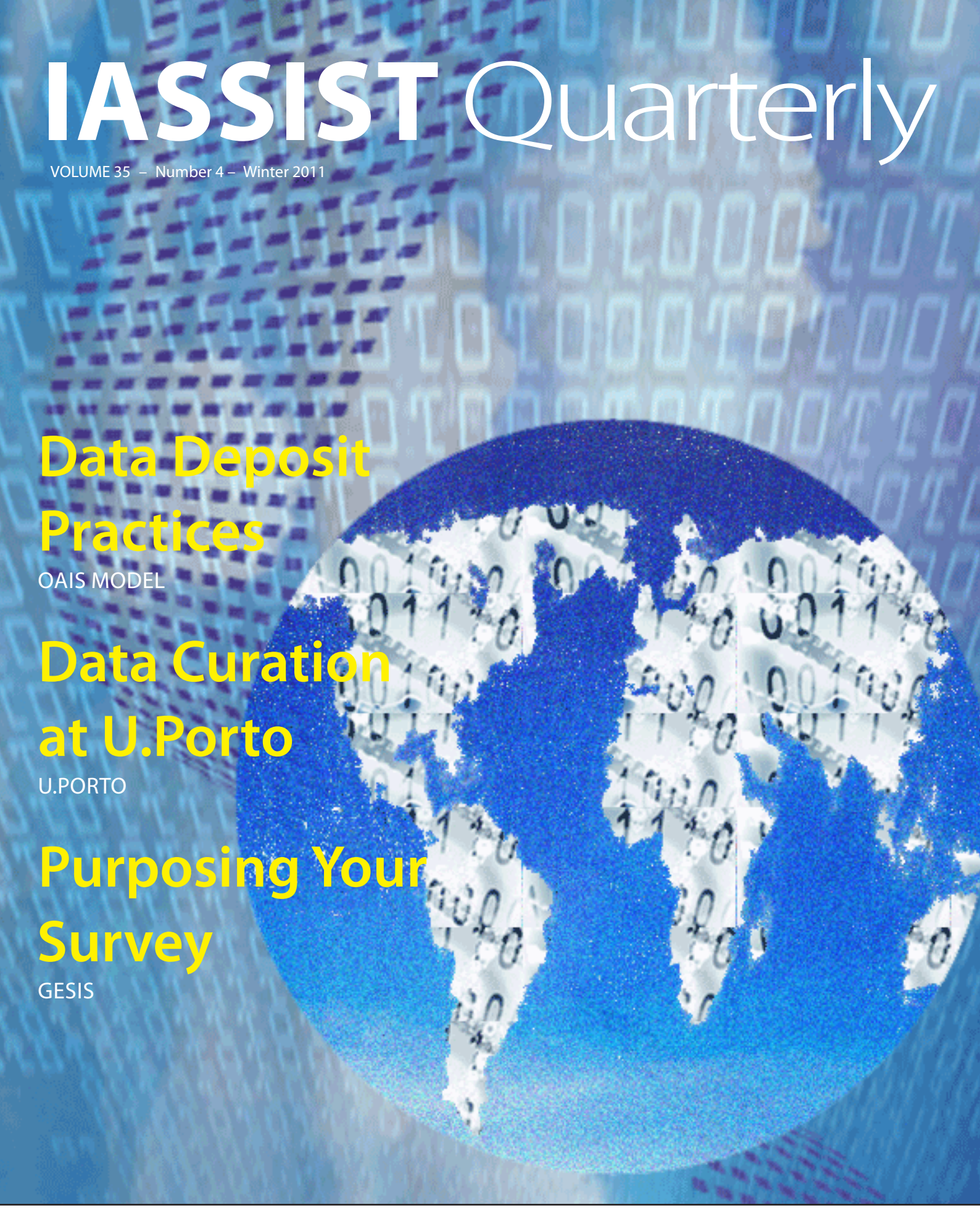
IASSIST Quarterly

VOLUME 35 – Number 4 – Winter 2011



Data Deposit Practices

OAIS MODEL

Data Curation at U.Porto

U.PORTO

Purposing Your Survey

GESIS

IN THIS ISSUE

3 Papers on Data

ON PAGE **5**

Editor Karsten Boye Rasmussen

ON PAGE **21**

Membership Information

Online at: iassistdata.org/iq

COLOPHON

IASSIST Quarterly

The **IASSIST Quarterly** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors

The Quarterly is normally published four times per year. Authors are encouraged to submit papers as word processing files (for further information see: http://www.iassistdata.org/iq/instructions-authors). Manuscripts should be sent to Editor: Karsten Boye Rasmussen.: Email: kbr@sam.sdu.dk

.Announcements of conferences, training sessions, or the like are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event

Editor

Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email: kbr@sam.sdu.dk

Deputy editor

Walter Piovesan,
Simon Fraser University

Website editor

Stuart MacDonald.
Unviversity of Edinburg

In this issue

5 **Editor's Notes**
Karsten Boye Rasmussen

6 **Examination of Data Deposit Practices in Repositories with the OAIS Model**
Social Science Context by Ayoung Yoon and Helen Tibbo

14 **Data Curation at U.Porto:**
Identifying current practices across disciplinary domains by Cristina Ribeiro, Maria Eugénia Matos Fernandes

18 **Purposing your survey:**
archives as a market regulator, or how can archives connect supply and demand? by Laurence Horton, Alexia Katsanidou

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137 **© 1997 by IASSIST. All rights reserved.**

Online @ http://www.iassistdata.org/iq

Editor's notes

Depositing data and placing it on the market

This issue (volume 35-4, 2011) of the IASSIST Quarterly (IQ) is the last of the 2011 volume. Many IASSIST members are now getting ready and looking forward to this year's conference. Probably it will turn out to be "the best ever!" and with interesting papers for the coming issues of the IQ.

This issue focuses on aspects of the depositing of data, the guidelines, regulations and formalities involved, and also on the connections between the deposit and the re-use of data.

The paper entitled "Examination of Data Deposit Practices in Repositories with the OAIS Model: Social Science Context" is written by Ayoung Yoon and Helen Tibbo from the School of Information and Library Science at University of North Carolina at Chapel Hill. The paper examines the requirements for depositing data in selected data repositories by analyzing the forms and guidelines for such deposits. The Open Archival Information System (OAIS) - an ISO standard - is used as a framework for this examination. The authors emphasize and reference others in arguing how "research data need to be available for use beyond the purposes for which they were initially collected, to make the results of studies using publicly funded data available to the public, to enable others to ask new questions of extant data and advance solutions for complex human problems, to advance the state of science, to reproduce research, and to expand the instruments and products of research to new communities". The authors use a method of content analysis in examining the requirements that exist within depositors' guidelines and deposit forms. The analysis is based upon 14 documents from 16 social science data repositories. The analysis is not looking into the actual content but registering the "required", "optional" and "not mentioned" requirements. It turned out that the documents varied significantly, including such surprises as not all repositories asked for the title or a description of the data study.

The second paper is authored by Cristina Ribeiro and Maria Eugénia Matos Fernandes from the University of Porto (Universidade do Porto). As the title outlines - " Data Curation at U.Porto: Identifying current practices across disciplinary domains" - we are now turning from comparing depositing at different repositories to the differences in data curation between different disciplines. The study has involved researchers collecting their views on data curation and data. The article includes a presentation of the University of Porto and the paper draws information from a local information system called SIGARRA (Information System for the Aggregated Management of Resources and Academic Records). This system supports authors in making their intellectual output centrally available as they sign contracts with publishers, so they maintain the right to self-archive their work in institutional open repositories that include a data repository prototype. The interviews with the researchers found that the design of a data repository should be determined by researchers' needs.

From curation and depositing of data, the authors Laurence Horton and Alexia Katsanidou from GESIS (GESIS-Leibniz Institute for the Social Sciences in Cologne) take a further step in their paper "Purposing your survey: archives as a market regulator, or how can archives connect supply and demand?". The authors start with the statement that researchers who are data creators and researchers who are data re-users have different needs and that archives mediate between them. The paper outlines the GESIS plan to create a research data management and archive training centre for the European research area. In their paper the authors give examples of how the re-use of data now has strong political support. The European Commission has committed itself to an open data policy and this is accompanied by statements like "Taxpayers have already paid for this information, the least we can do is give it back to those who want to use it in new ways…" and "Your data is worth more if you give it away". The arguments presented for data preservation and sharing are the technological and financial benefits. There are, however, continued obstacles that prevent data sharing viewed from the supply side of the social sciences. There are restrictions through law and ethics, and also a lack of incentives to share data.

As a publisher the IASSIST Quarterly supports the need to have institutional open repositories as mentioned in the second paper. We also support "deep links" where you link directly to your paper published in the IQ. Articles for the IQ are always very welcome. They can be papers from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. If you don't have anything to offer right now, then please prepare yourself for a future IASSIST conference and start planning for participation in a session there. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is much appreciated as the information in the form of an IQ issue reaches many more people than the session participants and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout:
http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
May 2012
Editor

Examination of Data Deposit Practices in Repositories with the OAIS Model

Social Science Context by Ayoung Yoon[1] and Helen Tibbo[2]

OAIS Model

Abstract
Given the significance of the role of data in research and the value of data for long-term use, researchers have been discussing the need for archiving and curating research data for future studies. To make data reusable, managing data in a reliable way and making them understandable to users is significant. This paper examines the current requirements for depositing data in selected data repositories by analyzing the forms and guidelines for such deposits. The Open Archival Information System (OAIS) is used as a framework for examining current requirements. Examining current data deposit requirements provides an opportunity to validate current data collection and management practices and provides insights into ways to improve such practices. .

Keywords: Social science data repository, data deposit, depositor requirements, ingest, OAIS model.

INTRODUCTION

The definition of "data" varies by discipline, and data can come in various formats and types. The National Research Council (1999) defines data as "facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors" (p. 15). The National Science Board (2005) uses the term "data" to refer to "any information…including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." (p. 13). The National Science Foundation classifies data into four types: (1) observational data (e.g., weather measurements and attitude surveys); (2) computational data (e.g., results from computer models and simulations); (3) experimental data (e.g., results from laboratory studies); and (4) records (e.g., from government, business, and public and private life) (Borgman, 2010, p. 19).

