Digital Data Archives as Knowledge Infrastructures

Borgman, Christine L.; Scharnhorst, Andrea; Golshan, Milena S.

2018

document version
Peer reviewed version

Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse

Submitted to the *Journal of the Association for Information Science and Technology*

Christine L. Borgman, UCLA Center for Knowledge Infrastructures, Los Angeles, CA
Christine.Borgman@ucla.edu, https://knowledgeinfrastructures.gseis.ucla.edu/
Andrea Scharnhorst, Digital Archiving and Networked Services, The Hague, Netherlands,
andrea.scharnhorst@dans.knaw.nl, https://dans.knaw.nl/en
Milena S. Golshan, UCLA Center for Knowledge Infrastructures, Los Angeles, CA
milenagolshan@ucla.edu, https://knowledgeinfrastructures.gseis.ucla.edu/

Abstract

Digital data archives play essential roles in knowledge infrastructures by mediating access to data within and between communities. This three-year qualitative study of DANS, a digital data archive containing more than 50 years of heterogeneous data types, provides new insights to the uses, users, and roles of these systems and services. Consumers are highly diverse, including researchers, students, practitioners in museums and companies, and hobbyists. Contributors are not necessarily consumers of data from the archive, and few users cite data in DANS, even their own data. Academic contributors prefer to maintain control over data after deposit so that they can have personal exchanges with those seeking their data. Staff archivists provide essential mediating roles in identifying, acquiring, curating, and disseminating data. Archivists take the perspective of potential consumers in curating data to be findable and usable. Staff balance competing goals, and competing stakeholders, in time spent acquiring and curating data, in maintaining current data and long-term stewardship, and in providing direct access and interfaces to search engines and harvesters. Data archives are fragile in the long run, due to the competing stakeholders, multiple funding sources, and array of interacting technologies and infrastructures on which they depend.

Introduction

Providing open access to data is a profound shift in scholarly practice, requiring researchers to treat data as products that can be extracted from the research process. Such extraction requires extensive labor, expertise, and expense beyond the conduct of the research per se. Incentives and motivations to share and reuse data are complex and vary considerably by domain, funding source, type of data, and other factors. Even in cases where researchers are motivated to release data, the infrastructure to ingest, curate, and sustain access to those data may be lacking. Similarly, the availability of data for reuse depends on infrastructure to make those data discoverable, retrievable, interpretable, and usable.

Digital data archives facilitate knowledge exchange between those who collect data, release data, and those who use data created by others. Little is known about the stakeholders and mechanisms

involved. Who are the individuals who contribute their data to digital archives? What are their expectations for stewardship, dissemination, and control of those data? Who are the individuals that search digital data archives, and what do they do with the data they acquire? What are their expectations for access, integrity, and stewardship of those data? How homogeneous are the communities of contributors to a digital data archive? How homogeneous are the communities of consumers? What is the degree of overlap between contributors and consumers? What role do the staff of digital data archives play in acquiring, curating, disseminating, and stewarding data? What roles do archives staff play in identifying, supporting, and bridging these communities?

Incentives to share data have received far more research attention than has the underlying knowledge infrastructure necessary to facilitate data sharing and reuse. Knowledge infrastructures are "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (Edwards, 2010, p. 17). They are living systems influenced by complex sociotechnical factors (Borgman, Darch, et al., 2015; Edwards et al., 2013; Karasti & Blomberg, 2017). While research data can be exchanged publicly or privately, with varying degrees of mediation, data archives are the mechanism preferred by most journals and funding agencies (Borgman, 2015; Wallis, Rolando, & Borgman, 2013).

Digital data archives take many forms and have many homes. Some collect only data of certain types and formats, such as genome sequences for biological research or survey data for the social and economic sciences. Others are more generic, collecting textual documents, static and moving images, audio, and other data types. Data archives range widely in mission from providing immediate access to replication datasets to long-term preservation. Accordingly, they vary in the degree of investment in data curation. Some institutions devote days or weeks of professional labor to curating each dataset before deposit; others rely on "self-curation," accepting data in whatever form submitted, with minimal review. The longevity of collections also varies from short-term grant funding to long-term commitments by universities, governments, or other agencies (Borgman, 2015; National Science Board (U.S.), 2005). Business models may be based on memberships, grant funding, institutional support, contributions, corporate for-profit entities, or a combination (Shankar, Eschenfelder, & Downey, 2016).

Despite the advances in research on data sharing and reuse, and advances in standards and practices through organizations such as the Research Data Alliance and Force11, relatively little research has addressed the mediating roles of digital data archives (Borgman, Darch, et al., 2015; Borgman, Darch, Sands, Wallis, & Traweek, 2014; Force11, 2018; Mayernik, Wallis, & Borgman, 2013; Mayernik, Wallis, Pepe, & Borgman, 2008; Pasquetto, Randles, & Borgman, 2017; Pasquetto, Sands, & Borgman, 2015; Research Data Alliance (RDA), 2018; Wallis et al., 2013). The three-year case study reported here addresses that gap to examine the roles of digital data archives in research data practice, and how they function as knowledge infrastructures for their communities.

Background and Literature Review

We briefly review knowledge infrastructures, motivations for sharing and reusing research data, uses and users of digital data archives, and context for our case study.

Knowledge Infrastructures

Infrastructures are difficult to study because they are complex, long-term, social and technical entities that are largely invisible (Edwards et al., 2013; Karasti & Blomberg, 2017; Star, Bowker, & Neumann, 2003; Star & Ruhleder, 1996). All infrastructures are fragile in the long term, although components such as digital archives may last for decades (Borgman, Darch, Sands, & Golshan, 2016; Edmunds, L'Hours, Rickards, Trilsbeek, & Vardigan, 2016; Lee, Dourish, & Mark, 2006). Some infrastructure components may be field-specific, such as ontologies or genome databases, while others may support multiple domains, such as archives of government statistics or general social surveys. Yet others may be essential resources for multidisciplinary fields, such as the International Ocean Drilling/Discovery Program that provides cores for biological and physical science research (Darch et al., 2015; Darch & Borgman, 2016).

Sharing and Reusing Research Data

Research data take many forms and may originate from observations, experiments, excavations, physical specimens, or other methods. Determining what are data is itself problematic, as one person's signal is often another's noise. Here we draw upon Borgman's definition that *data* refers to "entities used as evidence of phenomena for the purposes of research or scholarship" (2015, p. 29). Thus, almost anything can be treated as data in a research project.

Until recently, phenomenological concerns about the nature of data were left to philosophers and historians of science (Blair, 2010; Rosenberg, 2013). When data began to be treated as products of research to be released, shared, and reused by others, discussions about data and data practices accelerated. Stores of physical specimens for research in agriculture, botany, and earth sciences may date back several centuries, whereas digital data archives date to the mid-twentieth century. By the late 20th century, research policymakers began to promote or require release of research data on the grounds that "the value of data lies in their use" (Committee on Issues in the Transborder Flow of Scientific Data, National Research Council, 1997). Data management plans are now a standard part of research grant proposals. Depositing datasets at the time of submitting articles to journals also has become common, and often mandatory, research practice in many fields (Borgman, 2015).

Releasing, sharing, and reusing data are complex practices, wherein disincentives often outweigh the incentives for compliance. Incentives among stakeholders also are misaligned, as costs and benefits may be distributed in ways that discourage data sharing. While many researchers recognize the importance of preserving data, others question the long-term value of their data to themselves or to others, or ask whether potential reusers can understand someone else's data (Frank, Yakel, & Faniel, 2015; Mayernik, 2011, 2016; Tsoukala et al., 2015; Wallis et al., 2013).

Incentives for releasing, sharing, and reusing data are poorly understood and vary considerably by field (Borgman, 2015; Piwowar, 2011; Tenopir, Palmer, Metzer, van der Hoeven, & Malone, 2011; Wallis et al., 2013; Weber, Baker, Thomer, Chao, & Palmer, 2012; Zimmerman, 2008). Releasing data, whether by sharing directly with other persons or by depositing in a data archive, requires careful selection of data and the addition of metadata and other contextual information necessary for interpretation. Software or algorithms associated with data production may be needed to interpret or reuse datasets. The labor involved in documenting data for sharing is often extensive and unrewarded, and requires skills beyond the expertise of most researchers (Borgman, 2015; Mayernik, 2016; Mayernik et al., 2013; Pasquetto et al., 2017, 2015; Wallis et al., 2013). Researchers often maintain a sense of ownership over their data, regardless of legal status. Other reasons to control access to data include protecting privacy, cultural sites, endangered species, and intellectual property rights (Borgman, 2018; Eschenfelder & Johnson, 2014).

Uses and Users of Digital Data Archives

Despite the long history of studying information-seeking behavior in libraries and other institutional contexts (Case, 2006), uses and users of digital resources have proven hard to study. One reason is the multiple roles of individuals. Researchers may contribute data to archives, may be consumers of data in those archives, or both (Palmer, 2005).

Data sharing and reuse are context-specific, thus archivists wish to understand their communities sufficiently to maintain complex relationships of services and trust (Faniel, Barrera-Gomez, Kriesberg, & Yakel, 2013). While professional practice dictates that services, infrastructure, standards, policies, and practices of digital data archives be based on their designated community of users (Consultative Committee for Space Data Systems, 2012; Yakel, Faniel, Kriesberg, & Yoon, 2013), social science research reveals that communities take many forms and are difficult to characterize (Lave & Wenger, 1991; Wenger, 1998).

More studies have addressed why people search digital data archives than why individuals contribute data. As many data archives are field-specific, research tends to focus on searching behavior within specific domains such as archaeology, social sciences, or engineering. Not surprisingly, user behavior tends to correlate with existing data practices in a field, and archives tend to be tailored accordingly (Arbuckle et al., 2014; Faniel, Kansa, Kansa, Barrera-Gomez, & Yakel, 2013; Faniel & Yakel, 2017; Kansa, 2012; Kansa & Kansa, 2011, 2013; Kansa, Kansa, & Arbuckle, 2014). Archives often assist their communities by setting standards for data practice and aiding individual researchers in preparing data for contribution. Archaeology, for example, is promoting common standards to improve management of the wide variety of data types and formats in current use (Faniel & Yakel, 2017; Kansa & Kansa, 2011).

