Legal materials, SwaziLII, National library, UNISWA

INTRODUCTION

Swaziland is a landlocked country that shares borders with Mozambique and South Africa. It was founded by Bantu peoples from Mozambique in the 18th century and became a British protectorate when colonial rule was established in 1903. Swaziland was led to independence by King Sobhuza II in 1968, and is now a dual monarchy with a King (currently King Mswati III since April 25, 1986) and the Queen Mother.

Legal material is often complex in organization and presentation. It has a broad audience and is vital to users and to a democratic society. The potential loss of legal information, particularly primary legal materials and their supporting documents, threaten the very foundations of our democracy. So it is necessary to mandate the provision of online services to their citizens. The free flow of the information via the internet facilitates the transparency and accountability and it also increase the accessibility of government at all levels and encourages citizen’s participations in our democracy. Laws can always be amended. Case decisions can be overturned or sent back for further work or reversed. So you can't ever say that a legal document is final because perhaps it has been updated or supplemented or changed in some way. So preservation of legal information in all formats is all imperative.
University of Swaziland Library

The University of Swaziland library since its inception has a law collection section in the library which caters the needs of the law students as well as the law staff of the university and also legal professionals who join the library through institutional membership at a fee. With the coming of the digital age the library is trying its best to provide e-resources for students and the faculty by availing some legal databases through subscriptions on its website. Currently the library is subscribing to LexisNexis, JSTOR, Sabinet E-Journals, and linked to other free access online law databases. The law collection is not a separate entity from the main library it is at the main campus, at Kwaluseni. It is arranged in such a way that all legal materials comprising of law textbook, print law journals, law reports and legal dictionaries and encyclopedias are housed in a separate wing in the library. The library has started an Institutional Repository for digital preservation but still is in infant stage. We started with the thesis, dissertations and the reports that were submitted for the degree. It is high time to prepare for the born digital materials to preserve for future generations.

The Swaziland National Library

The Swaziland National Library Service (SNLS) like most National library services in the global south serves a dual purpose, mainly the National Depository library and Public library systems. SNLS operates two main public libraries in the major towns of the country (Mbabane and Manzini) and thirteen branch libraries located in the smaller towns. The Mbabane Public Library and the National Library are housed in the Swaziland National Library Service headquarters in Mbabane which is the country’s capital city. It is the nation’s leading documentary resource institution under the Ministry of Information and communication and technology started in 1970 in Manzini and the service began in 1971 under the Ministry of Home Affairs then known as the Ministry of Local Administration. The British Council, the United States Embassy and the Ranfurly Library Service were instrumental in its construction and in getting the SNLS functional information needs of Swaziland (Kingsley, 1991: 33). The SNLS was initially established for recreational purposes rather than to perform the functions of a National Library. Hence it was more of a Public Library than a National Library (Mkhwanazi, 1996: 1-2). The vision of the SNLS is “to develop the Swaziland National Library Service as the nation’s leading documentary resource institution” (Swaziland Government, 2008a: 1). In 1983 the Swaziland National Library was opened in Mbabane which is now the present headquarter. The library is run by 160 staff of which 27 are professional staff. As part of its mandate, the National Library collects Indigenous Knowledge from resourceful Swazi citizens for posterity. The Swaziland National Library Service has subscribed to the South African Bibliographic Network (SABINET) to increase user’s access to a variety of electronic databases. SNLS is a member of the Swaziland Library and Information Consortium (SWALICO) with the key partners in information industry. The National Library receives legal deposit on the National imprint in all the varying forms. This responsibility is shared with the University of Swaziland and the National Archives. The National Library deposits all materials on Swaziland with the Centre for International Collections, at Ohio University Libraries and the Library of Congress.
Information landscape is changing rapidly so Swaziland National Library Services has to form a strategy to analyze and evaluate its collection development. The National Library of Swaziland recognizes the intrinsic cultural value of born digital materials, and the importance of preserving this material for current and future generations.

SwaziLII (Swazi Legal Information Institute)

SwaziLII is an institute that is committed in conveying free access to legal information to individuals, the legislature, the executive, the judico-legal world, corporations and its international stakeholders to achieve universal, popular and well researched information. It was started as a project of the Swazi Judiciary in the year 2010 to create a sustainable organization providing free and open access to all Swazi law primarily via the internet. The operation is overseen by the SwaziLII Board whose members are drawn from the Judiciary, Attorney General's office, the Law Society, the University of Swaziland and CMAC. They make court judgments, court notices, court rolls etc. from the all superior courts in Swaziland available i.e. (High court, Supreme Court, Industrial court, Industrial court of appeal and CMAC awards. Users can access judgments by case name, citation, judgment date or by a specific area of law. Through their counterparts AfricaLII they received a donation from Freedom House in the form of a freedom toaster. This kiosk is for those who do not have internet access, the kiosk links with the website and anyone can access it since it is housed at the High court of Swaziland. A user can either bring a flash drive or a compact disk to copy judgments. SwaziLII is funded by the Open Society Foundations, African Legal Information Institute (AfricanLII), Swaziland Royal Insurance Corporation and the Judiciary of Swaziland. Since they are a donor funded project they are still looking for more sponsors to keep this project a going concern. You can access SwaziLII wherever you may be in the world by simply typing www.swazilii.org.

Objectives

The main objectives of the study are:

- To know when they started to collect the legal materials digitally.
- To know how the legal materials are preserved for future.
- To find out the usefulness of legal material to be available online.
- To know the source of funding for the digital collection initiatives.
- To know whether they adopt digitization for the legal materials.
- To know how the digital collections are created.
- To know whether they acquire born digital resources.
- To find out whether they outsource the materials.
- To find out how they help the public to use the information.
RESEARCH METHOD

To meet the main aim and the specific objectives of the study a quantitative research methodology study along with a comprehensive literature review were employed. A structured questionnaire was prepared for collecting the data. The questionnaires were collecting from the institute by email and the findings were disclosed. Data were analyzed using descriptive statistics.

FINDINGS & DISCUSSION

Law librarians have been concerned with preserving the material in their collection since their inception in the library. The major reasons are:

- Law materials represent a significant investment for every library.
- The unaddressed problems of the past are now coming to bear.
- The earliest legal materials are still used today since they may contain "good law."

The analysis is based on the questionnaire collected from the institute/organization that has legal collection of the country. The data collected from the institute/organization are analyze and tabulated as follows.