Given the importance of the role of data in research and the value of data for long-term use, researchers have been discussing the need for archiving and curating research data for future studies. Curating data (1) enables reuse of data for new research and new science; (2) enables retention of unique data that are impossible to recreate; (3) makes more data available for research projects; (4) enhances the ability to validate research results; (5) promotes the use of data in teaching; and (6) should be done for the public good. That data should be shared is almost universally agreed upon (Faniel and Zimmerman, 2011).

Research data need to be available for use beyond the purposes for which they were initially collected, to make the results of studies using publicly funded data available to the public, to enable others to ask new questions of extant data and advance solutions for complex human problems, to advance the state of science, to reproduce research, and to expand the instruments and products of research to new communities (Borgman, 2010; Hey and Trefethen, 2003; Hey, Tansley and Tolle, 2009).

> The OAIS reference model became an ISO standard in 2003 (ISO 14721:2003)

Despite the potential benefits of data reuse, controversies surround data sharing practices. Some argue over the ethics of sharing data and the methodological reasons

for not allowing it (Carlson and Anderson, 2007, p. 636). Others raise questions about how data collected or constructed by one researcher can be trusted or even understood by another, as data reuse generates a disconnection of the data from the people they represent, as well as from the researchers who collect them. Thus, to fill the gap generated by this disconnection and to make data reuse a common practice in scholarly communities, an explicit context for the production and establishment of appropriate systems for quality checks and assessments is essential (Carlson and Anderson, 2007, pp. 643-644).

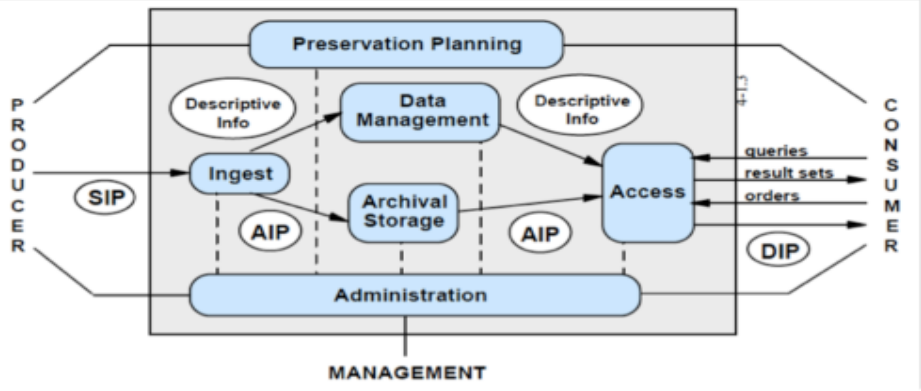


Figure 1. OAIS Functional Entities (CCSDS, 2002, p. 4-1)

TThis paper aims to understand the current requirements for depositing data in data repositories by analyzing the forms and guidelines for such deposits. The moment of deposit in repositories is key for trustworthy data management and long-term preservation. What is deposited in repositories is referred to as the Submission Information Package (SIP) in the reference model of an Open Archival Information System (OAIS), which is the first step in a data management cycle within the repository setting. Examining current data deposit requirements provides an opportunity to validate current data collection and management practices and provides insights into ways to improve such practices.

Data Deposits and the Role of SIP in the OAIS Reference Model for Data Curation

The keys to data curation are documenting, referencing, and indexing data with long-term value, enabling others to find and use them easily, accurately, and appropriately (National Academy of Science, 2009, p. 7). Because data without any a≠ccompanying necessary information concerning how and within what context they were created can be useless, all data should be well documented, associated with related materials, and linked to publications or other subsequent materials. Annotation is also significant in data curation to document changes that occur over time, allowing data to retain their long-term value (Lord and MacDonald, 2003, p. 45). For these actions to occur for curation purposes, data must be placed in a repository (Lord and Macdonald, 2003). Thus, an administrative framework must be developed that can provide mechanisms or channels for data deposit.

The OAIS reference model, which became an ISO standard in 2003 (ISO 14721:2003), provides procedures and requirements for data when they are deposited in repositories and is useful for managing any type of digital object in a "trusted" way. The OAIS reference model provides a framework that outlines archival concepts for long-term preservation and access, as well as relevant presentation information on digital objects (CCSDS, 2002). In the OAIS reference model, data from a producer[3] or creator packaged for deposit are referred to as a Submission Information Package (SIP). Within OAIS, SIPs are transformed into one or more Archival Information Packages (AIP) for preservation. AIPs are comprised of Content Information[4] and the associated Preservation Description Information (PDI).[5] Later, information from one or more AIPs becomes part of a Dissemination Information Package (DIP), which is the information package sent to the consumer in response to a request to the OAIS, enabling consumers to find and order the Content Information they are interested in (see Figure 1, CCSDA, 2002).

Each information package (SIP, AIP, and DIP) has its own role and significance in OAIS for long-term preservation and access. The implementation of the AIP can vary depending on the archives, but all required information contained in the AIP is essential for long-term preservation and access and to ensure that archival holdings remain valid. Considering the exact information content of the SIP and DIP and their relationship to the corresponding AIP, all relationships and procedures depend on agreements between archives, information producers, and consumers (CCSDS, 2002, p. 4-33). However, performing all necessary transformations of information is difficult without attaining proper SIPs, since SIPs provide a complete set of Content Information and associated PDIs to form an AIP, thus defining the fundamental significance of SIPs.

Thus, the interaction between a producer (or a depositor) and repositories is particularly critical during the process of acquiring information for a SIP. Ross and McHugh (2006) discuss the significance of the depositors' role in this process as well as the interaction between depositors (producers) and repositories. They insist that "depositors will be able to verify whether they are adequately informed when processes are completed and consulted about changes to repository procedures and services." According to them, the significance of a producer's role is determined by "the nature of the repository and its relationship with depositor" (Ross and McHugh, 2006).

In the OAIS reference model, the first interaction between OAIS and producer occurs when the OAIS preserves the data products created by the producers. The producer first establishes a Submission Agreement with the OAIS, which identifies the SIPs to be submitted and sometimes reflects a mandatory requirement to provide information to the OAIS, in contrast to sometimes voluntary offerings of information. According to the OAIS model, even if there is no formal Submission Agreement, such as in the case of websites, a virtual Submission Agreement can exist to specify file formats or other subject matter that the site will accept (CCSDS, 2002, p. 2-9).

This process of transferring information between a producer and a repository is well defined by the Producer-Archive Interface Methodology Abstract Standard (PAIMAS: ISO 20652). PAIMAS describes four main phases of the interaction: preliminary, formal definition, transfer, and validation phase (CCSDS, 2004). In the preliminary phase, all necessary preliminary information for data archiving is examined, for

instance, definition, volume of data, intellectual property, associated cost, and capability needs for ingest process. Then, a producer and a repository set the preliminary agreement. This phase should be undertaken as early as possible, even before data creation. Based on this phase, an entire process is detailed in the formal definition phase and results in the creation of a data dictionary, data model, and submission agreement. The transfer phase occurs when actual data transfer from the producer to the repository takes place, based on the previously planned agreement. When this SIP is received, the validation phase is followed, which can be automatic for some systematic parts such as file sizes or more in-depth for issues such as completeness of submission based on the plan (CCDSD, 2004, pp. 2-3 - 2-4).

The *Audit and Certification of Trustworthy Digital Repositories* (2011) also makes several recommendations regarding deposit and ingest processes to develop a trusted digital repository. Similarly to what is noted in PAIMAS, the repository should clearly specify the information that needs to be associated with specific Content Information at the time of its deposit, and should communicate clearly what producers need to provide. Although the repository is responsible for ensuring that it can extract information from SIPs and for verifying each SIP for completeness and correctness, it is recommended that the repository provide the producers or depositors with appropriate responses at agreed-upon points during the ingest process. This continuous interaction is important to ensure that the producer can verify that there are no inadvertent lapses in communications, which might result in loss of SIPs (CCSDS, 2011, p. 4-2, 4-6).

In the OAIS reference model, a typical SIP consists of the data inventory forms and actual data, or the Content Information. The inventory forms include (1) PDI (e.g., treatments, parameters measured, research subjects and IDs, date/period of collection, collection location, analysis phase, and comments) and (2) descriptive information (e.g., title, description, keywords, principal investigator's and co-principal investigator's names). Content Information is the original target of preservation in OAIS, and it refers to content data objects as well as representation information. It usually consists of physical samples, spreadsheets, final science reports, published articles, procedural documents, crew logs, photographs, videotapes, analog tapes, digital or printed images, and other types of digital data files (CCSDS, 2002, p. A-13). In the OAIS model, Content Information allows the data to be fully interpreted into meanings that can be understood by a Designated Community. If multiple data submission sessions exist, all representation information for each file should be provided, such as how frequently data submission sessions (e.g., one per month for two years) will occur and whether any access restrictions to the data exist (CCSDS, 2002, p. 2-9).

Because it is well known that compliance with OAIS would be aligned with a concept of "trusted digital repository," efforts have been made to build a system or process in compliance with OAIS. However, since the OAIS model intends to deal with digital objects in a general sense, archival communities or repositories need to translate OAIS concepts and terminology into their specific context. Of course, the elements of SIPs can differ depending on the nature of the SIPs. For instance, in a social science context, a typical example of a data object is a numeric survey data file and the associated technical information (codebook) that makes up the representation information used to understand and interpret codes in the data file. Representation information should not only include the information used to understand the numeric data (e.g., a codebook), but should also include information to enable the understanding of interpretive information. Thus, documentation on original instruments and explanations of methodology are needed to allow users to understand the question flow and determine how questions relate to variables in the resulting data file (Vardigan and Whiteman, 2007).

While efforts are made to understand data archiving processes in a certain repository and map them into the OAIS model to conform with the archival responsibilities of a trusted OAIS repository (Vardigan and Whiteman, 2007), examining this process is worthwhile in larger contexts such as social science data repositories. To respond to the growing need for the archiving and preservation of research data, examining the current status of data management practices, particularly in the SIP context, is critical to building more trusted repositories.

Methods

As previously noted, this study examines the requirements for depositors when they submitted data to repositories. To analyze the current practices among data repositories, a content analysis methodology was used, and a protocol was developed to examine the criteria or requirements that exist within depositors' guidelines or deposit forms.