Assessing the match between content, services, and communities is difficult enough when the users self-identify with a domain such as archaeology. Much harder to study are the archives that serve broad communities with diverse types of data, such as those that span the social sciences and humanities, or institutional repositories that serve all schools and departments of one or more universities. The more generic the data collection, the more diverse the community of users.

Thus, the set of users associated with multidisciplinary data archives are particularly hard to identify or to characterize.

DANS: Data Archiving and Networked Services

As an important exemplar of digital data archives, this study focuses on the Data Archiving and Networked Services of the Netherlands. DANS was founded in 2005 as an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and of the Netherlands Organization for Scientific Research (NWO) (DANS, 2017). DANS has cumulative responsibility for more than 50 years of digital research data in the social sciences and humanities from its predecessor organizations such as the Steinmetz Archive ("Steinmetz Archive," 1989). Currently, DANS provides three services: NARCIS, the Dutch Research Information System, Dataverse for sharing living data; and EASY, the Electronic Self-Archiving SYstem, for the long-term archiving of datasets. This paper focuses on the latter service. EASY is a legal deposit archive for archaeology data, which constitute the majority (about 75%) of datasets in the collection. Other domains in EASY include behavioral and educational sciences; law and public administration; life sciences, medicine, and healthcare; economics and business administration; and science and technology. 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 ("EASY: Published datasets," 2016).

DANS has conducted its own studies of users and usage, providing background information for this qualitative study. Among the factors that influence Dutch researchers' inclinations to share and reuse data are standardization needs, data appraisal and data backlog issues, a sense of ownership over the data, fear of misunderstanding and misinterpretation of their data, and policy requirements (DANS, 2010; Dillo & Doorn, 2011). Quantitative analyses of uploads and downloads of datasets show steady increases over time. More than 85% of datasets in DANS have been downloaded at least once (Doorn, 2017). The most downloaded individual datasets are from the social sciences, such as census data from Statistics Netherlands (Akdag Salah et al., 2012; Scharnhorst, Bosch, & Doorn, 2012). However, archaeology represents the largest category of overall downloads "in absolute terms" (Doorn, 2017).

Research Questions

The three-year case study reported here was developed as part of a larger research agenda to understand how practices vary between research domains and how these factors influence data sharing and reuse ("UCLA Center for Knowledge Infrastructures: Home," 2018). We focus here on the mediating role of digital data archives between contributors and consumers who share and reuse data, characteristics of these communities, and their expectations for digital data archiving services. The term "contributor" refers to the person who collected data that were deposited in DANS, regardless of whether the actual uploads of data were done by the contributor or by another person, such as a staff member, librarian, or a graduate student. "Consumer" refers to the person who retrieved or acquired the data from DANS, regardless of whether or how the data were subsequently reused.

Three research questions guide our study, using DANS as an exemplar of digital data archives:

1. Who contributes data to the digital archive? How, when, why, and to what effects?
2. Who consumes data from the digital archive? How, when, why, and to what effects?
3. What roles do archivists play in acquiring, curating, and disseminating data?

Research Methods

These research questions were developed in a series of visits to DANS over a three-year period when the first author was a DANS Visiting Scholar. Documents about DANS, many of which were in English, and publications by DANS staff and by those interviewed were acquired and analyzed. In cases where important documents were written only in Dutch, our DANS colleagues explained the main ideas of interest. This study followed the same general methodological framework as our other recent studies of data practices (Borgman, Darch, et al., 2015; Borgman, Golshan, et al., 2016; Darch & Borgman, 2016; Sands, 2017).

An early step in our research was to characterize the communities of contributors and consumers of DANS data, and the degree of overlap, by mining DANS transaction logs of system usage. DANS, like most digital service organizations, maintains weblogs for purposes of auditing, trouble shooting, and managing information. Users must register with DANS to contribute data or to retrieve certain kinds of datasets, thus creating a user database with a small amount of demographic information (e.g., name, institution, email address, discipline). Although transaction log data have a long history in information retrieval for studying user behavior, they can be difficult to interpret. Logs provide traces of what people do, but lack information about why they do so (Borgman, Hirsh, & Hiller, 1996).

Working closely with DANS data scientists, we selected transaction logs and the associated database of registered users from three fiscal years, October 2011 through September 2014 (FY 2012-2014). This was a period of consistent record keeping since the last major system upgrade. The logs contained sufficient information to identify contributors and consumers, but not frequency of system use (Borgman, Van de Sompel, Scharnhorst, van den Berg, & Treloar, 2015).

Interviews, averaging about one hour in length, were conducted in English, which limited the response rate. With a goal of 10 interviews from each group, we drew an initial sample of 50 contributors (from 3517 submissions during the sampling time frame) and 50 consumers (of 3401 registered during the sampling frame). After initial low levels of responses, more candidates were randomly selected from the existing pools. In total, we contacted 75 contributors, 9 of whom agreed to participate, and 112 consumers, 8 of whom agreed to participate. We also interviewed 10 members from the DANS team, which included all of the curators and archivists on staff at the time. While English is widely spoken in the Netherlands, especially in the academic community, it is likely that lack of conversational fluency reduced the response rate for consumers of DANS data, who spanned a much broader array of professional and academic backgrounds. Most DANS data contributors were academic faculty or researchers.

Most (21 of 27) interviews were conducted in person in the Netherlands, preferably in the offices of the interviewee. These interviews were conducted by one or two of the UCLA staff members. In most of the interviews with contributors and consumers, one DANS staff member participated in the interview, providing context and translation (usually of technical terms) as needed. The remaining five interviews of contributors and consumers were conducted remotely by videoconference by a UCLA team member. All interviews with DANS staff members were conducted at DANS by a UCLA team member, with minimal participation of DANS staff to maintain confidentiality. While participating DANS interviewers took UCLA training in human subjects' research and were certified to participate in the study, all data coding was conducted by UCLA staff and anonymized before sharing with any DANS staff. Interview subjects are anonymized, labeled by category from which identified in the weblogs (contributor, consumer) or DANS staff, and given a numerical identifier within the category (e.g., Contributor1)

Findings

The findings are framed first with a summary of the sample population and discussion of the content of the digital archives under study. Interview findings are organized by the three research questions, which seek to characterize the three sets of stakeholders: Contributors of data, consumers of data, and the staff of DANS who are responsible for curating and managing EASY, the digital data archive.

Sample Demographics

We conducted a total of 27 interviews with 28 people (one interview was conducted with two people from the same organization); 21 men and 7 women. The distribution of research subjects by category, domain, and occupation is presented in Table 1.

Table 1: Distribution of Interview Sample

Stakeholders /Participants	Number of Interviews	Domain Expertise	Occupation
Data contributors	9	Archaeology, history, and related fields (3) Labor economics (1) Linguistics (2) Oral histories (1) Scholarly communication (1) Plant biology (1)	Academic staff (7) Scholarly-professional society (1) Private company employees (2 persons; 1 interview)
Data consumers	8	Archaeology, history (4) Political science (2) Social science (2)	Academic staff (3) Cultural institution staff (2) Citizen scientists (1) Students (2)
DANS staff	10	Archaeology and humanities (7) IT development (3)	Archivists, project managers, and IT developers (10)

Data in DANS/EASY

DANS acquires data for EASY (Electronic Self-Archiving SYstem) from the humanities, social sciences, behavioral and educational sciences, law and public administration, life sciences, medicine and health, and economics and business administration, thus serving multiple communities. Data types include text, tables, images, graphs, maps, and audio-visual files such as oral history materials. The most common formats are PDF, DOC, and TXT for text; CSV and XLS for tabular data; TIF for images; and MP4 for audiovisual recordings. Archaeology reports, which constitute the largest portion of datasets in DANS/EASY, usually contain text, tables, photographs, graphs, and maps in a single PDF file (Mientjes, 2015). As of May 2015, when the interview sample was drawn, EASY contained 29,743 published datasets. Datasets average about 100 files each (Doorn, 2017); these files may have different level of access, as shown in Figure 1. Contributors have considerable flexibility in structuring their files and datasets. While this flexibility provides context-specific metadata and organization, it also limits the consistency of data structures within EASY.

Figure 1: Granularity of Access

Source: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:66658/tab/2

Perspectives on the types, formats, and content of data in DANS differ considerably by stakeholder, as reported in the sections below. Choices of data may vary by the stage of a research project, by format, type of representation, relationship to physical samples, degree of processing, level of access, and other factors.

Contributors of Data to DANS

The ethnographic fieldwork, weblogs, document analysis, and interviews with DANS staff revealed that "contributors" encompasses multiple roles in institutions and in the life cycle of research projects. In most cases, individual researchers contribute data that they collected. This is the simplest situation and the one where interviews provided the fullest explanation of the motivations for contributing, along with the specific practices, processes, and constraints. In other cases, individuals identified as contributors in the log files were intermediators between researchers and the DANS/EASY system. DANS' management distinguishes between "front office" services they provide directly to contributors and consumers, and "front office" services provided by other intermediators such as librarians. DANS refers to their own work in ingest and curation as "back office" functions. DANS' users tend to refer to DANS generically, not distinguishing between EASY and other services offered.

Weblogs listed only names and email addresses associated with uploading and downloading files, which was insufficient information to determine whether an individual was the data collector or an intermediary. To reach persons who were associated with data creation and reuse, we requested contacts with researcher-contributors in our interview solicitation. All of the ten individual contributors interviewed, in nine interview sessions, were involved directly in collecting data. Several of them identified other individuals in their organizations who sometimes upload data files to the DANS/EASY system on their behalf.

As documented in Table 1, seven of the nine contributor interviews were conducted with academic staff who were professors or researchers, whether from universities or scientific institutes. One of these seven had contributed data after retiring from the university. Two individuals from a private archaeology research company, a field researcher and his supervisor, were interviewed together in one session. Another individual led a scholarly-professional society. In cases where a person was identified by one category from the weblogs but found to have dual roles when interviewed, the person is tagged by his or her weblog assignment (e.g., Consumer4 is also a contributor of data). In this section, we explore the motivations and conditions for contributing data to DANS and other archives.