Table 1: What Year did your Institution Begin Collecting Digital Resources and do you have an IT Services?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>Year</th>
<th>IT Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>2012</td>
<td>Yes</td>
</tr>
<tr>
<td>SwaziLII</td>
<td>2010</td>
<td>Yes</td>
</tr>
<tr>
<td>National Library</td>
<td>2011</td>
<td>Yes</td>
</tr>
</tbody>
</table>

From the above table it is seen that all the institute/organization has begun collecting digital resources and have IT personnel to look after. UNISWA (University of Swaziland) library has started an Institutional repository recently to archive the University’s creative output. The service is offered to the University community for the management and dissemination of digital material created by them. It is an organizational commitment to the stewardship of the digital materials including long term preservation.

Table 2: Which, if any, of the Following IT Applications for Digital Collections Management does your Institution Currently Support?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital imaging</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Collection management system</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Publicly searchable collections database</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Digital asset management system</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 2 shows which IT application supports the digital collections management. SwaziLII and National Library support almost all application except Digital imaging for SwaziLII and Digital asset management system for National Library. The UNISWA library supports Collection management system and the publicly searchable collections database which is an e-resource of the library.
Table 3: Does Your Institution have a Digital Collection Initiative?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>Yes/ No</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>Yes</td>
</tr>
<tr>
<td>SwaziLII</td>
<td>Yes</td>
</tr>
<tr>
<td>National Library</td>
<td>Yes</td>
</tr>
</tbody>
</table>

All the institutes/ organization in the study has digital collection initiatives.

Table 4: How Many Staff and Volunteers Work on Your Digital Collection Initiative?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>No. of Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>4</td>
</tr>
<tr>
<td>SwaziLII</td>
<td>3</td>
</tr>
<tr>
<td>National Library</td>
<td>2</td>
</tr>
</tbody>
</table>

The table shows the number of staff and volunteers that work on the digital collection initiatives. UNISWA had the highest number of staff for the digital collection.

Table 5: What are the Sources of Funding for the Digital Collection Initiative?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fees from activities</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Fund raising</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Grants</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Organization’s operating budget</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 5 concerns about the source of funding for the digital collection initiatives UNISWA managed from the Organization’s operating budget which is included in the library budget. SwaziLII manage the funding from the fees from activities, Grants and also from Organization’s operating budget and National Library manages from Fund raising and Grants. SwaziLII are a donor funded project so they are still looking for more sponsors to keep this project a going concern.

Table 6: Have You Adopted Standards or best Practices for Digital Content Creation (Digitization)?

If so, Please Specify the Standards used, Link to the URL if Online or Attach a Copy if the Standards are Locally Customized or if best Practices or Guidelines have been Developed. Yes/ No

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>Yes/ No</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>Yes</td>
<td><a href="http://www.dspace.uniswa.sz:8080/jspui">http://www.dspace.uniswa.sz:8080/jspui</a></td>
</tr>
<tr>
<td>SwaziLII</td>
<td>Yes</td>
<td><a href="http://www.swazilii.org/sz">http://www.swazilii.org/sz</a></td>
</tr>
<tr>
<td>National Library</td>
<td>Yes</td>
<td><a href="http://www.africanlii.org">http://www.africanlii.org</a></td>
</tr>
</tbody>
</table>

All the institutes/organizations have adopted standards for the digital creation and their URL is given above in the table.

Table 7: How are Digital Collections Created?

<table>
<thead>
<tr>
<th>Institutes/ Organization</th>
<th>In-house</th>
<th>Outsourced to a Vendor</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>SwaziLII</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>National Library</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

The digital collections are created In-house only but SwaziLII created both in-house and also outsourced to a vendor. They outsources in South Africa.
Table 8: If Your Institution is Creating or Acquiring Born-Digital Resources, which of the Following Formats are you Creating or Acquiring?

<table>
<thead>
<tr>
<th>Institutes/ Organization</th>
<th>UNISWA</th>
<th>SwaziLI</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoded text (blogs, websites, listservs, PDF documents)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Photography or other still images</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Documentation or research data</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Other numeric data sets.</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

From the above analysis it is seen that the National library is creating the born-digital resources with the entire format. SwaziLI create using all format except with other numeric data sets. But UNISWA acquires with Encoded text only.

Table 9: Why are you Creating/ Acquiring Digital Resources?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>UNISWA</th>
<th>SwaziLI</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>For identification (i.e. collections management, cataloging)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>To generate revenue</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>For marketing and promotion of the institution</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>To increase access to the collection</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>To preserve the original by reducing handling</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>For publication</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>For study and use by local users</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>For study and use by remote users</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

They acquire the digital resources for different reason and purpose but all of them have one main purpose to increase access to the collection.

Table 10: Of all the Items you Consider Belonging to Digital Collections, are the Majority Unique or Replicated in Digital or Analog Versions?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>UNISWA</th>
<th>SwaziLI</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most are unique (single-copy works)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Most are replicated in both digital &amp; analog version</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

The items belonging to digital collection are unique for two institutes i.e. UNISWA and National library. UNISWA do have replication in both digital & analog version. SwaziLI has not identified it.

Table 11: What Formats are you Currently using?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>UNISWA</th>
<th>SwaziLI</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PDF/PDF-A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>TXT</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>HTML</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The formats currently used are different UNISWA used only PDF/PDF-A whereas SwaziLI used JPEG and HTML and the National library uses mix format PDF/PDF-A TXT and HTML.

Table 12: How do you Provide Access to your Digital Collections?

<table>
<thead>
<tr>
<th>Institute/ Organization</th>
<th>UNISWA</th>
<th>SwaziLI</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Through a website associated with our organization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Through a website associated with another organization.</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Through a content management system</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Local online access/intranet</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
The digital collections are provided through different modes. All of them provide access through a website associated with the parent organization. The rest depend on their management system whether they provide through a content management system or local online access or through a website associated with another organization like that of SwaziLII.

Table 13: Approximately what Percentage of your Digital Resources is Available Online?

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>2</td>
</tr>
<tr>
<td>SwaziLII</td>
<td>60</td>
</tr>
<tr>
<td>National Library</td>
<td>100</td>
</tr>
</tbody>
</table>

Based on the above analysis UNISWA has the least percentage of digital resources available online. The Institutional Repository of UNISWA is still in the infant stage. Most of the law collections are the thesis and dissertations of the students. The following graph shows the clear picture of the percentage of the digital resources available on line of the three institutes.

![Graph Showing the Percentage of Digital Resources Available Online](image)

Table 14: Digital Resources Require Metadata for Discovery, Access, Management and Preservation which Types of Metadata do you Create for your Digital Collections?