For this study, data depositors' guidelines or deposit forms were collected from social science data repositories in the United States. Data can be deposited either in the institutional repository (IR) or in discipline- or domain-specific data repositories, but this study limits its scope to domain-specific data repositories that contain social science data. While both IR and domain-specific repositories aim to preserve research materials and provide access to them, they are significantly different. IRs focus more on publication-related materials from multiple subject areas within a single organization, whereas domain-specific repositories manage collections grouped by type, subject, or discipline-oriented research needs (Green and Gutmann, 2007, pp. 39-40). In addition, because the diverse nature and types of data from different domains can affect data management requirements, this study only focuses on social science data.

Social science data repositories, which are not a part of IRs, were initially identified from the three lists provided by McGraw-Hill Ryerson, Data on the Net, and International Federation of Data Organization for the Social Science[6] The three lists provide names of 47, 85, and 32, respectively, (including redundant names across the lists) social science data repositories in the world. Data repositories outside the U.S. were first excluded from these lists, which left 46 repositories in the U.S. Among those 46 repositories, government organizations that only deal with census data and do not receive data from researchers were excluded. After eliminating them, publicly available depositors' guidelines or deposit forms were collected from the data repositories' websites, but few social science data repositories have publicly available deposit guidelines or forms. It was also unclear whether some repositories accept data from individual researchers or only from government or research institutions. Among repositories that mentioned data deposits, some did not provide information about the manner in which researchers could deposit data. If the organizations mentioned that they receive data from researchers but do not provide information regarding deposits, the repositories were asked if they had written guidelines or forms for depositors. When they were asked about depositing guidelines or forms, a few had forms but only provided them when asked; others either did not yet have a procedure or were in the process of developing one. Three repositories share one deposit guideline through the partnership, thus they are counted as one repository in this study. Throughout this process, 14 documents from 16 repositories were collected in October 2011.

To conduct the content analysis, an initial protocol was developed based on the SIP elements of the OAIS model. The initial protocol included requirements regarding (1) descriptive information (project or study level), (2) actual content (data) and related information, and (3) information on files. Each category contained detailed elements. However, since the collected guidelines or forms contained different elements or requirements, the protocol was modified throughout the coding process. The resulting elements that are seen in Tables 1-3 reflect the OAIS SIP data categories and contain the specific items found in the depositor guidelines.

Table 1. Requirements for Descriptive Information found in Deposit Forms (N=14 Unique Deposit Forms)

Requirements	Frequency		
	Required	Optional	Not mentioned
Title of study	4	-	10
Description of study	4	2	8
Subject/area of investigation	4	-	10
Time period of study	2	-	11
Principal Investigator (co-Principal Investigator)	6	-	8
Data producer (of creator), if different	3	-	11
Subject term	1	-	13
Agency/funder	5	-	9
Identifier	1	-	13
Copyright check	3	3	8
Donor/contact person/depositor	4	-	10
Study metadata in general (not specified)	2	1	-

Findings

All 16 repositories are university affiliated, having partnerships with either university libraries or departments. However, as already noted in the methods section, they are not part of university IRs, but rather social science domain-specific repositories. Learning about repositories' characteristics from the information publicly available on their websites was difficult because what and how much information was shared on the web differed greatly among repositories. For example, not all 15 repositories explicitly displayed information about collection size. Numbers of staff in the repositories were generally between five to nine for repositories which provided that information, but in cases in which social science archives are run as parts of university libraries, it was hard to determine the exact numbers of staff who work for the repositories. Except for one repository, all provide online search systems or online catalogs.

Study Level Descriptive Information Requirements

Project or study level information includes information about research projects that produce data submitted to repositories. Descriptive information about the projects creating the data is significant as it provides provenance for the data. The terms used to refer to this information vary, but the concepts are similar.

The 14 collected deposit forms varied significantly. While some asked for all detailed information about a study and provided specified requirements, others had only generic requirements and asked for metadata. In this case, the data depositors determined the metadata that should be provided.

Surprisingly, not all repositories asked for the title and description of the study. A study's title is fundamental; by not always asking for title information, repositories may be assuming that titles would come with submissions or it would not be necessary for all cases since they require to submit titles of data, as can be seen in Table 2. While a description of the study would enhance the understanding of the data and provide more context, only four repositories require this information, and two repositories state that it is optional.

Some elements of a study are not necessarily the same as the information on the data being deposited, such as the subject (or area of investigation) and the time period of study. The area of investigation refers to the topical subject area on which the research was conducted, and the time period of study refers to the entire study duration, which is different from the data collection time period. Only one repository required subject terms or keywords.

Three different categories of personnel information may be required: information on the principle investigator (PI) or co-principle investigator (co-PI), information on the data producer (if different from the PI), and information on the depositor (donor or contact person). Each of these categories usually requires a home address, telephone number, e-mail address, and fax number. Some repositories asked for information on the affiliated institution. One repository specifies all three and asks for information in case they are different, but usually repositories do not differentiate among PIs for investigators, data producers, and donors or depositors of the data. The definition of donor is sometimes not well defined and could refer to either the person who deposited or who owns the data. Repositories that include a depositor agreement form with the deposit form do not ask for duplicate depositor information. Interestingly, one repository requires donors to indicate that they are willing to help potential users with any problems that they would have.

Five repositories require affiliated agency and funder information. Three repositories require a grant number with the name of the grant agency, if the research was supported by a grant.

Content (Data) and Related Information Requirements

Repositories list the actual content required to be submitted with the data, as well as information associated with the data. In general, more requirements are found on deposit forms regarding actual data and related information. These requirements include descriptive information about the data, the actual data being submitted, some contextual information usually referred to as "supporting materials" or "document description," and provenance information, which tracks changes to the data from the moment of creation.

Eight repositories ask for descriptive titles of and types of data, and seven repositories require data collection dates. Three repositories require either one or more than three subject terms to describe the content of the data, and note that they use the submitted subject terms as subject categories in their data catalog.

Among the actual files that need to be submitted to repositories, all repositories naturally require the data file. Submitting a codebook and instrument is either required or encouraged by more than half of the repositories examined in this study. However, there are variations regarding the requirements for creating a codebook. While some repositories simply suggest, "submit a codebook," three repositories

provide detailed guidelines on what the codebook should include and how researchers should prepare it. One of the repositories requires that researchers "list all variables, variable descriptions, and information to understand variables" (R05). Another repository emphasizes the significance of a well-prepared codebook, since "it is critical to interpret data and output files" (R10), and asks for the "location of variables in data, name and value, exact question wordings with exact meanings, value labels, missing data codes, etc." (R10). Three repositories ask for a data dictionary that describes indexed or other constructed variables. Types and scales of variables and technical information about variables, which refer to information such as rows/columns of variables, variable length, numbers of variables, and weighted variables, are sometimes required in a codebook. Other repositories do not state that this information should be included in a codebook, but ask that it be provided as separate documentation. One repository specifically requires information regarding the relationship between variables or tables in a data set.

One repository asks for a methodological abstract in a codebook, and half of the repositories examined in this study (seven) require separate methodology documentation. The content of the methodology section also varies depending on the repository; some just require a description of the methods, and some ask about the mode of data collection (e.g., face-to-face, telephone survey, random digit dialing, computer-assisted telephone interview, mail, web survey), time span covered by the data, and dates the data were collected. Seven repositories also ask for information on sampling, which includes coverage, sampling techniques, response rate, or procedures.

Although tracking changes to data is critical, only three repositories require documentation on data edit/cleaning procedures, or information on how the data were changed from creation to the moment of deposit. In addition, four repositories require de-identification, although this process should be required for any data containing personal information, such as names, addresses, telephone numbers, and social security numbers. De-identification is a commonly required practice in social science research, but only four repositories ask for de-identified data or check to see whether de-identification was properly done.

Seven repositories require or encourage submitting final reports or publications if such documents result from the submitted data. Three ask for proper citations for the reports or publications along with the actual reports or publications. Five of these six repositories ask for final products and require or encourage providing information on analysis performed on data.

An OAIS recommendation calls for checking for access restrictions on the data when they are deposited. Half (seven) of the repositories require providing use of restriction information.

File Requirements

Given that the repositories studied are social science data repositories, most either have a requirement for data file formats, particularly regarding statistical data, or state the "preferred" file format for submission. One repository has "no required format" (R09). Three mention that the format should be "open standard" (R01), "user-friendly format" (R04), or "in ease of use" (R10). Preferred formats or accepted file types were usually ASCII, SPSS, SAS, STATA, Excel, and ArcGIS. However, only four repositories require information on the version of the software. One repository specifies the versions of the software that it accepts (for instance, SPSS version 7.x to 16.x (R11)). R10 states that it strongly prefers ASCII to maximize the use across different software packages because "files created with older versions may limit readability and usability in the future." Three repositories require spreadsheets with CVS but in tab- or comma-delimited format, and one (R13) states that the file "should be easily converted to open or non-proprietary formats meeting ISO standards." Only one repository requires submitting information about the platform environment, which affects the software being used.

Table 2. Requirements for Content (Data) and Related Information found in Deposit Forms (N=14 Unique Deposit Forms)

Requirements	Frequency		
	Required	Optional	Not mentioned
Description about content included	4	-	10
Title of data	8	-	5
Data collection date	7	-	7
Types of data	5	-	9
Subject terms for data	3	-	11
Final report/publication generated by data	4	3	7
Data file	14	-	0
Codebook	7	1	6
Instrument	6	1	7
Data dictionary	2	1	11
Data collection methodology	7	-	7
Types and scales of variables	4	-	10
Technical information about variables	5	2	8
Sampling	7	-	7
Data edit/cleaning procedure	3	-	11
Relationship between documents/tables/variables	2	-	12
Analysis performed on data	4	1	9
De-identification	4	-	10
Use of restriction check	7	-	7

Half of the repositories (seven) examined in this study have a required format for text document files (both text as data and text as documentation about data). Other repositories do not specify the media to be submitted (paper versus digital format) and assume that all files are digital; one repository requires both paper and digital format, whereas another states that it does not accept paper. The last repository states that it will take paper if that is the researchers' only option for submission. TXT and PDF are the most common file formats preferred by the repositories, but most repositories accept other formats, including Word files (DOC), ASCII, RTF, XML, and ODT (OpenDocument Text). Only three repositories mention image/audio/video file formats, possibly because those formats are not as common as data or text files in the social science repositories. Two of the repositories prefer TIFF, JPEG (one in particular mentions JPEG2000),

and GIF files, but the other accepts a greater variety of formats such as PNG, BMP, PCD, and PCD.
In general, except for the file formats, not much information is required and not many requirements exist regarding files. Although file compression is known to possibly affect bits of information (Heydegger, 2008; Panzer-Steindel, 2007; Wright, Miller and Addis, 2009), only one repository has requirements about file compression, stating that files can be compressed using 7-zip and WinZip (R13). Three repositories specify delivery methods and media formats for depositors to use, and two repositories have a system that allows depositors to directly upload all necessary files, although they also receive files from depositors. CDs are common across repositories (R03 specifies "IBM compatible CDs"), and other delivery methods include FTP and e-mail attachments.