Types and Origins of Data

While the interview sample was too small to achieve proportional representation by research domain, the nine interviewees spanned a wide variety of research areas, yielding a broad characterization of this diverse community. Most contributors discussed one or a few datasets they had input to DANS/EASY. The exception was the professional archaeology company, who discussed practices involved in contributing about 250 datasets per year. They are one of about 20 such companies in the Netherlands.

Most contributors were precise in describing their data, albeit by categories of concern to their own research. Consumer4, a university faculty archaeologist who also contributed data to DANS, distinguished between three types of data based on stages and types of research:

> "The first is excavation data, so field maps, digitized field maps and that sort of information. The second group of information is on finds, lists of finds made during excavations … the third category is reports on individual excavations… What type of data we are looking for depends on the type of project that we actually are carrying out."

Consumer4 further distinguished between digitized data that were born in analog form and data that were born digital:

> "Until 20 years ago, excavations were drawn at scale on large sheets of paper, and these sheets were stored in an archive…. Nowadays, more and more of these paper drawings have been scanned and are available as bitmaps, rasterized data… (In recent years), … (with) 'robotic total stations,' they make digital measurements in the field, and you have immediately your maps in a digital form."

Contributor5, of the private archaeology company, also mentioned the trend toward creating data in digital form. This company claims to have invested more heavily in digital data production than some others in the field. They are testing new data format standards that will automate parts of the ingest process to DANS/EASY and other repositories.

Contributor4, another faculty member, distinguished between types of data collection that acquire physical samples and the ways in which representations are created, processed, and coded. Boreholes result from drilling in a field, and cores are the samples that are taken from those boreholes. Each is described separately, as multiple cores may be obtained from an individual borehole. Some of these data are acquired from companies, as drilling boreholes and

collecting core samples are standard practice in assessing building sites. Contributor4 and his colleagues map these holes and cores, "which is high level interpreted data or information." His information products, created with geographic information systems (GIS), result in his own map system, which he shares "only with the direct workers." While his "basic map is just lines and polygons with labels," his coding system enables the team to run scripts that extract user product maps. "Then it becomes nicely colored maps…" The files contributed to DANS/EASY are large in number, consisting of map layers and parameters that can be used to reconstruct his maps or make new ones.

Some large oral history projects described by Contributor7 were deposited in DANS/EASY for stewardship rather than for access. These datasets contain audio recordings of interviews, transcripts, summaries, and extensive metadata. Each dataset encompasses all interviews on a topic or from a project. At least one dataset is stored in DANS as a dark archive, not accessible to anyone but the owners, due to political sensitivity. The open materials can be searched and streamed via a project website, delivered in the background by DANS. The audio materials also are being used as a testbed for research in computational linguistics.

Motivations to Deposit Data

The most common reason given for contributing data to DANS/EASY was to meet legal requirements. These requirements are explicit for archaeological data, especially those gathered by private companies. The *European Convention on the Protection of the Archaeological Heritage*, known as the Valletta Treaty, requires archaeological investigations prior to breaking new ground for any construction (Council of Europe, 1992). Netherlands law associated with the treaty requires that reports of these archaeological investigations be made public by deposit in DANS/EASY; some reports also are deposited in local archives and in the national library.

The archaeology company we visited deposits about five reports per week, or about 250 per year. Most of these are initial site surveys. The survey reports we examined average about 90 pages in length; these are PDF documents that contain many tables, maps, and charts, largely in color, with extensive narrative. The entire report is contributed as a single PDF file; internal tables and charts are not structured files. In about 10% of investigations, they find sufficient evidence of human activity such as pot sherds, weapons, or traces of old buildings that a more extensive study is legally required. In about 10% of the latter cases, a full archaeological dig is required, which may postpone building construction for a considerable time period. With a few exceptions, all of these reports are contributed to DANS/EASY.

Academic researchers also deposit data to fulfill requirements by their funding agencies. Contributor6 was explicit that "the reason we deposit the data there originally has to do with the contractual agreement we have with the people who pay for the data gathering… Because some of the data we gather we do not deposit at DANS…"

However, the majority of the academic researchers with whom we spoke deposit data in DANS/EASY as a means to share them with other researchers within or outside their academic community. As Contributor4 said, "another reason to put it in DANS was I don't want to sit on data until all the publications are ready." Similarly, from Contributor8, "I think it's very

important that we share all our data, … the data can be used so many times for so many different questions." Others, such as Contributor3, deposit data for preservation purposes: "I volunteered …when I retired, to make a database of it… So I hope it will never be lost."

Community Characteristics

We hypothesized that a considerable overlap existed in the contributors and consumers of data in DANS/EASY, which would suggest a coherent community of users associated with the archive. As the log data did not provide sufficient granularity to test this hypothesis, we sought answers in our interviews. Some, but by no means all, of the contributors of data were also consumers of data. Of those interviewed, the two individuals from the private archaeology company were the heaviest contributors and heaviest consumers of DANS/EASY. They began each new project by searching DANS for any prior archaeological studies of the region of interest. Their field reports cite prior studies, whether conducted by these or other companies, or academic publications.

Most of the academic contributors with whom we spoke collect new data for their own studies, and later may contribute those data to DANS or other archives. They might find background information in DANS, but did not necessarily incorporate others' data in their own work. Only two of the academic contributors said that they searched DANS/EASY for data to incorporate in their own research. Some researchers mentioned depositing data in DANS as a means to disseminate data to their colleagues and students, sending links on a regular or ad hoc basis.

Archaeological studies conducted by academic researchers overlap with private company studies in several respects. Academic researchers report their findings in journal articles or conference papers, with data deposit to follow. While such reports are less standardized than those of private investigations, the content is of mutual interest. Grant-funded research tends to be conducted on a longer time scale than private investigations; the latter are conducted and filed quickly to fulfill contract requirements and keep construction projects on schedule.

Individual personnel also overlap between universities and private companies. Some universities in the Netherlands have for-profit units that compete with private companies to conduct archaeological site studies. University faculty also consult for these companies when specialized expertise is required, such as types of pottery, military history, or geomorphology. We encountered one case where archaeology company employees would alert their former professor to interesting new finds in his area of interest. Continuity between generations also occurs when faculty inherit data from their advisors, as we found in another case.

Credit, Control, and Attribution

Whether or not required to deposit data, contributors often were motivated by some combination of credit, control, or attribution. Contributors may express a commitment to open access, as did Contributor6: "First of all, we are also believers of the open source, and open data policy, so we try to be as open as we can with the data." Yet they also express concerns: "Sometimes you want to keep some data until you have your stuff published, because otherwise people will publish instead of you, which is bad for us." Contributor6 and his team have an internal policy for archiving the data for their own reuse. Other researchers deposit their data to DANS for

preservation and open access, as Contributor3 said, “I hope some people will use it in the future…[and] they will make a publication.”

Individuals can search DANS/EASY anonymously or they can sign up for a personal account. Archaeology datasets are the only subset of DANS for which registration is specifically required, largely due to concerns about protecting sites from looters. To join the DANS/EASY archaeology group, individuals must demonstrate some professional affiliation with the field. DANS/EASY allows contributors to “lock” their files, making them available only upon request. Access requests go directly to the contributors, who can then negotiate with the requestor. The majority of academic contributors we interviewed preferred to restrict access to users of DANS who have registered by name. They may ignore anonymous requests or ask the requestor to register.

A contributors’ willingness to provide access to locked data files depends on an array of considerations. Contributors appear more willing to grant access to data for which they have completed all expected publications. In other cases, the contributor may request a collaborative relationship as a condition of providing access. Contributor4, for example, was more willing to release data to Dutch researchers than to those from other countries. This was at least partly a matter of convenience as the text was in the Dutch language and others would need translation assistance. In contrast, an academic researcher who ran a citizen science project wanted the data to be available as quickly as possible as a means to keep participants motivated.

Release conditions tend to be project-specific. Consumer4, speaking in his role as a contributor, explained:

> Normally I ask, "What precisely are you looking for, and for what purpose?" Because generally they ask permission to use all the data. And there are maybe 200 datasets within this project, and I can happily give them permission to inspect all these 200 datasets, but if they ask me what they want to know, then I can tell them, "Take a look in these two datasets." Or "Forget it, the data which you need are not there."

Factors such as data sensitivity and institutional ownership also influence release conditions. Control over data is more complex when co-authors have different opinions on open access or different levels of engagement with the contribution process. Contributor3, the retired faculty member, explained how his university had a “front office” relationship with DANS, where the librarians contributed data on behalf of faculty and researchers. Despite his preferences for open access, another faculty member and the librarian had locked his datasets, following institutional practice. As a test, he tried to obtain access to his own data from DANS and was unable to do so. With some investigation, he determined that the request went from DANS to the library staff, who did not realize the library had a responsibility to respond. Next, the request went to a co-author on the original project, who did not respond because he was not involved in the data deposit and unaware of his delegated responsibility. Contributor3 would have preferred to grant open access than impose a request process, saying “in my opinion, … if people want to steal our data without reference to it, let it be. …”

DANS management now encourages the adoption of Creative Commons licenses. Several people we interviewed mentioned receiving the license-change request and their explicit refusal. These

individuals would only contribute data to DANS (or other repositories) if they could maintain some degree of access control.

While we did not ask all of our subjects about data citation, the topic came up several times. DANS follows archival best practices by assigning a digital object identifier (DOI) and suggesting a citation for each dataset. Despite the ready availability of these metadata, individuals were inconsistent about citing data they have obtained from DANS or elsewhere, and inconsistent even in citing their own data. Contributor6, for example, sometimes reused his own team's data in a publication without mentioning that those data were available in DANS/EASY. When individuals requested data from him, he sometimes referred them to DANS and sometimes gave them the data directly. Contributor4, in contrast, said that his papers always cite his datasets.

Consumers of Data from DANS

It was more difficult than anticipated to find consumers of DANS/EASY data – those identified from the weblogs as having downloaded one or more datasets – who were willing to be interviewed in English. As discussed in the methods section, we contacted more than 100 individuals who had downloaded DANS/EASY data during the sampling period to obtain eight interviews. Ideally, we would have sampled both frequent and infrequent users, but the logs captured only the most recent use, so we were sampling anyone who had searched and downloaded. Four of the eight consumers interviewed were in the archaeology domain, whether associated with institutes, museums, or local activities. The other four were in other areas of the social and political sciences, including students, professors, and visiting scholars.