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive metadata (title, subject)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Administrative metadata (access privileges, rights, ownership of material)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

All of them create descriptive metadata (title, subject) which is the most important for discovery, access, management and preservation for digital resources. UNISWA and SwaziLII they create the Administrative metadata (access privileges, rights, ownership of material) also.

Table 15: Which Metadata Schemas are you using?

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARC</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Locally developed schema</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

UNISWA and SwaziLII used the MARC whereas national library used the locally developed schema.
**Table 16: How does your Institution Fund or Intend to Fund Preservation of Digital Collections?**

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Through the preservation budget</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Through another line in the operating budget</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Through the information resource budget</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Through grants</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

UNISWA has the funding through the preservation budget which is included in the overall budget of the library. SwaziLII intend to fund from another line in the operating budget and also from the information resource budget but National Library fund from another line in the operating budget and also through grants.

**Table 17: Which of the Following Digital Preservation Strategies has your Institution Implemented?**

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migration</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data backup</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Refreshing data</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The UNISWA has the strategies of data backup for digital preservation and SwaziLII has the strategies of migration but the National Library has all migration, data backup and refreshing data.

**Table 18: Are you Outsourcing to a Digital Preservation Service? Yes/No**

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNISWA</td>
<td>No</td>
</tr>
<tr>
<td>SwaziLII</td>
<td>No</td>
</tr>
<tr>
<td>National Library</td>
<td>No</td>
</tr>
</tbody>
</table>

From the above table none of the Institute/Organization is not outsourcing to a digital preservation service.

**Table 19: If you are Managing Digital Collections Locally, what Media are used for Storage?**

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Removable magnetic media</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Online magnetic media</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Optical media (CD, DVD)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tape</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

The media used for storage are all different in all the Institute/Organization. UNISWA uses removable magnetic media, online magnetic media and also optical media (CD, DVD). SwaziLII used only online magnetic media. National Library didn’t disclose what media they used for storing the digital collection.

**Table 20: If you Back up your Files, where are they Stored?**

<table>
<thead>
<tr>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-house</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>By a partner organization</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Outsourced to a storage site</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

The backup files are stored In-house in UNISWA and SwaziLII but in the National Library they outsourced to a storage site.
Table 21: At what Frequency do you Back up the Files?

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Institute/Organization</th>
<th>UNISWA</th>
<th>SwaziLII</th>
<th>National Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Once a week</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Daily</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Once a month</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

Frequency of backup depends on the parent organization. UNISWA do it once a month while SwaziLII do once a week and the National library do it every day.

**RECOMMENDATION**

The following recommendations are therefore proffered to enhance it:

- All the legal depot in the country should form a consortia.
- Digitization of legal documents is must.
- Librarian should be encouraged to buildup IR for their born digital materials.

**CONCLUSION**

Law is central to the development process, law reform could lead social change, and itself could be an engine of change. It promulgates the idea that diffusion of developed law to developing countries would aid in its modernization. They saw “law as a force which can be molded and manipulated to alter human behavior and achieve development” and so transplantation of new law according to law in developed society is a good way to perceive need for rapid and directed change. Frankly speaking, our current digital preservation strategies and systems are imperfect – and they most likely will never be perfected. That's because digital preservation is a field that will be in a constant state of change and flux for as long as technology continues to progress. Yet, tremendous strides have been made over the past decade to stave off the dreaded digital dark age, and libraries today have a number of viable tools, services, and best practices at our disposal for the preservation of digital content.

**REFERENCES**

Role of Digital Forensics in Digital Preservation as per the Indian Legal Requirements

Yogendra Tank¹, Bhavesh Gabani², Nikhil Padhiyar³ and Dinesh Katre⁴

¹,²,³Gujarat Technological University, Ahmedabad, India
⁴Centre for Development of Advanced Computing, Pune-411007, India
E-mail: ¹yogendra1911@gmail.com, ²gabanibhavesh9@gmail.com,
³nbpadhiyar@gmail.com, ⁴dinesh@cdac.in

Abstract—As per the Indian Information Technology Act 2000/2008 it is necessary to capture and preserve the digital evidences of digital information and electronic records to ensure its legal admissibility. Therefore, the authors of this paper have explored the digital forensics for e-mail, hard disk, and Android cell phone for capturing various digital evidences. The findings of this process are corroborated to infer and ascertain the provenance, integrity, authenticity and reliability which are key aspects of digital preservation as per the OAIS. It is found that digital forensics can be used as a part of the digital preservation process to meet the Indian legal requirements.

INTRODUCTION

Digital preservation of e-government records or any form of digital information stored on storage media or court case evidence such as mobile data requires to establish its provenance, integrity, authenticity and reliability [1,2,3,5] for legal admissibility. Same parameters are also important for digital preservation in addition to the parameter of "long term usability" of the digital information, as stated in the ISO 14721:2012 Open Archival Information System (OAIS) Reference Model. The legal requirements and digital preservation requirements are also defined in the institutional digital preservation policies [18]. Digital signatures, which are normally used for establishing the authenticity of digital information, have limited validity. Therefore, supplementary digital evidences become useful in establishing the authenticity which can be captured, preserved and corroborated when that can help in inferring or establishing the provenance, integrity, authenticity and reliability of the digital information and thus it can become legally admissible too. Therefore, in the next sub-section, we briefly discuss the requirements for preservation and legal admissibility in the Indian Information Technology Act 2000/2008.
Information Technology Act 2000/2008

The IT Act 2000/2008 provides very encompassing definitions for data, information and electronic records which covers various forms of electronic inputs and outputs. It also encompasses the digital information producing devices in generic terms and the medium of carriers and transmission in terms of networks and storages. In section 7, it specifically highlights the need to retain the electronic records or data or information as per the retention schedules and along with the details which will facilitate the identification of the origin, destination, date and time of despatch or receipt of such electronic records, data or information. The IT Act also requires that the electronic record is retained in the format in which it was originally generated, sent or received or in a format which can be demonstrated to represent accurately the information originally generated, sent or received. Section 65B of IT Act defines the conditions for legal admissibility of electronic records. It specifies the need to prove that the computer output containing the information was produced by the same computer during the period. There is also need to establish that the person who used the said computer and was having lawful control over it. Therefore, it is necessary to capture such evidences and preserve it along with digital information that requires to be retained for long durations.