Among the three repositories that ask for "data edit/ cleaning procedures," only one requires data file version and update frequency information. The repository does not ask for all different versions of a data file, but does ask for the version of the submitted data file and how frequently it is updated, if it is updated.

Table 3. Requirements for Files and Related Information found in Deposit Forms (N=14 Unique Deposit Forms)

Requirements	Frequency	
	Required	Not-mentioned
Data file format	11 (1*)	2
Document file format	7	7
Image file format	3	11
Audio file format	3	11
Video file format	3	11
File compression	1	13
Data file size	4 (2**)	8
Data file naming	3	11
Software name	4	10
Software version	4	10
Platform	1	12
Data file version	1	12
Data file update frequency	1	12
Numbers of file	1	12
Delivery (media) format	3(2***)	9

*One repository mentions that it has no required format

**Two repositories mention that there is no restriction on file size.

***Two repositories ask depositors to deposit directly to their system.

Regarding data file naming, while one repository requires a list of data file names, two ask that depositors follow a specific schema. One repository recommends using a consistent and descriptive file naming scheme, enabling files to be easily identifiable for reference purposes as well as to facilitate operation of the database system. The other provides a way to describe file names, which should consist of author(s), short name of data, years, and other information.

Discussion and Conclusion

Since this study examined only domain-specific, non-IR social science data repositories, the findings may not be generalized across all social science data repositories in the United States. For instance, characteristics of small-scale data repositories that are affiliated with university departments or collections that are a part of an IR might be qualitatively different, and thus might employ different practices in accepting data from individuals. The findings of this study, however, reflect current deposit requirements and practices for university-affiliated, social science data repositories. Overall, the requirements for data deposit, both regarding the content that should be submitted and the information that should be provided to repositories, vary from repository to repository. Requirements range from minimal wherein a repository just asks the user to submit data; to more elaborate guidelines for researchers regarding how to prepare data for deposit, with detailed requirements about file naming, file format, and all necessary information that should be accompany the data.

As already discussed, the OAIS model describes the SIPs as consisting of inventory forms, which are comprised of PDI and descriptive information, and Content Information, which contains content data objects as well as representation information. The OAIS states that the PDI must include information "describing the past and present states of the Content Information, ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered" (CCSDS, 2002, p. 4–27)., Because the PDI ensures that information stored is described sufficiently so it can be accurately retrieved for future users, having a requirement for it is significant for deposits. For Content Information, the four categories of PDI (reference information, context information, provenance information, and fixity information) are critical to the integrity of the information as well as being a good practice for preservation, according to *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (1996).

The collected elements from the deposit forms in this study include elements for creating PDI, which must all be presented in the AIP later. Provenance information documents the history of the Content Information, its origins, and chain of custody (Task Force on Archiving of Digital Information, 1996, p. 16). Among the deposit forms examined in this study, some descriptive information about data, data processing information (e.g., cleaning or editing history), data file versions, and updating information is part of provenance information. Context information about the relationships of the Content Information to its environment (CCSDS, 2002, p. 4–28) would include the technical context of information, linkages among information, and social environment factors (Task Force on Archiving of Digital Information, 1996, p. 19). Among the deposit forms examined in this study, some requirements for files (e.g., formats, software information, platforms, etc), the relationship between documents/tables, and the use of restriction checks would satisfy the efforts to document the context information of data. Reference information would include study-level descriptive information as well as some descriptive information of the data (e.g., data title, data collection date, data producer, etc.) so repositories can create bibliographic metadata as well as proper citations. Fixity information exists to check if the Content Information has been altered in an undocumented manner (CCSDS, 2002, p. 4-28). While it is relatively easy for a creator of digital objects to alter or retract previously released information (Task Force on Archiving of Digital Information, 1996, p. 14), checking the number of files, measuring byte counts, recording these counts, or recording length can be one way to ensure fixity once content is within a repository. Not much fixity information is required of depositors, but some elements are discussed—for instance, the numbers of files and data file size.

The requirements for Content Information also varied across repositories, but in general, there are more requirements for CI than the other types of required information. These more extensive requirements concerning representation information may be necessary, however, since it is critical to understanding not only what variables in a data file mean, but also the actual sequence of bits that makes up the file types, which makes it possible to render the file in the future (Vardigan and Whiteman, 2007, p. 77). Complete Content Information will allow the full interpretation of data, as the OAIS model suggests.

Although the components identified from the deposit forms collected in this study include minimum elements for inventory forms and Content Information, questions persist regarding how many repositories will adopt these elements and require them for deposit, and how much these requirements reflect compliance with the OAIS model. As already discussed, since the number of forms collected in this study is small, it is hard to make generalizations from the findings. However, the findings suggest implications for developing good practices for data deposit by examining current practices and mapping them into the OAIS model. By employing good practices when data come to repositories, repositories enhance users' trust, as "trust in data was intended to strengthen as good practices and standards are established" (Carlson and Anderson, 2007, p. 645).

Future Studies

As this study solely relies on the collected documents, it may provide a limited view of SIPs and the data deposit process. For instance, to examine the full process of communication between depositors and repositories, it is necessary to know how repositories follow up on submitted data. Both the OAIS model (CCSDS, 2002) and the *Audit and Certification of Trustworthy Digital Repositories* (CCSDS, 2011) state that it is a repository's responsibility to verify each SIP for completeness and correctness so all information can be extracted for AIP and DIP. In this study, seven repositories mention proof-edit or verification processes, while others do not mention any such things at all, although it is still possible they are doing so internally. Among those seven repositories, two state that they "do not edit or proof read the contents of deposited files" (R01) or "provide comments about the quality" (R09). The other four mention that they will verify the accuracy of final files, and depositors can be contacted to reformat or reorganize the data so the repository can meet its archival needs and goals. One repository says all submitted materials and accompanying metadata are subject to the approval of the repository, and metadata can be revised to enhance access. Thus, examining internal archival processes in data repositories is essential to fully understand current data deposit practices. For instance, close examination of metadata after data is processed in repositories and comparison with metadata when it is deposited would give an insight about what information is added. Interviewing data managers or archivists would be necessary in order to fully understand how decisions about what additional information is needed are made and how missing information is acquired.

References

Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, National Research Council. (1999). A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. Washington, DC: National Academy Press.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? Presented at the China-North America Library Conference, Beijing. Available at http://works.bepress.com/borgman/238

Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. Journal of Computer-Mediated Communication, 12(2). Available at http://jcmc.indiana.edu/vol12/issue2/carlson.html

Consultative Committee for Space Data System (CCSDS). (2002). Reference Model for an Open Archival Information System (OAIS). Washington, DC, USA: The Consultative Committee for Space Data Systems.

Consultative Committee for Space Data System (CCSDS). (2004). Producer-Archive Interface Methodology Abstract Standard. Washington, DC, USA: The Consultative Committee for Space Data Systems.

Consultative Committee for Space Data System (CCSDS). (2011). Audit and Certification of Trustworthy Digital Repositories. Washington, DC, USA: The Consultative Committee for Space Data Systems.

Faniel, I. M., & Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. International Journal of Digital Curation, 6(1). Available at http://ijdc.net/index.php/ijdc/article/view/163

Green, A.G. and M. Gutmann. (2007). Building Partnerships among Social Science Researchers, Institution- based Repositories and Domain Specific Data Archives. OCLC Systems & Services: International Digital Library Perspectives 23: 35-53.

Heydegger, V. (2008). Analyzing the impact of file formats on data integrity. Proceedings of Archiving 2008, Bern, Switzerland, June 24-27.

Hey, T., Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G.C. Fox, & T. Hey, (Eds.), Grid computing: Making the global infrastructure a reality. New York: Wiley.

Hey, T., & Trefethen, A. (2008). E-science, cyberinfrastructure, and scholarly communication. In G.M. Olson, A. Zimmerman, & N. Bos, (Eds.), Scientific collaboration on the Internet. Cambridge, MA: MIT Press.

Interuniversity Consortium for Political and Social Research (ICPSR). (December 2009). Principles and Good Practice for Preserving Data. IHSN Working Paper no 003.