Consumer interviews provided rich contrasts to what we learned from interviews with contributors and archivists and from the ethnography and document analysis that framed the study. Here we explore the demographics of these users and their motivations for searching, downloading, and using data from DANS.

Community Characteristics

The diversity of consumers interviewed shed light on the difficulties of locating willing subjects. While DANS datasets often have searchable English abstracts, the content is almost entirely in Dutch. The consumers interviewed used DANS intermittently, at most a few times per year. Several of them refreshed their familiarity with the system prior to meeting with us. We traveled to the far corners of the Netherlands to meet them, not only in their offices, but in their homes if they preferred.

While the domain expertise and professional roles of contributors and consumers are compatible, as shown in Table 1, their motivations and practices differ significantly. Just as the community of book authors is vastly smaller than the community of book readers, so is the community of data contributors much smaller than the community of data consumers.

Compared to DANS/EASY contributors, the consumers we interviewed were more diverse in demographics and motivations. Consumer3 contrasted most strongly with the contributors; he

was a computer professional whose avocation was guiding local history tours in his small town. He had become an expert on a particular type of regional archaeology, obtaining most of his material from DANS/EASY. He also searched DANS on behalf of his daughter, who was an undergraduate archaeology student. Over tea at his kitchen table, he showed us the database he had constructed and the blog site he maintained for local enthusiasts. A three-ring binder with maps, drawings, and other exhibits encased in plastic sheets was his guidebook for leading tours in the summer. In winter he did his research, and also was friendly with the town's official archaeologists and archivists. All of these materials are in Dutch, but he used his conversational English daily in his professional work in computing.

Consumer5, in another small town, was a curator at a regional museum. Our assumption, and his initial explanation, was that he was using DANS to obtain material for his museum exhibits. As the conversation progressed, a broader array of uses for DANS emerged. He also was using DANS/EASY as a digital library because the archaeological reports contained more extensive descriptive content and bibliographies than he could acquire from local libraries or from the collections of his museum. Outside of his professional time, often on evenings and weekends, he searched DANS/EASY for doctoral research; he plans to pursue his degree upon retirement from the museum. He was excited to show us a large collection of three-ring binders in which he had cataloged all known specimens of a certain type of bronze object found in the Netherlands. Each record had an image and extensive metadata of his own devising.

Information Seeking

Of those we interviewed, most were searching for data about a particular site or region in the Netherlands, or perhaps about a particular type of object, such as the museum curator mentioned above. The ability to search in the Dutch language for Dutch materials was a significant attraction, as Consumer8, an archaeologist, said, "There's really no other repository available in the Netherlands. There's only DANS… mainly of interest for Dutch resources, and it's the sites, and the raw data, it's only in Dutch..."

Searching by location is a difficult problem in information retrieval due to the many ways a place or site might be described. Examples include current or historical names of streets, towns, areas, buildings, or other entities, and descriptions of a historical event, an owner, or some other salient characteristic. The more languages involved, and the longer the time frame of the content, the greater array of naming possibilities. Terminology can be standardized to varying degrees by incorporating metadata from geography ontologies, e.g., (*Getty Thesaurus of Geographic Names*, 2017). However, the labor investment in curating records to this level of consistency is beyond the reach of most digital archives, including DANS/EASY.

Keyword searching in DANS for place names or locations is a daunting task due to the variety of data types and formats, the complexity of naming geographic locations, and the mix of Dutch language with English abstracts. Whether due to user interface challenges or simply the kinds of information sought, the consumers interviewed typically browsed content by category or geographic region. These users exhibited varying degrees of sophistication in geographic searching. Consumer8 referred to the "community name" ...where the excavation or the research took place." Consumer4, who also contributed archaeological data to DANS, used the more

technical term, to say that "datasets are usually stored under toponyms… if you cannot guess the toponym under which the dataset is stored, then you will not find it. … searching on the municipality, then you get enormous lists of datasets." Consumer3, the local guide, mentioned that he was willing to browse through 200 or so matches, which was four to five screen pages.

Consumer4 was one of several to identify the need for a "mapping facility" by which a searcher "can make geographical selections in the DANS archives." DANS recently had added a capability to search sites by drawing a region on an interactive map of the Netherlands. The new feature was not widely advertised, however, and not readily apparent in the interface. In several cases, we took a few minutes after the end of the interview to demonstrate this new feature, to the great interest of these interviewees.

Uses of Data

The uses of DANS/EASY data were as diverse as the set of consumers we interviewed. Consumer3, the local guide, is "mostly interested in maps combined with texts … so I can walk around and say to the people, 'This is... But this and this happened.'" Contributor5, of the commercial archaeology company, searches DANS by region to "make an inventory of the excavations and investigations in the neighborhood of our plan area." Consumer8 uses DANS archaeology reports to make comparisons in the field, "if it contains what I'm looking for, then I … check the field drawings and the photographs (to) see (if) what archaeologist wrote … is consistent with the way it looks like in the field, through my eye."

Consumer4, an academic researcher who also contributes to DANS, found "enormous profits" from DANS for "an assignment... that can only be done by combining very old data with very recent data." He frequently finds both the "very old data … And all the recent data" in DANS/EASY. His willingness to be interviewed was partly a payback for the benefits of access to DANS and its services.

Consumer4 also told us that he prefers "raw data, which I can connect in my own way, to very connected data, which are very difficult to entangle … rather 10 datasets with the various components of the data separate, than one big dataset, and I have to split all the data to get access to these parts… useful for my purposes." Thus, the lengthy packaged PDF reports are of great value to some consumers and of limited value to others.

DANS data also are used for course assignments, as Consumer1, a graduate student, reported about a statistics course. Of multiple assignments (about five per week) in the term, "about six or seven were with the databases of DANS... We had to download a file …a workbook ...with all kinds of hyperlinks to DANS' archives." Other students, or prospective students, may use DANS for research on dissertation topics, as Consumer5, the museum curator mentioned above. Academics who teach sometimes use DANS for sample datasets for their classes.

Consumer6, now a faculty member in her own country, had obtained data from DANS while working as a scholar in the Netherlands. She had not yet used those data, several years after acquiring them, but planned to use them in future research.

Credit, Control, and Attribution

As noted above, data contributors can maintain varying degrees of control over their data, and academic contributors typically released only to registered users. The contributors interviewed appear to assume that users interested in their data will register with DANS and make a request to use the data. As Contributor4 said, "they ask, 'Is your data open?' This is not open, open, open; it's open after you announce your name."

Contributors who lock their data also appear to assume a balanced relationship between contributors and consumers. Consumer4, who is also a contributor, is "prepared to invest some time in helping people who request his data and not just saying, 'Okay. There are 200 datasets and good luck with it.' Because it gives me a moral right to ask for some support when I need data as well, because the dataset may contain... thousands of files."

However, the consumers we interviewed, especially the non-academic ones, were seeking data that are easily accessible. As Consumer3, the local guide said, "when I can't access it, I don't know whether it's interesting. I could ask for access, but … I haven't tried, no…" While he is a registered archaeology user, he was finding enough material already and would rather limit his search to browsing unrestricted data.

Curators and Staff of DANS

We began the interview portion of the study with DANS staff, which helped to frame the larger questions and refine our specific interview questions for contributors and consumers. DANS staff included the ten individuals with direct responsibility for DANS/EASY as archivists, managers, software and systems development, and related roles. To maintain anonymity in this small sample, we report all as "DANS staff." We also met numerous times with DANS senior management, formally and informally, over the course of our three-year engagement for this study. We did not record interviews with them, preferring to focus on those staff most directly engaged with the contributors and consumers. In all cases, we emphasized that our goal was to study DANS as an exemplar of digital data archives, and not to evaluate the organization or the staff.

Support for Contributors and Consumers

DANS, which is funded by the Dutch government, is responsible for developing and maintaining the EASY collection within a generally specified scope. The staff process datasets that are offered to them, but they also seek data for the collection by reading journals in domains covered by DANS, by attending conferences and holding workshops, and by contacting prospective contributors directly. They bring a variety of expertise to their tasks, specializing in a domain area, metadata, legal contracts, or other aspects of data management. Archivists also assist DANS' users by staffing help desks and responding to queries by email or phone. They handle about 1,000 queries by email per year. By rotating help desk responsibility, staff are cross-trained in the many processes, practices, and problems that arise. Experience in working with users also contributes to software design and maintenance.

Contributors tend to require more hands-on assistance in creating metadata and working with the DANS interfaces to data than do consumers in searching. Given that consumers far outnumber contributors, the staff estimate that their time is about equally divided in service to these two groups. Some of the contributors, especially those most active, named their primary contact at DANS, whereas none of the consumers interviewed could identify DANS staff by name.

Acquiring and Curating Data

As with any data archive, staff make tradeoffs between time spent acquiring more data to grow the collection and time spent in curating data that are in the collection. Balancing tests differed among the 10 staff members. Several mentioned that DANS' management placed more emphasis on collection building, on the grounds that more data attracts more contributors and consumers, in a virtuous cycle. Several staff felt that more effort should be devoted to curation, on the grounds that datasets would be easier to find and use, thus increasing the value of their services. Staff also varied in their emphasis on who should be curating datasets. Some staff felt that contributors should provide most of the metadata, as they were the domain experts. Others preferred curation to be a staff activity, as they were the metadata experts and better at describing datasets in ways that others could find and use them. Curation activities mentioned include adding metadata, provenance, and documentation; migrating data to more standard formats; and other domain-specific improvements. Several mentioned the need to improve the searching capabilities of EASY.

DANS archivists review all files uploaded by contributors before publishing them to EASY. The amount of effort required to process data depends on the condition of datasets and the availability of relevant domain expertise. Some datasets arrive in standard formats that are easily ingested. Others require format migration, additional metadata, and various integrity checks. DANS staff take a comprehensive view that datasets are comprised of data, metadata, and persistent identifiers, whereas contributors tend to have a narrower perspective. DANSstaff5 mentioned that "it's sometimes hard to explain to people, to our consumers, or our contributors that I need to add something." DANSstaff2 provided a typology: "Dataset, in my perception, is the combination of adequate metadata in Dublin Core fields, possibly supplemented with specific metadata for a certain research specialization like language studies, we can accommodate that we place additional metadata with the dataset… The second element… is good documentation of the research project and its questions and its methodology and its problems perhaps. And the third element is of course an organized set of data files that are intelligible for somebody who finds them, that is organized in a good fashion, and that is in either accepted or preferred data formats."