International Trends

As per the DPC Technology Watch Report on Digital Forensics for Preservation, digital forensics is an important part of digital curation process and digital archeology [7]. University of British Columbia in Canada has introduced the subject of digital records forensics for their archival and information science curriculum [8]. The Guidelines on Cell Phone Forensics by NIST, provides specific guidelines for preservation of forensic evidence which includes securing, documenting, evaluating the scene and capturing, storing, packaging, and the custody of such digital evidence [9]. In India, digital forensics is increasingly being used for investigation of cyber crimes but its application for digital preservation is a new concept.

Methodology

Therefore, to capture various digital evidences using the forensics tools, we have taken up three different scenarios or case studies namely email forensics, disk forensics and cell phone forensics. Our main purpose of investigation in this research paper is to find the digital evidences, metadata and all extractable information that can help in preservation. Our approach is broadly defined as under:

- Apply open source digital forensics tools on the sample device / data to capture various evidences.
- Analyze and corroborate the evidences to establish the aspects of preservation in terms of provenance, integrity, authenticity and reliability of the digital information.
DIGITAL FORENSICS FOR E-MAIL PRESERVATION

Evidences in e-mail Header

As per the Indian Right To Information Act [4], the government organizations will also require to preserve the e-mail transactions within an organization so that the specific e-mails could be searched and reproduced in legally admissible manner to fulfill the right to information where required. Department of Electronics and Information Technology (DeitY), Government of India is already formulating the e-mail policy for all central and state government organizations in India for the purpose of e-mail data protection, which will also be subjected to preservation in various cases. E-mail messages consists of two major sections namely Header and Body, formats for which are standardized in RFC5322 [6]. As per the IT ACT, it is necessary to establish the origin, destination, date and time of despatch or receipt of an e-mail. The traces pertaining to these aspects are found in the header of the e-mail.

Origin, Date, Time, Destination and Identifier

It is a common knowledge to most of the users that an e-mail header captures the details such as from (sender), to (receivers), and date (time). If every e-mail is considered as a record then it must have a unique record identifier, a unique number by the specific e-mail could be identified. The header of an e-mail also includes the “Message-ID:” an automatically generated unique ID of the message. The sending email server generates a globally unique ID for each email going out from it, which generally depends on the algorithm used by the server. An example of the “Message-ID:” is shown below:

Message-ID: <52690F&F.6090308@cdac.in>

Trace Information

Further SMTP defines the trace information of the message, which is saved in header using following two fields. “Received:” when an SMTP server or Mail Transfer Agent (MTA) accepts or forwards an email message, it inserts trace information here, these field information is used in tracing the email back to the sender. This field generally includes name of MTA from which the email is received, IP address, the name of receiving MTA, protocol used in transfer and the time stamp of the transaction.

```
Received: from mailhub.pune.cdac.in (mailhub.pune.cdac.in [10.208.1.7])
   by mailhub2.pune.cdac.in (5.14.4/5.14.4) with ESMTP id z9coOg1q009464
   for <yogendra1911@gmail.com>; Thu, 24 Oct 2013 17:49:42 +0330
```

“Return-Path:” when the receiving SMTP server (Mail Delivery Agent – MDA) makes the final delivery of the email (to client), it inserts this field generally on the top of header.

```
Return-Path: <chandarakants@cdac.in>
```
Integrity

“Content-MD5:” MD5 checksum of the content, 128bit digest converted in base64 helps in checking the integrity of the email body data. The checksum is helpful in proving the integrity of the email.

Content-MD5: Q2h1Y2sgSW50ZWdyaXR5IQ==

Domain Authentication

“DKIM-Signature:” A domain of the sender can sign the email with its private key, claiming the originating domain of the email document. Used in verifying the domain of the sender. Not all the email services provide DomainKeys Identified Mail (DKIM) signature. Currently Yahoo, Gmail, AOL and Fastmail have implemented this feature. Any mail from these organizations should carry a DKIM Signature in the header of the email.

User Authentication

It is possible to digitally sign the outgoing messages which authenticates the identity of the sender. Digital signatures is mainly used for authentication of the source of the messages (i.e. the sender), along with integrity check of the message, and non-repudiation.

Capturing e-mail for Preservation

Single e-mail

EML which is basically ASCII text file, typically stores each message in a single file, and attachments may either be included as MIME content (base64 encoding) in message or saved in a separate file, referenced from a marker in the EML file. The EML format considering its open specification is suitable for e-mail preservation. XML is also another suitable file format for preservation of e-mail, which allows self-description. However, the preservation of attachments cannot be guaranteed unless the attached files are in compliance with open and standards based file formats. It is necessary to alarm the gulf which surrounds the attachments of the e-mail, which entirely depends on the sender.

e-mail Account

Most of e-mail clients support MBOX file format in which the e-mail account can be stored. MBOX saves all the mails in single file with MIME contents stored directly in the file in base64 encoding. The problem with MBOX file is that the corruption in file may harm clients' ability to read individual email or even entire folder.
Single e-mail can be preserved in XML, EML, TXT and PDF/A file formats. PDF/A provides an advantage of fixing the information contained in the e-mail, which prohibits editing or modification. However, for preservation of entire e-mail account MBOX is useful.

**DIGITAL FORENSICS FOR DISK PRESERVATION**

**Digital Forensics Aspects of Disk Image Capturing**

The digital repositories often receive miscellaneous electronic records, unorganized data from organizations for preservation purpose, which is deposited in the form of digital storage media like hard disks, flash drives, CD/DVDs, etc. In such cases, it is not possible to ingest the electronic records or data individually in an Open Archival Information System (OAIS) considering the unstructured or unorganized or miscellaneous nature of the data stored in the disk. In such situations, due to the obligations and need of preservation, it is often considered to capture the disk image of the storage media as is and then ingested into an OAIS. The digital forensics aspects of this process are discussed below -

**Disk Image**

Bit stream copy of storage media in forensics disk imaging format either in raw (dd) or advance forensics format (AFF) is used to prove reliability of storage media. A disk image is an exact replica of the source media which contains any slack space, unallocated space, and other metadata of the disk volume. Before taking bit stream images using any tools, care should be taken that the device is write protected to prohibit any changes.

**Serial Number**

The serial number of the storage media like hard disk or flash drives is extremely important for proving the source of the disk image, which is captured in the metadata.

**Hash Values**

Fixed size hash value derived from original disk image using irreversible hashing algorithms like MD5, SHA1 or SHA256 can be used to prove the integrity of the original source of information.

