Who uses the digital data archive? An exploratory study of DANS

Christine L. Borgman
Department of Information Studies
University of California, Los Angeles
christine.borgman@ucla.edu

Herbert Van de Sompel
Research Library
Los Alamos National Laboratory
herbertv@lanl.gov

Andrea Scharnhorst, Henk van den Berg
Data Archiving and Networked Services (DANS)
The Hague, Netherlands
andrea.scharnhorst@dans.knaw.nl
henk.van.den.berg@dans.knaw.nl

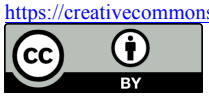
Andrew Treloar
Australian National Data Service
Melbourne, Australia
andrew.treloar@ands.org.au / @atreloar

ABSTRACT

Despite the advances in policy and practice for data sharing, surprisingly little is known about the uses and users of digital data archives, about relationships between users and the staff of data archives, or how these behaviors vary by discipline, geographic region, policy, and other factors. Digital data archives are not a single type of institution, however. They vary widely in organizational structure, mission, collection, funding, and relationships to their users and other stakeholders. We present an exploratory study of DANS, the Digital Archiving and Networked Services of the Netherlands, with the goal of identifying methods for studying the contributors, consumers, and role of archivists in digital data archives. Starting with transaction logs that serve management purposes, we present estimates of the distribution of uses and users of DANS. Units of analysis necessary to study user behavior, such as dataset, file, user, creator, and consumer, are difficult to glean from logs that were not designed for these inquiries. We recommend methods for improving the design of data collection instruments and outline the subsequent phases of our mixed-method research on the uses, users, policy, and practice of digital data archiving.

Keywords
Data, knowledge infrastructure, digital archives, user behavior, transaction log analysis, information policy

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.



INTRODUCTION

As open access to research data becomes a requirement of funding agencies and journals, digital data archives become essential components of scholarly communication and knowledge infrastructures. Data archives can take many forms and have many homes. Some are domain-specific, collecting only data of certain types and formats, such as genome sequences. Some have a broad remit within a domain, for example collecting social science survey data. Others are more generic, collecting surveys, textual documents, static and moving images, audio, and other data types. Data archives range widely in their mission, from long-term preservation to providing immediate access to replication datasets. They also vary in the degree of investment in data curation. Some archives devote days or weeks of professional labor to curating each dataset before deposit; others are "self-curated," accepting data in whatever form submitted, with minimal review. Yet another dimension along which data archives may vary is the longevity of the collection itself, ranging from short-term grant funding to long-term commitments by universities, governments, or other agencies (Borgman, 2015; National Science Board (U.S.), 2005).

Classifying digital data archives is a research project in and of itself, given the variety of entities and stakeholders involved. Only recently has this diverse array of players had forums such as the Research Data Alliance (founded 2013) and Force11 (founded 2011) to discuss common interests, policies, practices, and technologies that span research domains, countries, and communities (Force11, 2014; Research Data Alliance, 2015). Relationships between data archives and their stakeholders also vary greatly. For example, some funding agencies and journals designate a preferred archive to which data should be contributed, others require that a certain type of certified archive be used, and yet others allow a variety of data sharing options.

Libraries have a long history of studying the users of their collections, but these theories and methods have not translated easily into studies of digital services, much less of digital data services. Similarly, studies of the uses of physical archives do not transfer well to data or to digital collections. Research on how scholars create, use, and share data is expanding, but these studies typically end at the point that data are contributed to an archive or repository. The archive and its staff tend to be treated as black boxes.

Given the diversity of digital data archives, it is not surprising that little is known about who contributes data to these archives, who searches and retrieves data from them, and what uses are made of datasets retrieved. Similarly, little is known about the roles that staff of data archives play in acquiring data from potential contributors, in assisting potential data consumers in identifying, retrieving, interpreting, or using data, or in providing other kinds of services. Directories of archives provide only high-level overviews of institutions and services (Committee on Future Career Opportunities and Educational Requirements for Digital Curation, 2015; OpenDOAR, 2014; re3data, 2015).