National Academy of Science. (2009). Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Washington, DC: NAS. Available at http://www.nap.edu/catalog.php?record_id=12615

National Science Board. (2005). Long-Lived Digital Data Collections. Available at http://www.nsf.gov/pubs/2005/nsb0540/

Lord, P. & Macdonald, A. (2003). Data Curation for e-Science in the UK: An Audit to Establish

Requirements for Future Curation and Provision. Twickenham, England: 17-55

Panzer-Steindel, B. (2007). Data integrity. April 8, 2007. Available at http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797

Ross, S., & McHugh, A. (2006). The Role of Evidence in Establishing Trust in Repositories. D-Lib Magazine, 12. doi:10.1045 july2006-ross

Task Force on Archiving of Digital Information. (1996). Preserving Digital Information. Report of the Task Force on Archiving of Digital Information. The Commission on Preservation and Access. Available at http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED395602

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. Archival Science, 7, 73-87. doi:10.1007/s10502-006-9037-z

Wright, R., Miller, A., & Addis, M. (2009). The Significance of Storage in the "Cost of Risk" of Digital Preservation. International Journal of Digital Curation, 4(3). Available at http://www.ijdc.net/index.php/ijdc/article/view/138

Notes

1. A doctoral student at the University of North Carolina at Chapel Hill, School of Information and Library Science. 216 Lenoir Drive CB #3360 100 Manning Hall, Chapel Hill, NC 27599-3360, USA. ayyoon@email.unc.edu
2. An alumni distinguished professor at the University of North Carolina at Chapel Hill, School of Information and Library Science. 216 Lenoir Drive CB #3360 100 Manning Hall, Chapel Hill, NC 27599-3360, USA. tibbo@email.unc.edu
3. According to the OAIS definition, a producer is the role played by those persons, or client systems, that provide the information to be preserved (CCSDA, 2002, p. 2-2). The Interuniversity Consortium for Political and Social Research (ICPSR) (2009) supports a producer's role in data preservation, as it "generates or is responsible for data to be preserved and provides the data to the archive or unit responsible for preservation" (p. 7).
4. The OAIS model defines Content Information as "the set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc." (CCSDA, 2002, p. 1-8).
5. The OAIS defines PDI as "The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information" (CCSDS, 2002, p. 2-11).
6. A list provided by McGraw-Hill Ryerson: http://www.soc-sciresearch.com/r6.html; a list provided by Data on the Net: http://3stages.org/c/es2.cgi?search=dataarchive&file=/data/data.html&print=notitle&header=/header/archive.header; a list provided by the International Federation of Data Organizations for the Social Science: http://www.ifdo.org/network/index.html

Data Curation at U.Porto:

Identifying current practices across disciplinary domains by

Cristina Ribeiro, Maria Eugénia Matos Fernandes [1]

U.PORTO

Abstract

The University of Porto is the largest Portuguese university with more than 60 research centers that generate a significant part of Portuguese scientific production. U.Porto is currently concerned with the curation of and the access to the scientific data generated by its researchers. Researchers are motivated to keep their data assets alive as integral part of their published results, and the scientific impact derived from open datasets is also becoming apparent. We have followed the recommendations from well-known actions in research dataset auditing to lead a short study on available data at U.Porto. The study has involved researchers from a diversity of disciplines, collecting their views on data curation and sample data. As a result we have identified some generic use cases to inform the development of a data repository prototype. Our contacts with the researchers have revealed a great diversity of situations, from groups where data curation was already integrated in the research practice to others who were struggling to incorporate it into their workflows. Our experiment was focused on data auditing and use case identification, but we also concluded that in many groups there is a strong concern with the premature exposure of the data. The sample datasets provided by the researchers are being transformed into preservation-friendly archives to be part of a data repository. We will extend the repository infrastructure with data search facilities and expect feedback from the researchers to help define the research data management services at U.Porto.

Keywords: : Data curation, management of research data, data repositories

Introduction

The University of Porto (U.Porto)[2] is currently concerned with the curation of and the access to scientific data generated by its researchers. A steady growth in research activity in all domains, international cooperation initiatives and access to data that is either generated by local projects or available via joint projects has generated many ad-hoc data archives. Research cycles of projects and scholarships are very short-term from the data assets point of view: data generated in one project may, if there is no continuation project, be abandoned and lost in less than five years. The researchers' perspective on the longevity of such data is, in general, quite optimistic and the lack of national mandates for data curation favors the continuation of this state of affairs.

In this work we have followed the recommendations of pioneering actions in scientific dataset auditing to lead a short study on available datasets at U.Porto. An analysis of current initiatives in this area has shown that close contact with researchers is essential for getting a clear view on their needs (Ribeiro, et al. 2010). Our study involved researchers from a diversity of disciplines, collecting their views on data curation and sample data (Rocha da Silva, Ribeiro and Correia Lopes 2011). As a result, we have identified some generic use cases to inform the development of a data repository prototype.

Our contact with researchers revealed a great diversity of situations. There are areas where some form of data curation is already embedded in current practice, mainly due to the requirements of publication venues or the need to share data in international initiatives. Some researchers are motivated and aware of both the value of their data and the existing threats on it, but are still struggling to incorporate data curation into their workflows. Others, faced with the possibility of having their data curated in a repository, were extremely cautious with respect to privacy issues.

Our study focused on data auditing and use case identification, but we also asked researchers for samples of their data. The datasets provided by the researchers are being transformed into preservation-friendly archives to be part of a data repository. We are extending an existing repository infrastructure with data search facilities and expect feedback from the researchers to help define the data services for the U.Porto data repository (Rocha da Silva, Ribeiro and Correia Lopes 2011).

In this paper we provide a short overview of research at U.Porto, an outline of the goals for the data curation project at U.Porto and describe its preliminary results. We conclude with some reflections on the project results and the perspectives for the management of research data at U.Porto.

U.Porto: a research university

U.Porto is the largest Portuguese university. It comprises 14 schools, a business school, 30 libraries, 12 museums and about 70 R&D units, 31 of which have been regularly classified at the top ranks by a panel of international experts as part of the Portuguese research units evaluation. Its population consists of about 30,000 students, more than 2,366 teachers and researchers (76% PhD) and 1,689 technical and administrative staff.

U. Porto offers a large range of courses covering all levels of higher education and all the major areas of knowledge. There are over 670 training programs, including undergraduate, masters, integrated master, doctoral, continuing education and specialization courses. The number of foreign students under mobility programs represents more than 8% of the total number of students. U.Porto aims at becoming a national and international reference by the high level of its students and the production and dissemination of knowledge. It can be said that the target of being among the top 100 higher-education European institutions for its 100th anniversary in 2011 has been reached.

The physical dispersion is a characteristic of U. Porto, as the buildings of the University—schools, RD&I institutes, student residences, sports and cultural facilities—are located in three separate areas of the city of Porto. Moreover, there are research institutes and centers spread throughout the city and some of them even beyond its geographical boundaries.

The shortcomings of this geographical dispersion have been practically overcome by the SIGARRA Information System (Information System for the Aggregated Management of Resources and Academic Records). SIGARRA originated in the Engineering School in 1996 and its success led to the transversal implementation in the University from 2003 on. Currently, all the U.Porto schools, as well as the Rectorate, the Social Services and some of the research centers use SIGARRA.

This integrated system was conceived to facilitate the production, flow, storage and access to the information managed by the institution—contents of pedagogical, scientific, technical and administrative nature—and to promote internal cooperation and the cooperation with external academic, scientific and business communities. The SIGARRA system interacts with other applications and systems within the University, such as the libraries, the e-learning services, the student administration and the financial management systems and also with U.Porto institutional repository, built on a DSPACE platform. Figure 1

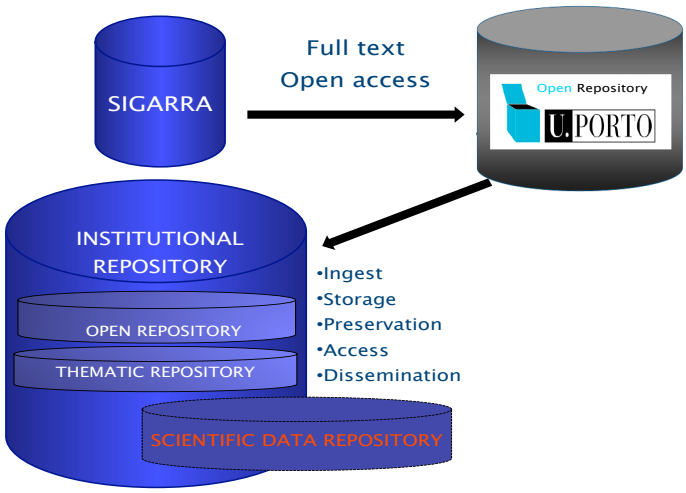


Figure 1 The SIGARRA information system and the institutional repository

illustrates the integration between the information system and the open repository.

The creation of the U.Porto Institutional Repository complemented the information management strategy by the end of 2007. The interface connecting the Information System and the Open Repository guarantees that the intellectual production of the academic and scientific community is transferred automatically from the Publications module of SIGARRA to the Open Repository. The authors just have to register and self-deposit the full text of their publications on their institutional pages, defining that they are public. The same interface also assures the connection between other applications used within the university to register and catalogue the library collections—such as Aleph—and the Open Repository, thus enforcing consistency of data across different applications and systems.

From the moment it was created, the number of publications of the Open Repository of U.Porto has grown steadily. At the beginning of 2008, the Repository had almost 1,800 full-text and open access publications. Three years later, the number of records has evolved to more than 18,000.

One of the missions of U.Porto is the creation of cultural, artistic and scientific knowledge within the academic community, composed by teachers, researchers and students. This concern has increased in the recent past due to a great variety of factors.

Beyond some aspects already mentioned above—such as the functionalities of the Publications module of the Information System and the benefits of the interconnection between SIGARRA and the Open Repository—, one cannot ignore the emphasis that has been placed on the recommendations made to the authors to make their intellectual outputs available, stressing the fact that these works are created in the context of their teaching and research activities. It is also important to highlight the suggestions made to the authors to have them consider, whenever possible, the "SPARC Author Addendum", when they sign contracts with publishers, so they maintain the right to self-archive their work in institutional open repositories, as well as the advice given to researchers to use the university-recommended format to register their affiliation.

Considering the last decade, 1/5 of the Portuguese scientific production was generated at U.Porto. Current figures show that U.Porto is responsible for more than 21% of the Portuguese scientific articles indexed in the ISI Web of Science.

Goals of the data curation project

U.Porto is currently concerned with the curation of and the access to scientific data generated by its researchers. There is a growing awareness of the fragility of personal digital archives and researchers feel that they need to keep their data assets alive as the research workflow becomes more sophisticated. The possibilities of scientific impact derived from open datasets are also becoming evident.

As a result of an identification task, we present a preliminary study on datasets that are being used in current research at U.Porto. The emphasis has been on diversity, picking examples from life sciences, engineering, social sciences and arts. The identification also provides insight on current models for data curation, both formal and informal, and on the sensitivity of researchers with respect to open access to their data (Scientific Data Curation at U.Porto, 2011).

The study has been complemented by the development of a data repository prototype. The purpose of the development is twofold: to provide services which address some of the requirements identified with researchers in a working tool and to establish the basis for a second round of interaction with the researchers, this time using the repository platform to illustrate the use cases in data curation and to test them with their end-users.