**Chain of Custody Information**

Most of the forensic disk image capturing tools, allow you to allocate a case number, which a unique identification number is given to disk image as an evidence. It maintains the list of tools used to collecting the digital evidences, examiner name, short description of evidence, the path where the evidence is stored initially, time of acquisition of evidence, and MAC times (b, m, a, c times) which is latest modification time of the file, last access time of the file, change time means metadata was changed like the file’s permissions or ownership were modified and birth or creation time means the time the image file was created. This information is very vital in proving the provenance, ownership of disk image which is being preserved.
Examples of Open Source Disk Imaging Tools

There are several open source forensics disk imaging tools available. Guymager is one of the open source tools used for media acquisition, which runs under Linux and it supports Multi-threaded and pipelined design. It can generate disk image in raw (dd), EWF (E01) or AFF formats.

![Screenshot of the AFF Image Metadata Captured by Guymager. As Part of the Notes Seen in the Screenshot, the Serial Number of Storage Media is Visible. Chain of Custody Information as Image Metadata and MD5 Hash Value is also Seen](image1)

Fiwalk is a library to extract file-level metadata from disk images which facilitates the identification of the origin, its MAC time values, etc. and output result in DFXML or key/value pairs which can be directly preserved in archive.

![Screenshot of DFXML Output from Fiwalk. MAC Time, File and Partition Information, Its Hash Value and Other Metadata Related to Particular File is also Visible](image2)
Raw format (dd) images are simple copies of storage media but it does not include metadata, we have to collect metadata separately but transportation of this format image is simple, we can use piping over netcat or ssh to transfer this kind of file over network. Advanced Forensic Format (AFF) is an extensible open format which includes the raw bit stream disk image and metadata. It also allows disk images to be compressed with a variety of algorithms [15,16,17].

It is important to note that there is no standardized format for forensics disk images. Disk imaging process is very time consuming and costly due to additional storage space requirements for copying and replicating, etc.

**DIGITAL FORENSICS FOR CELL PHONE PRESERVATION**

The cell phone device contains born digital information such as Contacts, Call History, SMS, MMS, e-mail, Device info, photographs, audio, video, passwords, address book, browser’s data, social networking information, etc. The cell phones are prone to misuse due to increasing computational power. For example, such mobile device can be hardened and used for criminal and destructive purposes [11] like data gathering of credit card scanning and for trigger bombs. In such criminal cases, the cell phones used in the crime need to preserved as a vital evidence for several years. Therefore, in this case study, we have used Android cell phone for forensic application.

**Digital Forensics for Android Cell Phones**

In Android cell phones the data can be found in number of locations such as NAND flash, removable storage and SIM cards. Not all phones are created equal, and what is extractable from a device is dependent on its capabilities [13]. Hence there are different methods and tools that can be used to forensically examine the device and extract the information. One of the essential step is to root the android device if it is not rooted. Root means to get root or super user access to that device. There are different methods or some exploits available for different device but it is riskier to root the device because of integrity issue. The forensic acquisition process has no standard method, but using multiple methods or combinations of different tools one can achieve it. So, only method we can say as manual extraction. This acquisition and extraction of forensics image of mobile phone can be scripted for desired use and application to fulfill the requirements using scripting language.

**Process**

Android device stores data into NAND flash memory. This memory contains logical block partitions for different purposes. These NAND memory uses YAFFS (Yet Another Flash File System), that is also open source and fully supports flash memory [12]. Android uses MTD(Memory Technology Device) to address different partitions on single chip of memory. Mtd is generally stored in/proc/mtd, and returns existing partition table of system. It depends on Version and device manufacture also. So aim is to locate the mount point and get image
of that device. There are different mount points such as system, cache, data or userdata, sdcard or extsd, etc. Among of these, userdata contains useful information. Using adb and dd tools this image can be captured and stored for preservation. Later this dd image can be used to extract useful information, and also this extracted information can be ingested along with image for preservation. Android provides rich and powerful API for development. There are some open source forensics applications which provide automated extraction from android mobile and generate well formatted report. AF Logical Open Source Edition is a free version, which pulls all available MMS, SMS, Contacts, and Call Logs from the given Android device. There are also other methods by which these details can be extracted and that can be converted in xml format for preservation.

**Process Outcome**

Bulk Extractor extracts information such as Email addresses, credit card numbers, exif information from jpeg, phone numbers, search history, urls, etc that might be useful for law enforcement authority. Bulk extractor also shows histogram for each search result that helps analysis of large amount data. AF Logical OSE gives trustworthy and well formatted information. It also produce one info.xml file that stores metadata. CSV file contains call numbers, call log, sms, mms, etc. that acceptable for preservation. Here as described in figure 3 we can prove identification of device for further related work with evidence. It can yield IMEI or MEID number, which is unique to device. ICCID is SIM card’s unique number, which is links to service provider and related useful information. Apart from that Device’s information such as maker, model, etc can be yield.

As the mobile industry is emerging in nature and hence limited number of tools that supports forensics for preservation. Also root the device is one biggest challenge. There are also some proprietary tools such as FTK MPE+, viaForensics viaExtract, Oxygen Forensic Suit, etc. Also AF Logical has Law Enforcement version that gives more features than OSE version and it is only be used for Law Enforcement agency.

In Android phone, for extraction different methods are there and each method yields different data. Android stores every information in SQLite database. This database can be used to extract useful information, and if multiple source of information yield same result then we can rely on such evidence. Such extracted data including DFXML [10,14], metadata, and related information, can be used for producing an Archival Information Package of the mobile data (AIP) preservation purpose as part of the ingest procedure in the OAIS.
CORROBORATING THE FORENSIC INFORMATION WITH PRESERVATION REQUIREMENTS

In this section, we have tried to corroborate the captured digital evidences for legal and digital preservation requirements. The metadata extracted using the forensic tools can also be processed for search, retrieval, sorting of digital information.

CONCLUSION

- As shown in Table 1, it is possible to capture various digital evidences by applying digital forensics methods to infer the provenance, integrity, authenticity and reliability of e-mail, disk, and cell phone.

- It is necessary to standardize the file formats for storing / preserving e-mail, forensic metadata and disk images.
• The forensic metadata will also be useful in identifying the obsolete or proprietary file formats, certain (limited) aspects of representation information related with the digital objects in the disk images for digital migration, representation or emulation if required. The future usability of all the data stored in the disk images cannot be guaranteed as it may be using several proprietary formats unless it is migrated into open and standardized file formats.

• Digital forensics should be incorporated in the procedures for preserving the electronic records of evidentiary value to meet the requirements of IT Act 2000/2008.