RESEARCH QUESTIONS

Our three-continent collaboration is centered at DANS, the Digital Archiving and Networked Services of the Netherlands Royal Academy of Arts and Sciences, where three of us (Borgman, Treloar, Van de Sompel) are visiting scholars and two are research staff (Scharnhorst, van den Berg). This poster is the first in a planned series of publications about uses and users of digital data archives, and of methods for studying them. These questions guide our overall research: Who contributes data to DANS? How, when, why, and to what effects? Who consumes data from DANS? How, when, why, and to what effects? What role do archivists at DANS play in acquiring and disseminating data? What kinds of research methods yield what kinds of indicators about uses, users, and staff roles in a digital data archive? Here we report exploratory work on research design challenges for our investigations. Specifically, we assess the degree to which management transaction logs can be used to describe the user population based on the traces they leave in the system and to identify criteria for drawing samples of users to interview.

RESEARCH METHODS

Our research methods are iterative and recursive, beginning with an exploration of available indicators. DANS, like most digital service organizations, maintains transaction logs for purposes of auditing, trouble shooting, and managing information. Users must register with DANS to contribute data or to retrieve most kinds of datasets, thus creating a user database with a small amount of demographic information (e.g., name, institution, email address, discipline). Transaction log data have a long history in information retrieval for studying user behavior. However, logging data must be used with care. Indicators collected for management purposes may not be suitable for research purposes; assumptions and context should be examined carefully. Second, as with any human subjects data, records should be anonymized in accordance with applicable regulations and guidelines; we are following Netherlands practice. Third, logs are unobtrusive indicators that reveal traces of what people do but not why they do so (Borgman, Hirsh, & Hiller, 1996).

The starting point for our exploratory study is to assess the reliability and validity of DANS transaction logs for studying user behavior. We are analyzing transaction logs and the associated database of registered users from three fiscal years, October 2011 through September 2014 (FY 2012-2014), a period of consistent record keeping since the last major system upgrade.

The second part of our exploratory study is to conduct interviews with DANS archivists, and later to interview contributors and consumers of DANS datasets. The interview studies are based at UCLA, for which the usual U.S. human subjects clearances have been obtained. We report here on initial findings of the transaction log analyses, which are helpful to understand both the distribution of user activities and the feasibility of exploiting such files to obtain indicators of user behavior. At the conference, preliminary findings from the interview studies also will be presented.

DANS: Digital Archiving and Networked Services

DANS, founded in 2005 as an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and of the Netherlands Organisation for Scientific Research (NWO), has cumulative responsibility for 50 years of digital research data in the social sciences and humanities from its predecessor organizations. DANS offers multiple services, including NARCIS, the Dutch Research Information System, and EASY, the Electronic Self-Archiving SYstem, for datasets. In 2014, DANS, in partnership with other Dutch data archives and research infrastructure providers, formed Research Data Netherlands, an alliance to promote best practices in data management and preservation. As of May 2015, EASY contains 29,743 published datasets. An EASY dataset is the equivalent of a "collection" in Dublin Core Metadata Initiative terminology. Datasets are tagged with one or more disciplinary classification codes. The majority of datasets in EASY originate in archaeology, for which EASY is a legal deposit archive. The majority of downloaded datasets, however, are from the social sciences, which include census data from Statistics Netherlands. (Akdag Salah et al., 2012; Scharnhorst, Ten Bosch & Doorn, 2012)

RESULTS

We devoted the first year of the project to exploring transaction logs, assessing the available indicators, policies,

and assumptions on which those indicators are based, and methods to normalize variables such as names, email addresses, and institutions. We consulted with archivists, researchers, technical staff, and managers at DANS, some of whom consulted with contributors or consumers of EASY datasets for further clarification. These insights were used to conduct iterative analyses of the database of registered users and transaction logs and to design the initial set of interview studies.

A series of questions emerged that could be addressed with user traces. The first set of questions provides a basic description of the user population. The user activity database contains only last-login dates; it does not create a cumulative record of an individual's interaction with the EASY system. While about 2000 new accounts were created in each of the three years studied, about half of these appear to be one-time visitors. About 4000 user accounts appear to be active. It has proven difficult to determine the number of unique users and the number of registrants. Demographic information in registration profiles is often scant, but is useful to normalize user names, particularly to identify duplicate accounts.