Despite being the domain experts, contributors often prefer to deposit data, "as is," with little understanding of the kinds of metadata necessary to make their data more useful to others. Only the most frequent contributors of data expressed familiarity with metadata creation and the ingest processes. Contributor4 complemented the DANS archival staff for their curation work to migrate data from a proprietary format that requires expensive software to a simpler format that is more widely used in the community. Contributor6 mentioned his great relief when a DANS archivist returned his dataset after finding personally identifiable data that were inadvertently left in the file. He cleaned the data and returned them to DANS for ingest.

Access to Data and Stewardship

A related tension faced by DANS staff is the balance between current investments in dataset acquisition and curation with investments in long-term preservation and stewardship, which includes data migration and maintaining data integrity. Datasets have varying lifecycles of usage, and these patterns often are difficult to determine in advance. The weblogs, our interviews with DANS staff, and other internal DANS studies confirm that dataset retrieval is unevenly distributed across the archive.

Another tension mentioned by DANS staff is how to balance support for contributors who are motivated by making their data available to potential users as quickly as possible and for those motivated to deposit for long-term stewardship. New archaeological finds may have short and long-term value. Datasets such as oral histories, some of which are embargoed for years, are an example of datasets deposited as cultural records for long-term stewardship by DANS.

Policies and practices for contributor control of deposited data highlight these tradeoffs. DANS' operating principle is "Open if possible, protected where necessary" (DANS, 2017). While the staff would generally prefer to release datasets openly under Creative Commons licenses, they recognize that they can only acquire some important datasets by allowing contributors to maintain a degree of control over access. Locked files appear to get less usage due to the overhead of registering and requesting permissions, which then are negotiated between contributor and potential consumer. DANS' staff may learn of access problems only when consumers contact them for assistance. Legal contracts govern data deposit, and include mechanisms for control of locked data to default to DANS after some period of time without response by the contributor. Such policies are essential, else datasets may remain locked indefinitely if the contributors do not respond, cannot be found, have left the institution, or are deceased.

Knowledge Infrastructure Activities

DANS/EASY, as a digital data archive, does not stand alone. Among its many stakeholders are the government agencies that contribute census and statistics data, institutions that contract with DANS to contribute other kinds of data, universities and libraries who partner with DANS on "front end" services, agencies who require deposit with DANS, and the Dutch funding agencies that provide continuing support. DANS provides services such as Dataverse as part of Netherlands consortia, and assorted other services and databases associated with their many research projects.

DANS also operates NARCIS, which is a portal into Dutch Research information including a name authority file for Dutch researchers (Reijnhoudt, Stamper, Börner, Baars, & Scharnhorst, 2012). The data contribution forms in EASY include a metadata field for the NARCIS identifier associated with the datasets. In our study of the DANS log files, we found that contributors were inconsistent in providing these identifiers, thus they were not useful for analyzing community demographics. Where available, the NARCIS identifier can be linked to datasets in EASY.

Metadata in DANS databases are harvested by other services such as Europeana, OCLC, and Google Scholar. They have partnerships with the Netherlands Royal Library for some kinds of datasets. DANS provides some identifier resolution services associated with their data. DANS staff must balance the interests of their many stakeholders in the design and maintenance of systems and services, including EASY.

Most contributions to EASY, at the time of our interviews, were manual uploads of datasets. DANS is developing automated ingest services for partners who contribute large numbers of datasets. Others have requested data export services. Many European universities have implemented Current Research Information Systems to manage the publication and research records of their staff ("euroCRIS | Current Research Information Systems," 2017). Some universities already have approached DANS about establishing links from an internal CRIS to DANS for data registration. Such linking services will use SWORD and other protocols. Several application programming interfaces (API) to DANS databases and archives are under development. Some contributors to EASY provide searching access via their own websites, delivering datasets from EASY in the background. Interoperability between DANS systems and the growing number of external systems with which they must interface is a continuing challenge for DANS staff.

Discussion

Our findings, drawn from weblogs, document analyses, ethnography, and interviews, provide a rich array of information about the practices, policies, motivations, and concerns of stakeholders in DANS. Having summarized the findings by research question in the prior section, the discussion is organized thematically. Here we examine who uses DANS, how and why they do so, characteristics and intersections of stakeholders, how knowledge infrastructures mediate access to data, value propositions, and metrics.

Who Uses Digital Data Archives and Why?

Meeting people in the offices and homes where they use resources such as DANS provides context not possible in surveys. These individuals showed us files on their computers, stacks of materials, and representations of artifacts they had created through their use of DANS systems and services. A number of them walked us through their departments or workspaces, explaining the flows of material in and out of DANS and other information systems. The museum curator gave us a private tour of the current exhibitions, which helped us to understand how digital materials contributed to these physical exhibits. Visiting people in large cities and in small towns also let us see how much, or how little, they relied on colleagues and on digital services for access to information resources. Open-ended questions allowed us to pursue unanticipated lines of inquiry and to expand on initial responses. Conducting interviews in person is far more time-consuming and expensive than distributing an online survey, thus the rarity of such studies is not surprising. However, given the importance of digital data archives to knowledge infrastructures, scholarly communication, and science policy, such investments in exploratory research are essential (Borgman, 2015; Tenopir et al., 2015; Wallis et al., 2013).

The more clearly bounded the community, the easier it is to study. Digital data archives in areas such as astronomy, archaeology, and genomics were developed by and for defined communities. Few domains have the resources or the clear boundaries to maintain their own archives, however. Most researchers, and most consumers, are likely to rely on more generic digital data archives that are operated by universities, governments, or other entities. The diversity of their collections and their user bases make this class of archives extremely valuable and extremely hard to study. Our study of DANS yields insights into these collections and communities, and to the methodological challenges of studying them.

Motivations to Share and Reuse Data

The breadth of motivations to share and reuse data in a sample of less than 30 individuals was striking. While many people contribute datasets to DANS because they are required to do so, they differ in the degree to which they are concerned with access or preservation. Some wanted their data released quickly and widely so that others could use them. Many were focused on preservation, wanting their data to be available to their research community and their country for the long term. The latter was true for archaeology, oral history, linguistics, and plant biology, for example. In oral history, datasets might be dark indefinitely due to long embargo periods on sensitive material. Archaeology is a middle ground where cumulative records are essential for research and for practice. Archaeology data contributors, whether academic, corporate, or private consulting, sought comprehensive information about sites they study. Access and preservation seemed to be prized equally.

Defining Data

In DANS/EASY as elsewhere, notions of data are in the eye of the beholder (Borgman, 2015). Contributors made fine distinctions in relationships between data types such as excavation data, lists of finds, and reports on excavations. An archaeologist distinguished between boreholes and the multiple samples that might be drawn from each borehole. Several interviewees distinguished between digitized versions of analog records, such as excavation drawings, and born-digital data such as robotic measurements of field sites. Lines and polygons may be useful data in themselves; to become maps, these elements must be extracted with customized scripts and algorithms, often with proprietary software. In oral history, audio recordings of interviews, transcripts, summaries, and metadata each may be viewed as data. When consumers sought background information about a site, an excavation report in the form of a 90-page PDF was an adequate unit of data. In other cases, consumers preferred "raw data" in the most discrete units possible so that data could be aggregated in other ways.

DANS staff took comprehensive perspectives on data. One distinguished between three components of a dataset: (1) adequate metadata, both generic Dublin Core fields and domain-specific metadata; (2) adequate documentation of the research project and methods; and (3) an organized set of data files. Staff set a high bar, adding that the data files should be "intelligible for somebody who finds them, … organized in a good fashion, and … in either accepted or preferred data formats."

The variety of data definitions mentioned by our research subjects reflect the range of data types, levels of description, heterogeneity of content and metadata elements, and diversity of contributors and consumers associated with DANS/EASY. Plentiful content in Dutch, usually with English abstracts, makes DANS a rich resource, but it also limits the ability to provide standardized vocabularies, units of measurement, or common search features. Digital archives that serve heterogeneous data to diverse user communities must balance flexibility, standardization, and ease of use.

Uses and Reuses of Data

Among the fundamental reasons to share research data is to create collections that can be mined and combined (Arzberger et al., 2004; Borgman, 2015; Boulton et al., 2015; Committee on Issues in the Transborder Flow of Scientific Data, National Research Council, 1997; Holdren, 2013). When researchers acquire datasets and employ them to create new knowledge products, these are "foreground uses" of data (Wallis et al., 2013). In only a few cases in these interviews did we encounter specific foreground uses in which data were being extracted as digital entities to be recombined with other data. These included the consumer who wanted data in the most discrete units possible, and contributors who locked their data files to avoid mass downloading of datasets that might be reused without attribution.

In several cases, consumers were extracting data from DANS/EASY to create other kinds of data products. The local archaeology guide used DANS/EASY data in blog posts for other guides, and made tables and images for the printed binder he displayed to his tour participants. The practitioners in the archaeology company mined excavation reports for data on sites of interest. As the reports are in PDF format, extraction remains a largely manual process.

Background uses of data are those in which investigators seek general information about a site or a problem, such as calibration or history data (Wallis et al., 2013). DANS/EASY was an important source of background data for those we interviewed. The archaeology company started every investigation with a search of DANS/EASY to learn what is known about a site or region, for example. The local guide, the museum curator, and many others browsed DANS/EASY for data of interest. Knowing what had been done before might be more important than identifying a dataset for reuse. Background uses of data are especially valuable at the early stages of research, but are particularly hard to identify, as they go unmentioned in research methods sections or in citation lists.

An interview with an oral history contributor revealed complementary background and foreground uses of a large dataset. Whereas most users seek oral history records for their content, whether recordings or transcripts, this researcher was a computational linguist who mined data for linguistic structure. He also participated in developing the corpus.