Table 1: Summary of the Findings from Digital Forensics of E-mail, Disk and Cell Phone from Preservation Perspective

<table>
<thead>
<tr>
<th>Aspects of Preservation</th>
<th>e-Mail Forensics</th>
<th>Disk Forensics</th>
<th>Android Cell Phone Forensics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provenance</td>
<td>IP Addresses</td>
<td>Serial number of the storage media</td>
<td>IMEI or MEID number, Product details (handset) ICCID (SIM card’s unique number)</td>
</tr>
<tr>
<td>Authenticity</td>
<td>Sender’s Digital Signature</td>
<td></td>
<td>SIM Card details</td>
</tr>
<tr>
<td>Inferential authenticity (date, time, provenance, actions)</td>
<td>Time stamp, Message-ID, Trace information, Information about sender, receivers</td>
<td>Time stamp, Message-ID, Trace information, Information about sender, receivers</td>
<td>Text file containing e-mail addresses, credit card numbers, exif information from jpeg, phone numbers, search history, urls, CSV file containing call numbers, call log, sms, mms Corroborate the records from the mobile service provider</td>
</tr>
<tr>
<td>Reliability</td>
<td>DKIM-Signature</td>
<td>Same information can be extracted using different methods and compared</td>
<td>Same information can be extracted using different methods and compared</td>
</tr>
</tbody>
</table>

REFERENCES

Role of Digital Forensics in Digital Preservation as per the Indian Legal Requirements ♦ 247

[19] XENA (Xrn1 Electronic Normalising for Archives).
[20] dd
  UNIX tool that copies data from one file to another.
[21] AF Logical OSE
  https://viaforensics.com/resources/tools/android-forensics-tool/
[22] ADB (Android Debug Bridge)
[23] Fiwalk
  http://www.forensicswiki.org/wiki/Fiwalk
[24] Bulk Extractor
  http://www.forensicswiki.org/wiki/Bulk_extractor_viaForensics_viaExtract
    https://viaforensics.com/products/viaextract/
[25] guymager
Value from Digital Preservation
Preservation and Value: Support based on a Coherent View of Preservation

David Giaretta
Alliance for Permanent Access
E-mail: director@alliancepermanentaccess.org

Abstract—The recognition of the importance of the vast amounts of data which are being collected and the potential of using it for the benefit of society is being reflected in many policies and activities. We face a deluge of newly created information every day. Some of this could be re-created at any time, although costs may hinder this in many cases; however much of the data cannot be re-created since they depend on the precise circumstances at their creation, for example Earth Observation data. If such sets of data are to contribute to the value they must be preserved. However the actual value of any particular piece of data is far from clear and yet preservation requires the expenditure of resources.

The value of a piece of data can very often be enhanced, and new information generated, by combining with other data, perhaps from many different sources, disciplines and user communities.

Repositories therefore will need to demonstrate not only that they serve their current users well, but also that they can serve wider communities now and into the future.

This paper discusses the work that carried out to create a centre of excellence which will be able to help repositories preserve and add value to their holdings, based on a coherent and consistent view of preservation.

Keywords: Digital Preservation

INTRODUCTION

The “Riding the Wave” report [[1]] addressed the question of how Europe can gain from the rising tide of scientific data. McKinsey [[2]] estimates of the value of big data include €250 Billion per year for Europe’s public sector administration. To estimate this value they considered “only those actions that essentially depend on the use of big data – i.e. actions where the use of big data is necessary (but usually not sufficient) to execute a particular lever.” In order to extract the value a number of issues must be addressed including legal, security related, technological, and organisational. McKinsey noted that “legacy systems and incompatible standards and formats too often prevent the integration of data and the more sophisticated analytics that create value from big data” and that there is an increasing “need to integrate information from multiple sources.”
A good deal of the observational data is unique while much of the rest is too costly to be re-created. If such sets of data are to contribute to the value they must be preserved. However the actual value of any particular piece of data is far from clear and yet preservation requires the expenditure of resources.

The report of the Blue Ribbon Task Force [[3]] addressed this issue, and while not providing the solution it did provide a number of useful concepts and suggested steps to take, including the need to “leave open options for future stakeholders” recognizing that if the current option is “unsustainable over time, decision makers would need to revisit their options and make a different choice.”

Perhaps most importantly the report noted that “When the future value of an asset is uncertain, the likelihood of long-term preservation can well depend on current use cases. Preserving materials with clearly defined current uses implicitly creates the option of having the assets available for as-yet-unknown uses that may emerge in the future. Therefore, to the extent possible, value propositions should focus on the benefits generated for current users—about whom decision makers are understandably most concerned—rather than focusing too much on benefits to future generations and unknown future uses. But sometimes this is just not enough.”

An important question on which this paper attempts to cast some light is the way in which one might be able to increase the potential value of data, and what support is available to do this.

**Demand for Value**

We argue that the steps to take for preservation can also help increase potential value. This is because in addition to the problems arising from “legacy systems and incompatible standards and formats”, preservation must counter other threats which the PARSE.Insight project [[4]] identified. Amongst these were that “Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved” and “The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity”.

**Opportunities for Data**

The first type of threat mentioned above is usually regarded as irrelevant for documents or images, in that simply being able to continue to render the document or image is regarded as adequate in terms of preservation. However it most certainly is not adequate for data, where the meaning of the numbers is an absolutely vital consideration, otherwise one may try to combine apples with oranges i.e. combine two incompatible quantities, thereby creating nonsense instead of something of value. Even for documents and images it may be that in order to preserve the meaning more than rendering is required, for example explanation of specialist terms or concepts.
OAIS (ISO 14721) [[5]] introduced a number of concepts, in particular Representation Information, defined as the “information that maps a Data Object into more meaningful concepts.” This includes not just detailed descriptions of the format but also information about the semantics of the data and software which can operate on the data.

Ensuring the usability of the data requires, among other things, ensuring there is adequate Representation Information for the Designated Community, as its knowledge base changes over time.

Further Representation Information can be added beyond that needed by the Designated Community, if one wishes to ensure that another, perhaps wider, community of users can understand and use the data. Note that there need not be any commitment to maintain understandability and usability for this wider community.

What Works in Preservation

Just as the above argues that what works for (simple) documents, in terms of preservation, does not work for the preservation of data, so we need to ask what works for preserving different sorts of data. Clearly if we choose techniques that do not work then we will run a serious risk of failing to preserve, and in particular failing to leave options for future stakeholders. For example if we choose a technique for preserving some digitally encoded information but future stakeholders discover that we have managed to preserve the numbers but have failed to preserve the meaning of those numbers, then the future users will have few or no options for creating value from that data. Furthermore the resources used in that failed attempt of preservation will themselves have been wasted.