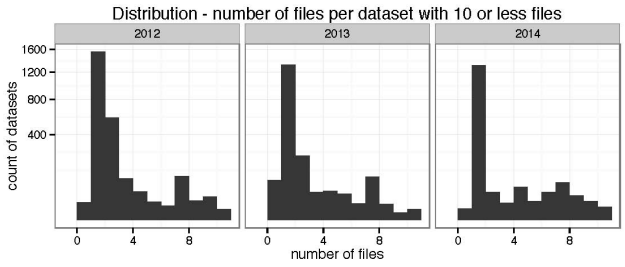


Figure 1. Number of uploaded datasets with x files for three consecutive years

The second set of questions addresses the activities of DANS/EASY users, such as the distribution of visits to the system, overlap between contributors and consumers of datasets, specifics of user queries, session time and length, and scenarios of usage patterns. As shown in Figure 1, about 2500 datasets are contributed each year. The size of datasets is highly skewed, ranging from 0 to more than 50,000 files in a single dataset. The mode for files/dataset is 1 (4209 datasets with 1 file each). Some of these are batch uploads. The number of datasets contributed per creator also is highly skewed, ranging from 1 to 547. Of the 2040 distinct creators in the three year period, about half of them (1040) contributed only one dataset.

Download statistics also are skewed, as shown in Table 1. These statistics are difficult to interpret, as individual files within datasets can be downloaded. Users typically download one dataset or one file per visit, but one user clearly wrote a script to download approximately 14,000 datasets. Most of the datasets in EASY belong to an archaeological sub-collection, whereas most of the download requests concern datasets from the social sciences and history. Among the most popular datasets are WoON2012 (Annual residential living conditions research) (549 downloads), Geological-Geomorphological map of the Rhine-Meuse delta, The Netherlands (404 downloads) and Nationaal Kiezersonderzoek, 2010 (National voter survey) (391 downloads). Users come from various regions, mostly from locations with universities in The Netherlands. Some are from the U.S. and other parts of the world.

Year	Multiple Files	Single File
2012	7,890	40,378
2013	7,907	43,005
2014	6,958	73,659

Table 1. Number of download requests per year.

DISCUSSION AND CONCLUSION

The transaction log files are a rich source of descriptive information about who uses the system, when, how, and how often. As common with most kinds of "big data" analyses, these transaction logs are deficient in a number of ways. However, log files are the best data sources that most archives currently have, and are no worse than many other archival and historical sources. We are assessing these unobtrusive data cautiously, to exploit available sources and to make reasonable inferences. Our findings suggest ways to instrument systems that will yield more comprehensive descriptions of user behavior.

Logs intended to monitor web traffic proved difficult to match to archival functions based on the OAIS reference model (Consultative Committee for Space Data Systems, 2012) or to specific user behaviors. Close inspection reveals the difficulty of explicating basic units of analysis such as dataset, file, user, creator, and consumer that are essential for studying user activity. The means by which datasets are contributed to DANS also proved difficult to distinguish in aggregate web statistics. Datasets can be contributed to DANS/EASY individually through the web interface, but at least two other protocols are available for batch ingest by partner institutions. Similarly, contributors may have their own criteria for when to combine multiple files into one dataset and when to treat each file as an independent dataset. Units can be arbitrary, and anomalies arise such as cases where a dataset contains zero files.

Such flexibility in combining data files makes statistics on downloads of datasets difficult to interpret. Consumers may download whole datasets or selected files within datasets. Contributors to the system may be the researchers themselves, or may be staff members uploading datasets on behalf of the team or organization. Free text metadata fields allow contributors to add other responsible persons to the record, such as creator, owner, or rights holder. A substantial portion of the datasets in DANS/EASY are

restricted access. For some classes of datasets, such as parts of the archaeology collections, users must be authenticated for professional credentials. For other classes of restricted datasets, the contributor determines whether to provide access on a case-by-case basis. For yet other classes of datasets, DANS archivists authenticate requestors based on established legal criteria.