Degrees of Openness

Data archivists mediate between contributors and consumers not only in acquiring data and making them available, but in determining who has access to what content. This mediating role is more akin to that of physical archives or special collections. Institutions make contracts with

contributors about terms of use, and users requesting access to archival materials are vetted for their qualifications and ability to meet contract terms (International Council on Archives, 2011). Digital libraries, in contrast, tend to be available under a publishing license to all members of a community, such as those of the university library (Borgman, 2007; Calhoun, 2014; Cox, 2016).

DANS/EASY, like many digital data archives, is not simply a publishing platform in which contributors deposit data for anyone to use. Most of the academic contributors interviewed deposited data on the condition that they could maintain some continuing control over who had access to those data and for what purposes. By locking datasets, they could force potential consumers to register with DANS by name and to contact the contributor directly to request access. Direct contract allows the contributor to learn about who is interested in the data, why, and for what purposes. In the best cases, a fruitful conversation leads to selective sharing of appropriate datasets, and perhaps to collaboration. Because data are so difficult to interpret outside their original context, these personal relationships can be essential to reuse (Pasquetto et al., 2017; Wallis et al., 2013).

The value of personal contact in data exchange reflects the differences in open access to publications and to data. “Data publishing” is a misnomer as an analogy to publishing journal articles and books (Borgman, 2015, 2016; Parsons & Fox, 2013). Whereas journal publishers have extensive rights to disseminate their content, digital data archives may cede continuing control to the contributors. This can be a difficult position. Archives have the data, but cannot release them directly to potential consumers, at least until certain default conditions about release are satisfied.

Characteristics of Communities

One goal of our study was to characterize the user community of DANS/EASY, especially with regard to the degree of overlap between contributors and consumers of data. The question of whether DANS’ designated community (Consultative Committee for Space Data Systems, 2012) is one or several communities could not be answered from the weblogs. The archaeology company whose staff we interviewed were heavy users, contributing about 250 reports per year and searching DANS/EASY at the start of each site investigation. The rest of the contributors interviewed had submitted only one or a few datasets to DANS over the course of their careers. These findings affirm DANS internal studies that contributions are skewed in a typical long tail curve, with a small number of organizations that submit many data sets and a large number of individuals who submit a few.

While one of the consumers interviewed was also a consumer, he was the exception. The rest of the consumers interviewed had never contributed any datasets to DANS/EASY, or probably to other data archives. They were data users, akin to searchers of digital libraries or physical archives. We were somewhat surprised to find data contributors who were not searchers of DANS/EASY. These were academic researchers who collect new data for their own projects and do not make foreground use of datasets produced by others. Methodological differences in these communities appear to explain this finding. Some DANS/EASY users, such as archaeologists, always need background information about sites from external sources. Others search DANS/EASY for foreground data to use in their own studies or practice.

We conclude that the overlap between contributors and consumers of DANS/EASY data is minimal. These are largely distinct communities, more like book authors and readers than like a community of genomics researchers for which system input and output might be more equitable. Contributors tend to be researchers in universities or practitioners in archaeology companies. Consumers are a diverse array of individuals in universities, research institutes, museums, libraries, cultural institutions, and individuals who have avid avocations in areas where DANS/EASY has content. They may search infrequently, but rely on the system as an essential resource.

How Knowledge Infrastructures Mediate Data Sharing and Reuse

Infrastructures are difficult to study because they are most visible when they break down and least visible when functioning well (Borgman, 2000; Edwards et al., 2013; Karasti & Blomberg, 2017; Star et al., 2003; Star & Ruhleder, 1996). We explored the ways in which DANS, as an established digital data archive, mediates access to data and supports processes of data sharing and reuse. They provide technical, human, and policy infrastructure for their communities, each of which has characteristics of durability and fragility.

Relationships between Stakeholders

DANS technical architecture is clearly mapped in internal documents that are beyond the scope of this study. However, even that architecture depends on technical and institutional relationships with partners in the Netherlands, the European Union, the U.S., and elsewhere. Most data are stored in national computing centers, for example, rather than in the DANS office complex in The Hague.

That DANS is a node in an international knowledge infrastructure for data provision is probably news to most of their contributors and consumers. DANS maintains interoperability agreements with many other services that harvest its data, such as OCLC, Google Scholar, and Europeana. Some of these are organizational agreements, where others depend on technical capabilities to allow particular kinds of data to be indexed and retrieved. In some cases, datasets in DANS are delivered in the background through websites maintained by contributors. Such arrangements enable institutions to provide specialized searching services or access to data from multiple providers. These arrangements may relieve DANS of specialized interface requirements, but make them invisible to those data consumers. DANS is also developing APIs (application programming interfaces) to support some of these services. All of these infrastructure activities require staff with high levels of expertise in technical standards, software engineering, law, and science policy.

Also invisible to most users is the human infrastructure that DANS provides to identify, acquire, curate, and maintain access to datasets, and the staff expertise required to assist people in searching, accessing, and using datasets. DANS contributors, especially those who contribute frequently or submit large datasets, may have personal relationships with staff. DANS staff are much less visible to consumers, who may encounter a staff member only when asking for assistance from the help desk or attending an information session. If consumers need assistance

in accessing locked datasets for which the contributors have not responded in some reasonable period of time, staff engagement may intervene.

The array of relationships between these stakeholders is presented in Figures 2 and 3.

Figure 2: Relationships between Stakeholders

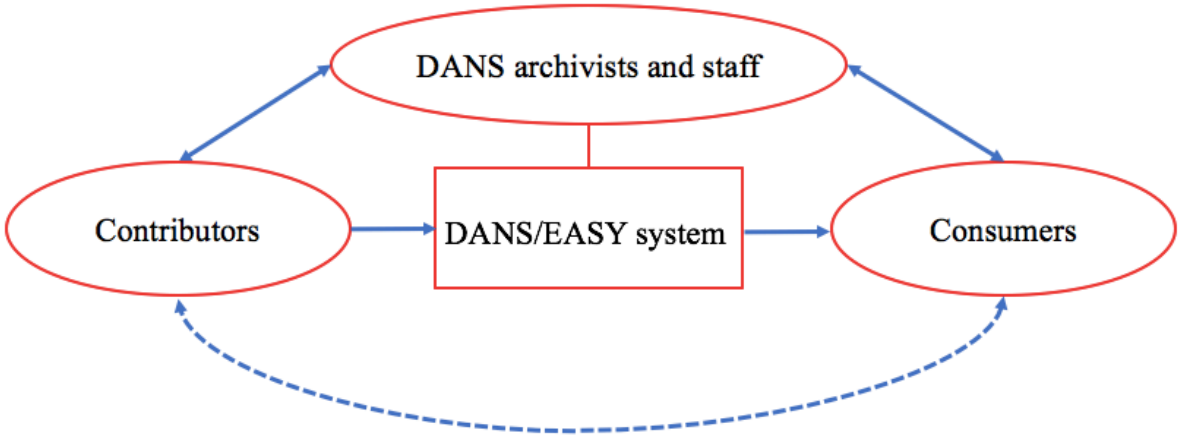
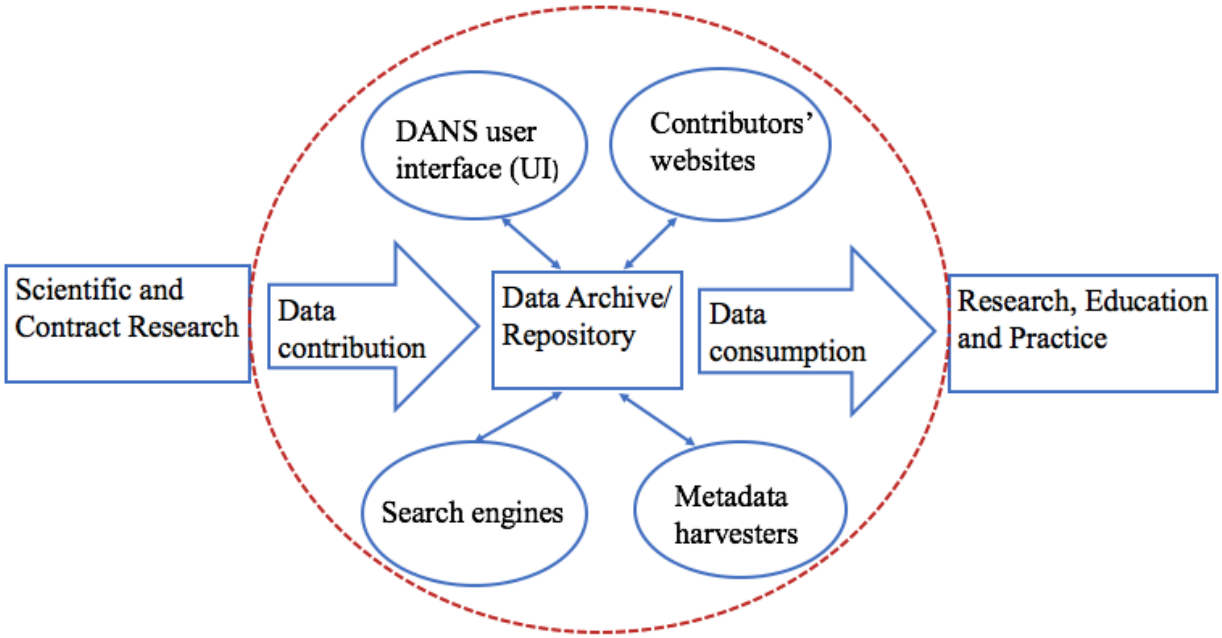


Figure 3: Flow of Data



Value Propositions and Metrics

DANS, like any other public entity, must demonstrate "value for money." The value of data reuse is hard to evaluate in economic terms because uses are far downstream from data creation and ingest. Rare is the feedback loop to data creators, given the long provenance chains from data deposit to reusers who may mine and combine data from many sources. DANS, like other digital data archives, collects statistics on almost everything that can be counted. What is valuable and what is countable rarely coincide in knowledge infrastructures, however. Staff balance time spent on acquiring new datasets and time spent on curating those datasets to make them more accessible and more durable for long-term preservation. Effort spent on acquisition and curation, in turn, is balanced with time spent serving and recruiting consumers of their data. Policy makers and funders tend to place higher value on acquisition than on preservation, which is less visible and often more expensive. Datasets, once acquired, are stored, backed up, and migrated to new generations of technology. Acquiring data is not a one-time activity or cost; it is a long-term commitment. Even if data preservation agreements expire, managing those data and disposing of them requires money and effort.