Preservation involves not just technical solutions but also for example estimates of the resources needed, the legal requirements and the possible value opportunities. The interplay between so many aspects and activities makes preservation decisions very complex. Yet most research in preservation has tended to involve studies of particular aspects while ignoring many more aspects.

The latter part of this paper describes some of the work which has been undertaken to collect the evidence needed to help in making the correct decisions about preservation within a consistent and coherent framework.

Defragmenting Preservation

The APARSEN project [[6]] has worked since the start of 2010 to bring together various studies in preservation. It has sought to perform this integration in a staged way, bringing together researchers in digital preservation, vendors, manufacturers, international and national scientific repositories, major national and international libraries.
Silos and Staged Integration

The silos which APARSEN divided the digital preservation world are shown below, grouped into 4 topics, trust, sustainability, usability and access.

Table 1: APARSEN Topics and Separate Areas

| Trust          | • Certification of repositories  
|                | • Reputation and trustability of datasets, publications and people  
|                | • Authenticity  |
| Sustainability | • Business cases  
|                | • Preservation services  
|                | • Cost/benefit analysis  
|                | • Storage solutions  
|                | • Scalability  |
| Usability      | • Intelligibility  
|                | • Use by common tools  
|                | • Cross domain usability  
|                | • Interoperability  |
| Access         | • Identification of datasets, publication, people  
|                | • Rights and responsibilities  
|                | • Policies and governance  |

Within each silo the work which has been carried out inside or outside the partner organisations has been collected and compared. The aim is to see which elements are correct and to identify the commonalities. In many cases there have been advocates, or even the originators, of particular approaches within the consortium. To resolve conflicts the strategy adopted was to agree that all the approaches probably had some validity, but in each case that validity is limited. The issue therefore was to discover those limits. This was done by examining the evidence.

For example there are a large number of tools which claim to help in preservation. The question about applicability can be resolved by identifying examples where the particular tool has been used successfully, while seeking (much rarer) stories of failures and in some cases conducting thought experiments, where it would not work. For example if a tool ignores the semantics of a dataset then it cannot by itself preserve data.

In the case of estimating costs of preservation, there are many cost models available in the literature. Within the cost model silo attempts have been made to test many of these models and it seems that none of the models can be tested using data from other sources. Moreover, using a landscape based on the metrics set out in ISO 16363:2012, Audit and Certification of Candidate Trustworthy Repositories [[7]], it appears that the existing cost models ignore many important factors.

The next level of integration was in terms of the topics. For example trust has many interrelated aspects. Trust is fundamental to the working of society – in particular when it comes to unfamiliar digitally encoded information, especially when it has passed through several hands over a long period of time. To ascertain whether a digital object can be trusted, one might ask:

- Has it been preserved properly?
- Is it of high quality?
Has it been changed in some way?
Does the pointer get me to the right object?

Providing coherent answers to these questions requires us to bring together many of the silos mentioned above. A readable summary of this work is available on the APA/APARSEN web site [8].

For the sustainability topic one can argue that those who have responsibility for preservation may well have asked themselves the following:

- How long must the resources be committed?
- How can I justify the resources needed for digital preservation?
- How can I estimate the resources needed and how can I keep these to a minimum?
- How can I plan to cope as the volume increases over time?

The integration in each case brings the answers to these questions together.

**Value and Usability**

At the time of writing the topics of usability and access had not been integrated. However an important conclusion can be drawn, based on the comparison made above of documents vs data. As noted from McKinsey, integration of data, from many sources, and the more sophisticated analytics creates value from big data. Therefore one can increase the value of data by ensuring it can be used more widely, and one can do this by adding Representation Information.

**OVERALL INTEGRATION**

Bringing together all the considerations we can draw a diagram which summarises an overall view.

Fig. 1 illustrates the basic sequence of activities to implement a sustainable business process centred in the preservation of a digital object, to be incorporated in the overall business cycle of organisation responsible for securing the future usage of such assets.

Fig. 1: Cycle of Value for Preservation
Note that the focus here is on preservation. There are a number of other models ([9],[10],[11]) with which one may be tempted to make comparisons; these tend to be focussed on the creation of digital objects and the publication of results, or the academic lifecycle, but those models tend to ignore the business model aspects.

It should be borne in mind that in reality there may be a number of iterations. For example to create a Business case, Value may be re-visited and revised as may be Usability; these iterations are omitted in the flow shown above for the sake of clarity.

The activities may be summarised as follows:

**Preserve the Object by a Variety of Sub-processes**

- Ingest.
- Store.
- Plan preservation, including identifying the designated community (ideally this should be done at the earliest opportunity—certainly before the creation of the digital objects, if we want to secure the best conditions for future usage and we must secure a proper value justification to secure financial resources flows).
- The basic steps in preservation to counter changes are:
  - Create adequate Representation Information for the Designated Community and/or.
  - Transform to another format if necessary or.
  - If preservation cannot be carried on by the current organisation then hand over to the next organisation in the chain of preservation.
- Evidence about the authenticity of the digital objects must also be maintained, especially when the objects are transformed or handed over.
- Confirmation of the quality of preservation can come from an Audit (with possible certification).

**Usability**

- Additional Representation Information may be created to enable a broader set of users to use and understand the digitally encoded information
  - Other communities may use different analysis tools and it may be convenient to transform the digital object to a more convenient format. This will itself require its own Representation Information; the semantic RepInfo may be unchanged but new structural RepInfo will certainly be needed.
• The digital objects should also be discoverable in some sensible way – bearing in mind that some information will be publicly available whereas other information will be restricted.

**Value**

• The portfolio of Value proposition/s will provide the core of the answers to “who pays and why?”

• Value propositions must be created by the identification, classification and quantification of the expected benefits which may be obtained by the targeted communities of customers and users from the continuous usage of the preserved objects, which in turn depends on the needs of the users and the usability conditions created for such preserved objects.

• The objects will probably be more useful to one type of user community than to another, and this may change over time. These differences and changes must be addressed by a portfolio of Value propositions (as well as by the design and implementation of adequate business models).

• Rights may be associated with the objects, perhaps arising from the value or potential value of the object.

**Business Case**

• This is needed to justify: the need for objects to be preserved, the potential benefits derived of their usage, the costs involved in the preservation, as well as other resources required for preservation.

• It will be embedded within a particular business model.

• There will almost certainly be options for trade-offs between costs, risks and capabilities.