We were quite aware, from the start, of the many challenges of the project, but also of its strengths. The study conducted in the context of the national repository project (Ribeiro, et al. 2010) located similar initiatives (Rice 2009, Martinez-Uribe 2009) and existing recommendations that have helped to establish the main lines of the data audit experiment (University of Glasgow, DCC 2009). The commitment of the Rectorate and Digital University Services of U.Porto to the development of the Institutional Repository has provided a solid ground for supporting an experimental data repository. On the other hand, and in spite of the absence of mandates for data curation plans in national projects, we were able to find many researchers concerned with the management of their data and committed to sharing them within their research groups and in the context of international projects in which they are involved.

Data Auditing and Dataset Collection

IThe data audit at U.Porto has followed the recommendations issued by similar initiatives, namely the methodology proposed in the Data Asset Framework (University of Glasgow, DCC 2009). Considering that this is the first approach to data curation at the university level, we have decided to give preference to the diversity of domains. The choice of research groups to include in the study has followed a mixed strategy, selecting some groups due to personal contacts by the team members and others resulting from a call issued by the university Rectorate and Digital University Services to the directors of schools and research institutes. The first contact with the researchers led to an appointment of interviews at their laboratories, based on a script that allowed for many open questions. In cases where the researchers were willing to provide sample datasets, a follow-up interview was scheduled to discuss data formats, the definition of data and their terms of use. We adopted the recommendations of the Data Asset Framework (University of Glasgow, DCC 2009) to prepare an "Interview Guide" (U.Porto 2011) and a "Comprehensive Questionnaire" (U.Porto 2011) that were used to collect the researchers profiles, some general information on their datasets, preservation actions and expected use cases for the scenario of a university-level data repository.

There was no imposition on researchers to provide data, but most (8 out of 13) volunteered to provide sample datasets, knowing that the data would be used to design and prototype the system and that there was no agenda for a repository service, so they could not expect any immediate benefits from the collaboration.

Table 1 lists the nature of the collected datasets and the access conditions established by the researchers. Interviews were a rich source of information for their needs, where we can highlight data preservation and data exchange with research partners, either internally at U.Porto or externally in international projects and partnerships.
The collected datasets provide a first view on the research data at U.Porto, with data obtained from science, engineering and social sciences resulting from either automatic acquisition or direct collection by the researchers and access conditions ranging from open data to data useable in research but whose origin must be kept anonymous due to pending contracts. Most of the datasets under consideration were originally created as a result of research projects, but there were also data collected by external institutions with which U.Porto holds service contracts and data collected by national institutes, such as the census data created by the national statistics institute.

Table 1. Domains and access conditions for data

Domain	Dataset	Access
Astronomy	Gravimetry	Free
Chemical Engineering	Pollutant analysis	Contract pending
Mechanical Engineering	Material fracture	Embargoed
Civil Engineering	High-speed railways	Embargoed
Educational Science	Interviews	Embargoed
Psychology	Interaction records	Embargoed
Economy	Population	Embargoed
Ecology	Plant distribution	Embargoed

The interviews with the researchers confirmed our initial assumption that the design of a solution for a data repository should be determined by researchers needs, rather than by any abstract data management convenience (Borgman 2011). The interviews showed more concern with functionalities such as data browsing and querying than with strict data preservation or management.

For the 8 sample datasets provided by the researchers we created basic Dublin Core descriptions to ease their deposit into the upcoming data repository.

Future Directions for Data Management at U.Porto

The data audit at U.Porto has exceeded our expectations with respect to the commitment of researchers with data curation. In some areas

with established practices of deposit in international repositories, the data curation problem can be considered solved, but this is not the case in most domains. The resources required for this small-scale experiment are indicative of the effort required for setting up a data curation service at an institution with the size of U.Porto.

The use cases identified in this study are being used to define the requirements for the U.Porto data repository. The data samples are the basis for the design of data models where the tradeoff between generality and usefulness must be considered to make the curation process practicable. An experimental repository is being developed to test the requirements. As soon as we have a platform where some datasets are deposited and can be queried, researchers can explore it, detect the shortcomings of the proposed approach in their own domain and engage in future developments.

This work has raised even more issues than initially expected and many questions remain unanswered. We have observed that in several areas researchers are willing to participate in data curation, even in a scenario where they cannot expect any immediate benefits. This proves that we will be able to stimulate their cooperation in the following steps, but there must be some perceived gain for the researchers in order for this commitment to be sustained. A scenario where people are motivated to participate and get no practical results may ultimately compromise this and future initiatives.

The technological support for a research data repository is another open issue. The maturity of software for institutional repositories shows that we do not have to start from scratch and that basic functionality can be taken for granted. But, on the other hand, the use cases for research data are much less clear and less uniform than those for an institutional repository.

Another issue worth reflection and experimentation is the nature of data curation services. There are currently no data curation services in Portugal so there is no experience with respect to their integration in a research institution. Libraries are experienced with many of the issues in curation, but not equipped with the highly technological expertise it requires. Computing centers have complementary expertise, but their mission is centered in very different services.

Maybe the most critical aspect for the success of a data curation project is compliance with researchers needs. Institutional repositories have flourished due to the adoption of repository technology, originally created to satisfy very specific needs, by the more traditional library community. There are currently no well-established generic platforms for research data management but many custom-designed systems already exist. Experience and successful developments will show whether generic platforms can cater to the needs of researchers in different domains or if they have to be more specialized by discipline.

References

"U.Porto Comprehensive Questionnaire." UPData. 2011. http://science-data.up.pt/doc/ (accessed November 2011).

"Interview Guide (in Portuguese)." UPData. 2011. http://sciencedata.up.pt/doc/ (accessed November 2011).

Scientific Data Curation at U.Porto. Edited by João Rocha Silva. 2011. http://sciencedata.up.pt/updata/ (accessed November 2011).

University of Glasgow, DCC. The Data Asset Framework Implementation Guide. October 2009. http://www.data-audit.eu/ (accessed November 2011).

Borgman, Christine L. "The Conundrum of Sharing Research Data." Journal of the American Society for Information Science and Technology, 2011: 1-40.

Hey, Tony, Stewart Tansley, and Kristin Tolle, . The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft, 2009.

Martinez-Uribe, Luis. Using the Data Audit Framework: an Oxford case study. University of Oxford, http://ie-repository.jisc.ac.uk/300/, 2009.

OECD. OECD Principles and Guidelines for Access to Research Data from Public Funding. http://www.oecd.org/dataoecd/9/61/38500813.pdf, 2007.

Ribeiro, Cristina, Eloy Rodrigues, Maria Eugénia Matos Fernandes, and Ricardo Saraiva. Repositórios de Dados Científicos: Estado da Arte (in Portuguese). Project Report, Porto: RCAAP, 2010.

Rice, Robin. DISC-UK DataShare Project: Final Report. Project Report, Edinburgh: University of Edinburgh, 2009.

Rocha da Silva, João, Cristina Ribeiro, and João Correia Lopes. " UPData-A Data Curation Experiment at U.Porto using DSpace." Proceedings of 8th International Conference on Preservation of Digital Objects, iPRES 2011. iPRES, 2011.

NOTES

1. Cristina Ribeiro, DEI-Faculdade de Engenharia da Universidade do Porto/INESC TEC, Rua Dr. Roberto Frias, s/n, Porto, Portugal, mcr@fe.up.pt. Maria Eugénia Matos Fernandes, Reitoria da Universidade do Porto, Universidade Digital Praça Gomes Teixeira, Porto, Portugal, efernand@reit.up.pt.

2.U.Porto: Homepage. http://www.up.pt/

Purposing your survey:

archives as a market regulator, or how can archives connect supply and demand?

by Laurence Horton, Alexia Katsanidou[1]

GESIS

Abstract

What do researchers need from archives? What do archives need from researchers? These questions cover two types of researchers that encounter data archives: those who create the data (data creators) and those who re-use it (data re-users). These groups have different needs and archives mediate between them.

The role of an archive for creators and re-users is to support them in producing quality data, metadata and documentation and to facilitate wide and multipurpose data dissemination. By supporting multipurpose reuse, to the fullest extent possible, archives help realize the value of public investment in academic research.

This paper discusses the optimization of research data management training and support for research data creators, and data dissemination and long-term preservation for social science data archives. It outlines the GESIS plan to create a research data management and archive training centre for the European research area, to cater to both data supply and data demand.

The training centre will look to ensure excellence in the creation and long-term preservation of reusable data in the European Research area, contribute to promoting and to the adoption of standards in research data management, and promote data availability and reuse. Finally, the centre will provide and coordinate training on technologies and tools used by data professionals.

Keywords: : Archives, research data management, incentives, sharing, training.

Introduction[2]

Social science data archives connect two primary audiences. One is data creators––those who bring social science data into being. In this category, we place principal investigators of studies as well as researchers who work in data collection procedures. The other audience is data re-users. Here we mean researchers who either use data they themselves created some time ago or use data created by others to examine social phenomena.

The ligaments connecting these audiences are data archives: organizations that facilitate data ingest and dissemination. By accepting data into their catalogue for preservation and reuse, then furnishing the research community with that data, the archives establish a connection between the two audiences. However, it is a dynamic relationship fashioned by two forces: a movement towards data sharing for reuse and a set of resistances to data reuse.

In this paper, we discuss these forces and we highlight actions to promote data sharing and reuse. The basis of our perspective is a supply and demand model of data archives and thus the basis of our proposals are for both audiences. We focus on attempts to introduce practical policy suggestions to facilitate an easier relationship between creators, archives, and re-users primarily within the CESSDA-ERIC consortium of European social science data archives.

The Data Sharing Movement

The contemporary movement towards data sharing for reuse is a trend enabled and assisted by technological innovation. The means by which one can share data and collaborate on research have become cheaper and easier to utilize. Negating the barriers towards reuse and collaboration posed by time, distance, cost, and logistics are developments in instantaneous means of communication, large capacity data transfer, cheaper digital storage costs, and the power of data analysis software packages. Today we can do more research with more data in less time and at less cost. Indeed the range, scope and potential applications of data created, available, and analyzed can reach such a size that it may even

challenge the primacy of the experimental hypothesis approach in doing social research (Anderson, 2008).