Archaeology data are the largest portion of the archive, largest portion of acquisitions, and largest overall number of downloads (Doorn, 2017), but the most downloaded individual datasets are from the social sciences, such as census data from Statistics Netherlands (Akdag Salah et al., 2012; Doorn, 2017; Scharnhorst et al., 2012). DANS serves those data as part of their mission. The originating agencies do most of the processing, thus these datasets get high usage for relatively low investment in curation by DANS staff. The vast majority of DANS datasets are downloaded at least once, which affirms their acquisitions strategy (Doorn, 2017). However, the contrast between investing in archaeology for preservation and in government statistics for access highlights the complexity of their mission.

DANS can claim a broad user base by counting registered users, uploads, searches, and downloads. To enumerate the community fully would require forcing users to register, to identify themselves uniquely, and for DANS to track user activities in detail. Such monitoring would likely run afoul of E.U. privacy regulations, in addition to creating barriers to easy use of their systems.

Archives and libraries generally avoid such tracking out of concerns for personal privacy. More complete logs would have improved sampling for this study, but not necessarily have characterized the communities much more fully. Individuals with "hands on" access to DANS were often surrogates for actual contributors. Librarians, graduate students, research staff, and company documentalists are contributing data on behalf of researchers who collect those data. Similarly, librarians, students, parents, and members of the public are searching DANS on behalf of themselves, their supervisors, their clients, their colleagues, their children, and others.

Another way to evaluate digital data archives is to track citations to data in publications. Despite advances in metadata practice to make datasets more citable, such as assigning Digital Object Identifiers and including "how to cite this dataset" notifications, data citation is not yet common scholarly practice (Brase et al., 2014; CODATA-ICSTI Task Group on Data Citation Standards Practices, 2013; DANS, 2010; Uhlir, 2012). We found that DANS consumers rarely cite the data

they acquire or cite DANS as a source. More problematic is that data contributors do not consistently cite their own datasets in DANS, which would make those datasets more visible to readers of the contributors' publications. Whatever the units counted, the numbers probably severely underestimate the actual use of DANS/EASY and the long-term value of these data.

Conclusions

Digital data archives such as DANS/EASY serve multiple communities and have many stakeholders. Despite following best management practices for tracking usage through weblogs, help desks, and other forms of outreach, it remains difficult to learn who uses digital data archives, why, or how. Anecdotes and surveys abound. Extensive qualitative studies are expensive, time-consuming, and rare. Such studies are worthwhile, as demonstrated by the rich characterization presented here of one archive's community and stakeholders. The sample represents a small portion of the archive's users, but is typical in size for an interview study. Subjects were drawn randomly from lists of registered contributors and consumers, and the full population of DANS archivists was interviewed, plus several other staff associated with DANS/EASY.

While many of those interviewed contributed datasets because they were required to do so by legal, funding, or contractual arrangements, they also expressed a commitment to open access. Openness remains a contested concept, understood and employed in varying ways among the users and stakeholders of DANS. Academic contributors preferred to maintain a degree of control over their datasets after submitting. Some protection was due to concern for indiscriminate downloading of their data without attribution, but locking datasets also forced prospective users to identify themselves. These contributors sought a direct relationship with data consumers, whether to guide them to the most effective selection and use of their data, or as a collaborative partnership. Several contributors mentioned the desire for a quid pro quo relationship, suggesting mutual sharing between contributors and consumers. However, contributors appear to have less in common with consumers than they imagine. Few of the consumers interviewed were data producers with anything to share in return. They were unlikely to request access to locked datasets. In sum, individual users may have little in common other than their interests in DANS data. Identifying them for user studies is difficult and mobilizing them in support of the archive would be a challenge.

DANS staff provide essential value in mediating access to data between contributors, consumers, and other stakeholders. DANS/EASY is far more than a database or technical infrastructure for providing access to datasets. The staff seek data from all the domains in their purview, then work individually with contributors to curate and document their data for submission. They rotate time on the helpdesk to cross-train between communities and data types. Datasets, in their view, exist only in the context of adequate metadata and other forms of description that make those data usable and interpretable by others. Staff and management make difficult tradeoffs between time spent in acquiring new datasets and curating datasets to make them more useful, and between current acquisitions and investments in long-term stewardship. The architecture must serve many stakeholders in addition to the immediate community of contributors and consumers. Data are

served from multiple data centers and from websites of some of their contributors. Data are harvested by multiple search engines and by partner digital data archives. They have multiple funding sources, each of which may have its own contractual requirements and deadlines.

DANS, like most digital data archives and host organizations, appears extremely durable with its 50+ year collection of datasets, official role in Dutch data services, and array of national and international partners. Yet, like most digital data archives and host organizations, the knowledge infrastructure is fragile in the long run (Borgman, Darch, et al., 2016). Communities and stakeholders are multiple, diverse, and competing. Data standards and policy evolve rapidly and continuously. Their systems depend on many technologies, all of which are evolving rapidly. Interoperability and maintenance are constant challenges. Funding and staffing are pieced together from multiple sources. Senior management devotes much of their time to sustaining continuity of people and money. Sustaining access to data requires knowledge infrastructures that function well and adapt readily to the fluid environment in which they exist. Toward this end, more work is needed to open the black box of digital data archives, to understand more about their communities and the organizations that play essential roles in scholarship and in information practice.

Acknowledgements

We gratefully acknowledge the support of KNAW, the Netherlands Academy of Arts and Sciences, for Visiting Scholar appointments of Christine L. Borgman, Herbert Van de Sompel, and Andrew Treloar at DANS. Peter Doorn, Director of DANS, graciously opened the DANS doors, physically and digitally, for these investigations, and provided Andrea Scharnhost's time for the project. Additional support for conducting and analyzing interviews is provided by grant # 201514001, 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. COST TD1210 also supported dissemination and meetings. Interviews with human subjects in the Netherlands are covered under UCLA IRB # 15-001291. Most interviews were conducted by Christine L. Borgman and Andrea Scharnhorst. Data analysis was conducted by Milena Golshan and Christine L. Borgman. Weblog analysis was conducted by Henk van den Berg.

Andrew Treloar and Herbert Van de Sompel contributed to the initial design of the study. Ashley Sands participated in several interviews and provided extensive comments on drafts of the paper. Sally Wyatt provided expert guidance throughout the study and commentary on interim drafts. Willeke Wendrich and Deidre Whitmore provided expertise on archaeological uses of data. Peter Darch, Irene Pasquetto, and Bernadette Boscoe also offered guidance in project design, data analysis, and comments on drafts. Most of all, we acknowledge the generosity of DANS staff for their thoughtful discussions of work practices, and of DANS/EASY contributors and consumers who welcomed us to their offices and homes, or traveled to meet us elsewhere in the Netherlands.

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* (pp. 7–10). New York, NY, USA: ACM. https://doi.org/10.1145/2390867.2390870

Arbuckle, B. S., Kansa, S. W., Kansa, E. C., Orton, D., Çakırlar, C., Gourichon, L., … Würtenberger, D. (2014). Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey. *PLoS ONE*, *9*(6), e99845. https://doi.org/10.1371/journal.pone.0099845

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G. C., Casey, K., Laaksonen, L., … Wouters, P. (2004). An International Framework to Promote Access to Data. *Science*, *303*(5665), 1777–1778. https://doi.org/10.1126/science.1095958

Blair, A. M. (2010). *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven [Conn.]: Yale University Press.

Borgman, C. L. (2000). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Borgman, C. L. (2016). Data Citation as a Bibliometric Oxymoron. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication* (pp. 93–115). Berlin, Boston: De Gruyter Mouton. Retrieved from https://www.degruyter.com/view/product/379257

Borgman, C. L. (2018). Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier. *Berkeley Technology Law Journal*, *33*(2).

Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. In *Proceedings of the Association for Information Science and Technology* (Vol. 53, pp. 1–10). ASIS&T. Retrieved from http://dx.doi.org/10.1002/pra2.2016.14505301057

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3–4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. In *2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (pp. 257–266). London: IEEE Computer Society. https://doi.org/10.1109/JCDL.2014.6970177

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & Randles, B. M. (2016). Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation*, *11*(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

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. https://doi.org/10.1002/(SICI)1097-4571(199607)47:7<568::AID-ASI8>3.0.CO;2-S

Borgman, C. L., Van de Sompel, H., Scharnhorst, A., van den Berg, H., & Treloar, A. (2015). Who Uses the Digital Data Archive? An Exploratory Study of DANS. In *Proceedings of the Association for Information Science and Technology* (Vol. 52). St. Louis, MO: Information Today. https://doi.org/10.1002/pra2.2015.145052010096

Boulton, G., Babini, D., Hodson, S., Li, J., Marwala, T., Musoke, M. G. N., … Wyatt, S. (2015). *Open Data in a Big Data World: An International Accord* (Outcome of Science International 2015 meeting). ICSU, IAP, ISSC, TWAS. Retrieved from https://twas.org/sites/default/files/open-data-in-big-data-world_short_en.pdf

Brase, J., Socha, Y., Callaghan, S., Borgman, C. L., Uhlir, P. F., & Carroll, B. (2014). Data Citation: Principles and Practice. In J. M. Ray, *Research Data Management: Practical Strategies for Information Professionals*. West Lafayette: Purdue University Press.

Calhoun, K. (2014). *Exploring digital libraries: foundations, practice, prospects*. Chicago: Neal-Schuman, An imprint of the American Library Association. Retrieved from http://www.alastore.ala.org/detail.aspx?ID=4247

Case, D. O. (2006). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (2nd ed.). San Diego: Academic Press.