**Business Model**

• Decisions about the mix of sources providing the financial resources required for implementing and operating the preservation business process will be based on the characteristics of the users and customers base (the target groups), the competition in the provision of the preserved assets as well as in the nature and dynamics of the formulated business case.

• The resources may be used at the very start to create new digital objects, which will presumably have been created for a specific purpose and which then may be either disposed of or be preserved.
A selection process will be needed to decide what is to be preserved. This will presumably be based on business case and risk considerations. It may also depend on the interest of other possible curators of the information.

This financial resourcing may be (perhaps should be) part of the budgets needed to create the digital objects. However some or all of the objects created may be disposed of rather than preserved.

Each of these steps will be assisted by the use of tools and/or services, such as the ones the APA/ APARSEN Centre of Excellence (CoE), described in the next section, should be able to offer.

The underpinning components are first the use of a consistent terminology ([12]), the OAIS terminology with extensions to cover those aspects outside the OAIS remit and second the training modules covering all aspects of the common vision.

**APA/ APARSEN Centre of Excellence**

Based on the integrated view described in the previous section, a centre of excellence (CoE) is being created, based on the Alliance for Permanent Access (APA) ([13]), extending it and adding services, consultancy, tools and training to the existing membership organisation. It is important to note that much of these offerings will be provided by the member organisations, leveraging the in-depth knowledge that they possess. Because of the importance of the contribution of members, the European Union uses the term “virtual centre of excellence”.

The SCIDIP-ES ([14]) project is developing an important set of tools and services which were proven and prototyped in the CASPAR ([15]) project, and which address threats to digital preservation. These threats were confirmed in the PARSE.Insight project, through very large surveys, as being the ones which are recognised by researchers, data managers and publishers across disciplines and around the world.

The SCIDIP-ES tools and services are also ones which can help address the types of improvement which the test audits undertaken during the finalisation of ISO 16363 (see other papers in this conference) showed to be rather commonly required. These included the construction of Archival Information Packages, defining the Designated Community, ensuring that there is adequate Representation Information and ensuring that adequate evidence is collected to justify claims about authenticity. It is worth underlining the point that the SCIDIP-ES tools and services are open source and are designed to supplement existing repository systems to help them improve their practices and share the effort of preservation.

The SCIDIP-ES software provides a way to add value to digital objects by increasing usability through adding extra Representation Information which enables those objects to be usable by people and systems for which the objects are unfamiliar. Fig. 2 provides an indication of the way in which the contributions of outcomes of the APARSEN, SCIDIP-ES work as well as the expertise from members, are helping to provide capabilities and expertise in depth for the tools, services, training and consultancy offered by the APA.
CONCLUSION

Value and preservation are closely linked. We have argued here that the techniques that are needed for preservation, namely adding Representation Information, allow one to add value to data. This can help to satisfy the demands for value from data as well as justify the resources needed for preservation while one strives to leave options for future stakeholders.

At the same time a coherent and cost-effective approach must be taken to preservation.

The APA centre of excellence is designed to provide a coherent set of tools, services, advice and training which can help repositories to improve the way in which they preserve their digital holdings, helping them to become a more trustworthy repository. It will of course work closely with, and we hope will complement, the Centre of Excellence for Digital Preservation, C-DAC, India, and others.

ACKNOWLEDGEMENT

The work reported here is result of the efforts of the large number of organisations and people involved in APARSEN and SCIDIP-ES.

APARSEN and SCIDIP-ES are funded by the Seventh Framework Program (FP7) of the European Commission (EC) under the Grant Agreements 269977 and 283401 respectively.
REFERENCES

[4] PARSE.Insight project http://www.parse-insight.eu, and documents are also available on the APA web site
[14] Science Data Infrastructures for Preservation – with an initial focus on Earth Science (SCIDIP-ES) – information is available on the web site http://www.scidip-es.eu
[15] CASPAR project website http://www.casparpreserves.eu, and documents are also available on the APA web site.
Author Index

**A**
- Abichandani, Payal, 220
- Ambacher, Bruce, 3

**B**
- Behara, Suman, 47
- Bothale, Vinod, 128
- Brocks, Holger, 78

**C**
- Chaubey, O.N., 169
- Conrad, Mark, 3
- Crompton, Shirley, 78

**D**
- Devi, Thiym Satyabati, 228
- Dillo, Ingrid, 72
- Downs, Robert R., 3
- Duranti, Luciana, 23

**E**
- Engel, Felix, 78

**G**
- Gabani, Bhavesh, 237
- Gaikwad, Sandeep, 169
- Garrett, John G., 3
- Giaretta, David, 3, 78, 142, 251
- Graf, Roman, 87
- Griffin, Elizabeth, 133

**H**
- Hemmje, Matthias, 142
- Hughes, John S., 3
- Hussain, T., 202

**J**
- Jansen, Adam, 39
- Jha, Pratapanand, 163

**K**
- Kapre, Milind, 169
- Katre, Dinesh, 11, 47, 169, 202, 237
- Kesavan, R. Venkata, 187
- King, Ross, 87
- Kirchhoff, Amy, 98
- Koriya, Sourabh, 47
- Kumar, K.V. Rathna, 128
- Kumar, M.V. Ravi, 128

**L**
- Lambert, Simon, 3, 142
- Longstreth, Terry, 3

**M**
- Maboreke, D., 198
- Marelli, Fulvio, 78
- Mathabela, Ntombikayise Nomsa, 228
- Matthews, Brian, 78
- Milic-Frayling, Natasa, 71
- Mittal, Neeraj, 169
- Morrissey, Sheila, 98
- Muchefa, L., 198
- Muralikrishnan, S., 128

**N**
- Naik, Srinu, 47
- Nimmo, Emily, 155

**P**
- Padhiyar, Nikhil, 237
- Pawar, Jayshree, 47
- Prakash, Rishi, 220
- Puntamkar, Shashank, 202

**R**
- Rao, V.V.S. Nageswara, 128
- Riestra, Ruben, 142

**S**
- Saminu, Attahiru, 208
- Schlarb, Sven, 87
- Sharma, Ankit, 202
- Sharma, Vakul, 217
- Shiers, J.D., 123
- Shoan, Arif, 78
- Siemer, Barbara, 3
- Sinha, Mukul K., 107
- Sontakke, Minakshi R., 176
- Sunder, Gaur, 58

**T**
- Tank, Yogendra, 237
- Tibbo, Helen R., 3

**W**
- Wadnerkar, Vaishali B., 176
- Witkamp, Paula, 72
- Wittenberg, Kate, 98