In recognition of these phenomena, the European Commission commissioned a report on how to best direct this changing data environment towards scientific and economic innovation. Its High Level Expert Group on Scientific Data envisioned

> …a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense [...] the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance (European Union, 2010 p.4)

The belief that technology is changing patterns of research and publications has a normative basis in the argument that publicly funded data is a public good and that funders can maximize the value of research they support with a requirement that data be shared to the fullest extent possible. This argument is based on the position of the Organization for Economic Co-Operation and Development (OECD) that publically funded research data should as far as possible be openly available to the research community for re-analysis, repurposing, and long-term preservation (OECD, 2007).

The *Riding the Wave* report (European Union, 2010) echoed an expectation of transparency in data creation. An expectation that the methods of generating and manipulating data be clear so data is comprehensible to others outside of, and remains comprehensible to as time passes, the original data creators themselves. In addition, there is an acceptance as the norm in good scientific research that findings be based on data that is available (where legally and ethically possible) for independent verification, analysis, and reuse. This is a movement accelerated by a requirement of some academic journals that publication of articles is dependent on the authors' making available the underlying data if it is not already accessible. We find an example of this trend in Dryad. Dryad is an open data repository for articles published in the natural sciences and lists a number of journals as partners for which it either holds, or works with, to preserve and disseminate data (Dryad, 2011). An additional example is European Data Watch Explained (EDaWaX) (European Data Watch Extended, 2011) This German-based project examines the absence of incentives in economics for the replication of results and data reuse with the intention of creating a publication data archive. A similar project for political science, but with narrower focus is the GESIS Data Infrastructure team's Data Policy availability project. This project empirically investigates data policies of all top academic journals in political science, analyses their content and finally proposes policy guidelines.

In an era of tight pressures on public spending, the political attraction of these arguments is clear. The European Commission has committed itself to an open data policy that it estimates would provide an extra €40 billion a year to the EU economy. "Taxpayers have already paid for this information, the least we can do is give it back to those who want to use it in new ways..." stated Commission Vice President Neelie Kroes. "Your data is worth more if you give it away" (European Commission, 2011a) she added. However, the EC policy is tied to public sector data, not publicly funded academic research data which remains exempt (European Commission, 2011b). Yet this too can be, and is, considered a public investment to be shared thereby maximizing its value. We find examples of this belief in the emergence of policies that mandate data sharing be addressed as an aspect of proposals seeking public funding.

The United States National Institutes of Health (NIH) enforced a data sharing policy in 2003, with a requirement for funding applications to include a plan for data sharing (National Institutes of Health, 2003). The National Science Foundation (NSF) followed in early 2011 by adopting a similar requirement to produce a data management plan for sharing (National Science Foundation, 2011).

In the American environment it is often institutions that provide a preservation and dissemination service. Examples include University of California-San Diego (2010), University of Illinois at Urbana-Champaign (2005), Cornell University (2005), Massachusetts Institute of Technology (2005), and University of Rochester (2008). However, these approaches have been institution-specific rather than national infrastructure tools as NIH and NSF aside, the United States lacks the regional, national and supranational level funding regime of European countries such as the United Kingdom and Germany.

Similar developments have occurred in Europe. In May 2011, Research Councils UK––the strategic partnership agency of the United Kingdom's seven main research councils––published a set of common principles on data policy intended to provide an overarching framework for individual council policies on data reuse. The principals include an explicit statement that:

> Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property. (Research Councils UK, 2011)

UK councils may vary in the specifics of data, but this principal holds across the field. The Economic and Social Research Council (ESRC), Natural Environment Research Council (NERC) and the British Academy all mandate research data be offered to data centers. In the case of the ESRC (Economic and Social Data Service, 2011) and NERC (Natural Environment Research Council, 2011), through council funded data centers. Other UK funders expect or encourage data sharing but do not mandate places of deposit. The Engineering and Physical Sciences Research Council (EPSRC) has introduced a policy (from May 2015) mandating that institutions ensure well documented data is preserved and available for a minimum of 10 years from last request for access by a third party (Engineering and Physical Sciences Research Council, 2011). From an institutional perspective, University of Edinburgh, followed by the University of Hertfordshire (2011), became the first UK universities to adopt an institutional Research Data Management Policy. This included, in Edinburgh's case, a commitment that:

> Research data management plans must ensure that research data are available for access and re-use where appropriate and under appropriate safeguards (University of Edinburgh, 2010).

In Germany, the main publically funded research organizations have adopted a set of principles for the handling of research data. This 2010 agreement does not take as strong a tone as its RCUK equivalent; however, it does support long-term preservation and the "principle" of open access to research data, as well as the development of subject-specific requirements, standards, and metadata to facilitate interdisciplinary research and supporting infrastructure (Alliance of German Science Organisations, 2010). These principals drew, in part, from an earlier set of proposals submitted by the German Research Council (DFG) that encourage researchers to take into account data management issues. Reinforcement of this invitation is by guidelines promoting data sharing for experts on review panels.

The DFG raise the issue of data management and demand secure preservation and visibility for those data publically funded and used for publications, but limit this demand to a ten-year period (Deutsche Forschungsgemeinschaft, 1998). Since then, greater effort has occurred to promote effective and consistent data management but not explicitly formulated in an official publication of the German Research Council.

Thus, the causes of a movement towards data preservation and sharing are clear: technology and financial benefit. Furthermore, the demand is there. Two of the largest data archives, the UK Data Archive (UKDA) as part of the Economic and Social Data Service (ESDS) and in the United States the Inter-University Consortium for Social and Political Research (ICPSR) (2011a), have both seen significant increases in orders for data they hold since offering online access to data (Economic and Social Data Service, various). A similar phenomenon is apparent in the GESIS Leibniz-Institute for Social Science's user statistics––specifically for Eurobarometer data, for which the number of datasets distributed has jumped between 2005 and 2009 (GESIS Leibniz Institute for the Social Sciences, 2010).

Resistances to Data Reuse

However, let us look at the supply side in the social sciences. Here there are still obstacles that prevent data sharing. Primary limitations are those placed by law and ethics. Neither data archives nor funding agencies believe in sharing all data with everyone, or even within the academic community. The policies and recommendations presented above recognize, as we do, that there has to be protection of intellectual property, professional credit, and critically––moral and ethical protection of research participants.

However, alongside these recognized limitations there are additional resistances to data sharing. Opposition remains to the idea of sharing research data. This phenomenon in the social sciences can draw on a range of arguments.

Low-level (researcher-level) ignorance as to why others would want to use their data. This was a reason cited by a small number of researchers interviewed for the UKDA's Data Management Planning for ESRC Centres and Programmes (UK Data Archive, 2010 pp. 17-21). It is not resistances to data reuse itself, but an inability to imagine that the type of data generated would be of interest to anyone else. We can overcome this problem through more interaction within the scientific community and open presentation of opportunities for data sharing.

Additional to the ignorance of researchers about potential reuse of their data, there are also epistemological concerns. These cover congruence, reflexivity, and context. Essentially, data creators holding this objection claim understanding and value of data can only exist in the specific context of their creation. They are concerned that their data, abstracted from the methodologies and ontologies adopted at the time of creation cannot adapt into a different research project. These problems of course need proper consideration particularly where the reflexive relationship between researcher and participant is critical to understanding the data, but given appropriate documentation, they should not prevent future reuse[3]

A clear problem is the lack of incentives to share data. As long as the main metric of career progression remains publications and citations of publications, data sharing will be a secondary concern. However, data creation requires the investment of a lot of scientific effort and expertise. Reusing an existing dataset builds on the scientific work of other researchers who should be not only acknowledged, but also credited for their achievements. Widespread recognition and implementation of a system for acknowledgement of data citations as an indication of research quality and establishing them as equivalent to publication citations would remove a reservation against data sharing.

Data creators often have concerns as to the ethics of reuse concerning research participants. Specifically, a concern of compromised anonymity and confidentiality of participants emerges when disseminating data to other researchers. There are ways to anonymize data but some data are extremely sensitive and easily trackable. Thus, researchers can be reluctant to share on principle of protecting their participants' anonymity.

We propose that the character and structure of the current social science research environment determines attitudes to reuse. Outside of large-scale surveys, the concept of data reuse is not dispositional. There is still no established culture of archiving, sharing and reuse. The environment described above is situational. A strong situational determinist research environment should not only coerce researchers into creating reusable data, but also give them confidence to do so, thereby creating a researcher disposition towards creating reusable data. Using the colloquial metaphor that seems to be prevalent in research data management discussions, the current situation is mostly sticks and few carrots, and we need more carrots.

Promoting reuse: Cognition vs. Emotion

There is a case to be made, and has been made by funding councils and institutions, that data management and reuse be addressed as a mandatory requirement in any funding application. The reasoned argument for data management stands clear: it is fundamental to transparent, high quality sustainable data generation. Therefore, in psychological terms, data management for reuse is a "cold", cognitive task – an intellectually conscious, controlled process based on explicit learning (Kahneman, 2003). However, often the resistance to reuse draws not so much on logic, but sources that are more emotive. Drawing on movements within political psychology, what we feel should not happen is to dismiss emotive impulses.

We believe that emotions should be brought into the discussion between data creators and re-users. This is predicated on the belief that emotional responses are great motivators. Emotions can be harnessed to aid decisions, for example, the emotion to care. Ambition, incentives, professional acclimation can all be connected with data sharing and help researchers reach their decision to share. Researchers make an effort in collecting and working with data, and therefore they should develop an affective relationship with them. They are their intellectual creators and they should be given reason and tools to present them to the community in the same way they do with publications. If we can tie good research data management and data sharing into recognized career advancement, we can bring with it esteem of peers not just for the publications but the data underpinning publications. If we can instill professional pride in replication and peer scrutiny of data creation like the academic community has instilled in journal publications, then by sharing data researchers will be a more "important" with wider recognition than those who do not share because they will help advance the state of their discipline. Those who chose not to, however, will have another emotion to mange – fear: the fear of professional irrelevance (King 1995, p.445). For without emotions such as care, or fear, what incentive––and as we have suggested, incentives are currently lacking––is there to think of the consequences of actions? Through

this, we could hope to see a dispositional environment towards data sharing emerge.

Support for data sharing procedures is an important factor in facilitating sharing as lack of awareness can be a serious obstacle. While resources exist to support data creators in generating reusable data, they are often not discipline-specific. For example, the first versions of the Digital Curation Centre's (DCC) Data Management Planning Tool (Digital Curation Centre, 2011) or the Australian National Data Service's data management planning advice (Australian National Data Service, 2011) offer detailed but generic support. Although discipline-specific focuses are emerging, promoted in part through programs like JISC's Managing Research Data (Joint Information Systems Commission, 2009), as most are either generic tools or pure data management projects, these resources do not occupy the brokerage positions that data archives can assume.