Identifying individual users also can be problematic. Conventions such as user registrations are helpful indicators of activity, but cannot be assumed to provide an accurate record of unique visitors, given the ability for individuals to have multiple user accounts and email addresses. People interact with the system in different roles, whether as creator, contributor, consumer, archivist, researcher, policy maker, or student. As individuals change institutions and roles over the course of a career, their credentials may change accordingly. Users are requested to provide their Dutch Author Identifier (DAI, which is similar to an ORCID) if they have one. However, so few users comply that the DAI is of limited value in identifying unique users of DANS.

Research on the uses and users of digital data archives requires a mix of quantitative and qualitative methods. Statistical analyses of log files provide insights into who uses the system, when, how, and how often. However, technical logs may not be able to distinguish between humans and scripts, especially if scripts are well designed to mimic human activity. Interpreting statistics and traces of activity requires an intimate knowledge of archive policy, legal conditions at the national and local level, and variations in practice and policy by discipline. Mining the traces of activity in valid and reliable ways for research purposes requires different parameters, policies, and practices than transaction logs that are designed for system monitoring and maintenance. While the logs are less useful than expected for drawing samples of users to interview, these insights suggest parameters needed for the next generation of user transaction logs in data archiving. We will combine available log data with expert judgment to select contributors and consumers to interview. The interviews, in turn, will help us to interpret the transaction logs. In the longer term, we will make recommendations for the design of transaction logging systems that will provide useful research data, and will generalize our research methods to offer guidance for user behavior studies of other digital data archives.

ACKNOWLEDGMENTS

The participation of Christine L. Borgman, Herbert Van de Sompel, and Andrew Treloar in this research is supported by KNAW, the Netherlands Academy of Arts and Sciences, which is gratefully acknowledged. Peter Doorn, Director of DANS, has graciously opened the DANS doors, physically and digitally, for these investigations. Additional support for conducting and analyzing interviews is provided by a gift from Microsoft Research and grants from the Alfred P. Sloan Foundation to UCLA, Christine L. Borgman, Principal Investigator. The EC funded FP7 project Impact-EV has provided support for data mining, analytics and methods. Sally Wyatt and Ashley Sands provided comments on earlier drafts. Milena Golshan of UCLA provided technical and bibliographic support.

REFERENCES

Akdag Salah, A. A., Scharnhorst, A., Ten Bosch, O., Doorn, P., Manovich, L., Salah, A. A., & Chow, J. (2012). Significance of visual interfaces in institutional and user-generated databases with category structures. In *Proceedings of the Second International ACM Workshop on Personalized Access to Cultural Heritage - PATCH '12* (p. 7). New York: ACM Press. doi:10.1145/2390867.2390870

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press. http://mitpress.mit.edu/big-data

Borgman, C. L., Hirsh, S. G., & Hiller, J. (1996). Rethinking online monitoring methods for information retrieval systems: From search product to search process. *Journal of the American Society for Information Science*, 47(7), 568–583. http://doi.org/10.1002/(SICI)1097-4571(199607)47:7<568::AID-ASI8>3.0.CO;2-S

Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System* (No. Issue 2). Consultative Committee for Space Data Systems. http://public.ccsds.org/publications/RefModel.aspx

Committee on Future Career Opportunities and Educational Requirements for Digital Curation. (2015). Preparing the Workforce for Digital Curation. Washington, D.C.: National Academies Press. http://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation

Force11. (2014). About Force11. https://www.force11.org/about

National Science Board (U.S.). (2005). Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Arlington, Virginia: National Science Foundation. http://www.nsf.gov/pubs/2005/nsb0540/

OpenDOAR. (2014). OpenDOAR: The Directory of Open Access Repositories. http://www.opendoar.org/

re3data (2015). Registry of Research Data Repositories. http://www.re3data.org

Research Data Alliance. (2015). About RDA. https://rd-alliance.org/about.html

Scharnhorst, A., Ten Bosch, O., & Doorn, P. (2012). Looking at a digital research data archive - Visual interfaces to EASY. *Preprint. Computer Science. Digital Libraries. arXiv:1204.3200[cs.DL]*. Digital Libraries; Physics and Society. http://arxiv.org/abs/1204.320