CODATA-ICSTI Task Group on Data Citation Standards Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12*, CIDCR1-CIDCR75. https://doi.org/10.2481/dsj.OSOM13-043

Committee on Issues in the Transborder Flow of Scientific Data, National Research Council. (1997). *Bits of Power: Issues in Global Access to Scientific Data*. National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=5504

Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Recommendation for space data system practices No. CCSDS 650.0-M-2 Magenta Book). Washington, D.C. Retrieved from https://public.ccsds.org/pubs/650x0m2.pdf

Council of Europe. (1992). European Convention on the Protection of the Archaeological Heritage: Valletta Treaty No. 143. Retrieved May 19, 2017, from https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/143

Cox, J. (2016). Communicating New Library Roles to Enable Digital Scholarship: A Review Article. *New Review of Academic Librarianship*, *22*(2–3), 132–147. https://doi.org/10.1080/13614533.2016.1181665

DANS. (2010). *The first five years of Data Archiving and Networked Services: Self-assessment DANS 2005 - 2010* (p. 53). The Hague: DANS. Retrieved from https://www.knaw.nl/shared/resources/actueel/bestanden/DANS_Self-assessment_report_2010.pdf

DANS. (2017). DANS: Organisation and policy. Retrieved July 12, 2017, from https://dans.knaw.nl/en/about/organisation-and-policy

Darch, P. T., & Borgman, C. L. (2016). Ship space to database: emerging infrastructures for studies of the deep subseafloor biosphere. *PeerJ Computer Science*, *2*, e97. https://doi.org/10.7717/peerj-cs.97

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and

beyond. *International Journal on Digital Libraries*, *16*(1), 61–77. https://doi.org/10.1007/s00799-015-0137-3

Dillo, I., & Doorn, P. K. (2011). *The Dutch data landscape in 32 interviews and a survey*. Retrieved from http://depot.knaw.nl/10090/1/The_Dutch_Datalandscape_DEF.pdf

Doorn, P. (2017). *Datametric Analysis of DANS Data Archive* (No. version 1.0). DANS.

EASY: Published datasets. (2016). Retrieved September 19, 2016, from https://easy.dans.knaw.nl/ui/browse

Edmunds, R., L'Hours, H., Rickards, L., Trilsbeek, P., & Vardigan, M. (2016). Core Trustworthy Data Repositories Requirements. *Zenodo*. https://doi.org/10.5281/zenodo.168411

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: The MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges* (p. 40). Ann Arbor, MI: University of Michigan. Retrieved from http://hdl.handle.net/2027.42/97552

Eschenfelder, K. R., & Johnson, A. (2014). Managing the data commons: Controlled sharing of scholarly data. *Journal of the Association for Information Science and Technology*, *65*(9), 1757–1774. https://doi.org/10.1002/asi.23086

euroCRIS | Current Research Information Systems. (2017). Retrieved July 28, 2017, from http://www.eurocris.org/

Faniel, I. M., Barrera-Gomez, J., Kriesberg, A., & Yakel, E. (2013). A comparative study of data reuse among quantitative social scientists and archaeologists. In *iConference 2013 Proceedings* (pp. 797–800). https://doi.org/10.9776/13391

Faniel, I. M., Kansa, E. C., Kansa, S. W., Barrera-Gomez, J., & Yakel, E. (2013). The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 295–304). New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467712

Faniel, I. M., & Yakel, E. (2017). Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation. In L. R. Johnston (Ed.), *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository* (pp. 103–126). Chicago, Illinois: Association of College and Research Libraries. Retrieved from http://www.oclc.org/research/publications/2017/practices-do-not-make-perfect.html

Force11. (2018). About Force11. Retrieved June 26, 2015, from https://www.force11.org/about

Frank, R. D., Yakel, E., & Faniel, I. M. (2015). Destruction/reconstruction: preservation of archaeological and zoological research data. *Archival Science*, *15*(2), 141–167. https://doi.org/10.1007/s10502-014-9238-9

Getty Thesaurus of Geographic Names. (2017). Los Angeles, CA: Getty Research Institue. Retrieved from http://www.getty.edu/research/tools/vocabularies/tgn/index.html

Holdren, J. P. (2013, February 22). Increasing Access to the Results of Federally Funded Scientific Research. Executive Office of the President, Office of Science and Technology Policy. Retrieved from https://www.usaid.gov/sites/default/files/documents/1865/NW2-CCBY-HO2-Public_Access_Memo_2013.pdf

International Council on Archives. (2011, September 1). ISAD(G): General International Standard Archival Description - Second edition | International Council on Archives.

Retrieved from https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

Kansa, E. C. (2012). Openness and archaeology's information ecosystem. *World Archaeology*, *44*(4). https://doi.org/10.1080/00438243.2012.737575

Kansa, E. C., & Kansa, S. W. (2011). Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing. In *Archaeology 2.0: New Approaches to Communication and Collaboration* (pp. 57–92). Cotsen Digital Archaeology series. Retrieved from http://escholarship.org/uc/item/1r6137tb

Kansa, E. C., & Kansa, S. W. (2013). We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication. *Journal of Eastern Mediterranean Archaeology and Heritage Studies*, *1*(1), 88–97.

Kansa, E. C., Kansa, S. W., & Arbuckle, B. (2014). Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology. *International Journal of Digital Curation*, *9*(1), 57–70. https://doi.org/10.2218/ijdc.v9i1.301

Karasti, H., & Blomberg, J. (2017). Studying Infrastructuring Ethnographically. *Computer Supported Cooperative Work (CSCW)*. https://doi.org/10.1007/s10606-017-9296-7

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.

Lee, C. P., Dourish, P., & Mark, G. (2006). The Human Infrastructure of Cyberinfrastructure. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (pp. 483–492). New York, NY, USA: ACM. https://doi.org/10.1145/1180875.1180950

Mayernik, M. S. (2011, June). *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators* (PhD Dissertation). UCLA, Los Angeles, CA. Retrieved from http://dx.doi.org/10.2139/ssrn.2042653

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, *67*(4), 973–993. https://doi.org/10.1002/asi.23425

Mayernik, M. S., Wallis, J. C., & Borgman, C. L. (2013). Unearthing the Infrastructure: Humans and Sensors in Field-Based Research. *Computer Supported Cooperative Work*, *22*(1), 65–101. https://doi.org/10.1007/s10606-012-9178-y

Mayernik, M. S., Wallis, J. C., Pepe, A., & Borgman, C. L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. In *Proceedings of iConference 2008: iFutures: Systems, Selves, Society*. Los Angeles, CA. https://doi.org/http://hdl.handle.net/2142/15119

Mientjes, A. C. (2015). Archeologische Begeleiding, Protocol opgraven, Kruittoren 17 - 25, Tholen, Gemeente Tholen. *(SOB Research)*, *DANS*. https://doi.org/https://doi.org/10.17026/dans-23t-32e4

National Science Board (U.S.). (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (No. US NSF-NSB-05-40). Arlington, Virginia: National Science Foundation.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*, *56*(11), 1140–1153. https://doi.org/10.1002/asi.20204

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*, WDS32-WDS46. https://doi.org/10.2481/dsj.WDS-042

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*. https://doi.org/10.5334/dsj-2017-008

Pasquetto, I. V., Sands, A. E., & Borgman, C. L. (2015). Exploring openness in data and science: What is "open," to whom, when, and why? In *Proceedings of the Association for Information Science and Technology* (Vol. 52, pp. 1–2). https://doi.org/10.1002/pra2.2015.1450520100141

Piwowar, H. A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE*, *6*(7), e18657. https://doi.org/10.1371/journal.pone.0018657

Reijnhoudt, L., Stamper, M. J., Börner, K., Baars, C., & Scharnhorst, A. (2012). NARCIS: Network of Experts and Knowledge Organizations in the Netherlands. Map. › KNAW Research Portal. DANS-KNAW. Retrieved from http://hdl.handle.net/20.500.11755/88be28c1-4293-42e8-85b4-088433c03ae7

Research Data Alliance (RDA). (2018). About RDA. Retrieved June 26, 2015, from https://rd-alliance.org/about.html

Rosenberg, D. (2013). Data before the Fact. In L. Gitelman (Ed.), *"Raw Data" is an Oxymoron* (pp. 15–40). Cambridge MA: MIT Press.

Sands, A. E. (2017, January 1). *Managing Astronomy Research Data: Data Practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects*. UCLA. Retrieved from https://pub-jschol-prd.escholarship.org/uc/item/80p1w0pm

Scharnhorst, A., Bosch, O. ten, & Doorn, P. (2012). Looking at a digital research data archive - Visual interfaces to EASY. *arXiv:1204.3200 [Physics]*. Retrieved from http://arxiv.org/abs/1204.3200

Shankar, K., Eschenfelder, K. R., & Downey, G. (2016). Studying the History of Social Science Data Archives as Knowledge Infrastructure. *Science & Technology Studies*, *29*(2). Retrieved from http://ojs.tsv.fi/index.php/sts/article/view/55691

Star, S. L., Bowker, G. C., & Neumann, L. J. (2003). Transparency beyond the Individual Level of Scale: Convergence between Information Artifacts and Communities of Practice. In A. Bishop, N. A. Van House, & B. P. Buttenfield (Eds.), *Digital library use: Social practice in design and evaluation* (pp. 241–270). Cambridge Mass.: MIT Press.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Steinmetz Archive: Dutch Social Science Data Archive. (1989). *Historical Social Research / Historische Sozialforschung*, *14*(1 (49)), 118–121.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Tenopir, C., Palmer, C. L., Metzer, L., van der Hoeven, J., & Malone, J. (2011). Sharing data: Practices, barriers, and incentives. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4. https://doi.org/10.1002/meet.2011.14504801026

Tsoukala, V., Angelaki, M., Kalaitzi, V., Wessels, B., Price, L., Taylor, M. J., … Wadhwa, K. (2015). *Policy RECommendations for Open Access to Research Data in Europe: RECODE Project*. Seventh Framework Programme for Science in Society. Retrieved from http://recodeproject.eu/wp-content/uploads/2015/02/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf

UCLA Center for Knowledge Infrastructures: Home. (2018). Retrieved from https://knowledgeinfrastructures.gseis.ucla.edu/

Uhlir, P. F. (Ed.). (2012). *For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, D.C.: The National Academies Press.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Weber, N. M., Baker, K. S., Thomer, A. K., Chao, T. C., & Palmer, C. L. (2012). Value and context in data use: Domain analysis revisited. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901168

Wenger, E. (1998). *Communities of practice : learning, meaning, and identity*. Cambridge: Cambridge University Press.

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. *International Journal of Digital Curation*, *8*(1), 143–156. https://doi.org/10.2218/ijdc.v8i1.251

Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology & Human Values*, *33*(5), 631–652. https://doi.org/10.1177/0162243907306704