The "brokerage" role of data archives – the supply and demand model

The responsibility of a broker is as a third-person facilitator to bring "sellers" and "buyers" together. We can therefore think of the brokerage role for an archive in terms of facilitating the "buying" (acquisition) and "selling" (dissemination) of data between data creator and data re-user. Archives know their "market" for data, and have established relations with creators "sellers" and re-users "buyers", they are institutions that talk to both communities from acquisition to dissemination via ingest. Consequently, they become important regulators of this data market. They regulate the inflow and the quality of data on the supply side by encouraging data creators to share, leading the move to professional credit for sharing by making data citation possible and advising and supporting data creators on avoiding unnecessary obstacles to creating shareable data. However, they also regulate the output of data towards the demand side by disseminating them, increasing their visibility, and providing a service for responsible reuse of data.

To highlight four cases, the UKDA (2011), the ICPSR (2011b), in the Netherlands the DANS (Data Archiving and Networked Services, 2011a) and the IQDA (Irish Qualitative Data Archive, 2010) are national archives that have produced resources to aid data creators as well as providing data and dissemination support.

However, archives do not only regulate supply and demand. Through division of labor and specialization, they also add value to the data life cycle. Archives undertake tasks that enhance data quality and data survival in an uncertain technological world. Though not exhaustively, data archives provide long-term preservation of data with a strategy to ensure readability as file formats and technologies change. In addition, archives add value to data through structured metadata, catalogue records, and harmonization with comparative data collections. Archives develop networks for secure and easier access of data for reuse.

Nevertheless, to provide high quality data, archives must adopt modern technologies and standards, ensure cooperation between same-discipline archives across countries, and promote dialogue with archives operating in other disciplines. Through systematic interaction, archives can be the critical ligament that facilitates data sharing.

Incentives

The role of the archive is to build incentives for both audiences to adopt best practices when dealing with data. From the supply side, it is important to increase the cognitive and emotional incentives for data sharing. We have already stated the important enticement for creators in making data available for reuse is their publications record, as their rewards and career advancements depend on that. The first step is then to make data citable. To do so, we need to provide the infrastructure and technology that allow the efficient referencing of data files. The most commonly used form of identifier is the Digital Object Identifier (DOI®) System (International DOI Foundation, 2011). These persistent identifiers are codes that connect a digital object such as a dataset, with accompanying metadata that includes author names, year of data collection and other important information of relevance. DOIs digitally identify journal articles, thus researchers are already familiar with their basic uses and functions. By having a DOI allocated to a dataset, the researcher can be sure that by using that specific DOI they refer to the same dataset. Therefore, referencing a dataset within the publication used to create it becomes effective. A reader of this publication can then identify the very same dataset with no alterations and replicate the analysis. This ensures research quality and the primary investigator is acknowledged.

GESIS is a data archive that has a project providing persistent identifiers for data files in its collection. One example, hosted by GESIS, is the da|ra project (GESIS Leibniz Institute for the Social Sciences, 2011) GESIS's registration agency for social science research data. The da|ra infrastructure lays foundations for permanent identification, storage, and localizing to create citable research data. Initiated in 2010 with a pilot phase, on entering 2012 the project is now in an upgrade phase. An expansion phase from 2013 to 2014 will centre on the development of useful services like user statistics, citation indexing, peer review possibilities for data, and registering of other data formats.

Another project of note is the effort by DANS (Data Archiving and Networked Services, 2011b) to produce a competitive alternative to the DOI. The Dutch archive is involved in the design and implementation of a persistent identifier (PI) infrastructure in cooperation with the infrastructure-oriented SURFfoundation [sic], and Koninklijke Bibliotheek (National Library of the Netherlands). This collaboration seeks to establish a mechanism called the National Resolver that would translate the PI into the current URL of the object.

In highlighting the projects and arguments we have presented thus far, it is our main goal to encourage researchers to take pride in their data creation activity, not just the outputs, and to invest time in making it reusable and archivable. We also aim to encourage researchers to value the work of other researchers who collect data, and to acknowledge this process as important and equal to other publication activities. To do that we focus on a new innovative data management training facility which we are involved in developing at GESIS: the Archiving and Data Management Training and Information Center (GESIS, 2012)

A concept for training

A new development that builds on the supply and demand model is the GESIS plan to create a research data management and archive training centre for the CESSDA-ERIC European area. This area is inclusive of data archives in twenty European nations (CESSDA, 2011)

The training centre will provide a central reference point for European researchers and archives, containing original resources and links to significant external resources, with the aim to ensure excellence in the creation and long-term preservation of reusable data, contribute to promoting the adoption of standards in research data management, and to advance data availability and reuse. The centre will also provide and coordinate training on technologies and tools used by data professionals.

By networking, and through surveys of demand for training needs, we are identifying themes and developing training concepts through potential collaborations with expert instructors. These concepts will be the basis on which courses are developed. The idea is to build resources around them using mixed and matched smaller thematic units depending on the needs of each specific course.

Our website will hold resources created by us, links to external resources, and will host information on the training center's consulting activities. The Virtual Centre of Competence will allow for consultation on best practice in research data management and archiving, including personal development and the promotion of skills training, provide information on our training activities, and offer structured teaching and self-learning materials.

Specifically, the centre will support data creators in implementing international standards of metadata and documentation. Information for data creators about the importance and uses of persistent identifiers and will be given, plus advice on ethics and consent, details on issues of data ownership, and an overview of archiving software systems.

The main support for data reuse is through the training of data archive staff to provide quality user support and to deal with increased volume of support requests. In addition, there will be information for archive professionals about new projects, new technologies, data discovery, and dissemination tools. Furthermore, the presentation of projects on data harmonization will enable archive professionals to add to the value of data for their users and create an online user community engaged in task of harmonization.

Finally, and perhaps most importantly, the training centre looks to support other archives, libraries and repositories in ensuring state of the art data-related functions and in keeping up with the constant development of new technologies. This feature is not only useful for institutions either in a formative stage or that are not specialized social science resources, it is essential for all institutions operating in the data world to keep up with innovation, establish clear workflows, and strive for internationally accepted standards.

The centre seeks to bring together the best examples and expert individuals to provide training. Training will not only have the traditional form of workshops. It will be an active form of community building and incentive development through all communication channels provided to us by the new technologies. The core of our training concept is to negate all the reasons outlined in this text that allow researchers to sit on their data without sharing, and this can only be done with systematic incentive building.

This training centre is only one way to augment the incentives of data sharing by bringing the subject closer to researchers' hearts. However, the other driving factors mentioned and analyzed in this paper have to be pushed forward in order to ensure the emotive connection of researchers to sharing data, and to establish it an integral part of the scientific contribution. In the world of data, the imperative to share is clear. We have enough sticks; it is time to cultivate the carrots.

References

Alliance of German Science Organisations (2010) "Principles for the Handling of Research Data" http://www.allianzinitiative.de/en/core_activities/research_data/principles/

Anderson, C. (2008) "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired Magazine, 16(8) http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Australian National Data Service (2011) "Data Management Planning" http://ands.org.au/guides/data-management-planning-awareness.html

Bishop, L. (2009) "Ethical Sharing and Reuse of Qualitative Data", Australian Journal of Social Issues, 44(3), pp.261-267

CESSDA: Council for European Social Science Data Archives (2011) "Member Organisations" http://www.cessda.org/about/members/

Cornell University (2005) "Registry of Digital Collections" http://rdc.library.cornell.edu/search/index.php?mode=browse&type=Collection

Data Archiving and Networked Services (2011a) "Data Management Plan" http://www.dans.knaw.nl/en/content/categorieen/diensten/data-management-plan

Data Archiving and Networked Services (2011b), "Persistent Identifiers" http://www.dans.knaw.nl/en/content/categorieen/diensten/persistent-identifiers

Deutsche Forschungsgemeinschaft (1998) "Proposals for Safeguarding good scientific practice Wiley-VCH" http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Digital Curation Centre (2011) "DMP Online" https://dmponline.dcc.ac.uk/

Dryad (2011) "Dryad Partners" http://datadryad.org/partners

Economic and Social Data Service (2011) "About the Economic and Social Data Service" http://www.esds.ac.uk/about/about.asp

Economic and Social Data Service (various) "Annual Reports" http://www.esds.ac.uk/news/publications.asp

Engineering and Physical Sciences Research Council (2011) "Implementing the Delivery Plan" http://www.epsrc.ac.uk/PLANS/IMPLEMENTINGDELIVERYPLAN/Pages/default.aspx

European Commission (12 December 2011a) "Digital Agenda: Turning government data into gold" http://europa.eu/rapid/pressReleases-Action.do?reference=IP/11/1524&format=HTML&aged=0&language=EN&guiLanguage=en

European Commission (12 December 2011b) "Digital Agenda: Commission's Open Data Strategy, Questions & answers" http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/11/891&format=HTML&aged=0&language=EN&guiLanguage=en

European Data Watch Extended (2011) "About EDaWaX" http://www.edawax.de/about/

European Union (2010) "Riding the Wave: How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data" http://cordis.europa.eu/fp7/ict/e-infra-structure/docs/hlg-sdi-report.pdf

GESIS Leibniz Institute for the Social Sciences (2010) "Eurobarometer Data Service – International Data Sets Distributed via Archive Networks, 2005-2009" http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/eurobarometer/contacts/eb-user-stat_archives_2005-2009_v2.pdf

GESIS Leibniz Institute for the Social Sciences (2011) "Über da|ra" http://www.gesis.org/dara/home/ueber-dara/

GESIS Leibniz Institute for the Social Sciences (2012) "Archiving and Data Management Training and Information Center" http://www.gesis.org/archive-and-data-management-training-and-information-centre

Inter-university Consortium for Political and Social Research (2011a) "ICPSR Usage Statistics" http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/usage.jsp

Inter-university Consortium for Political and Social Research (2011b) "Guidelines for Effective Data Management Plans" http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/index.jsp
International DOI Foundation (2011) "Welcome to the DOI® System" http://www.doi.org/
Irish Qualitative Data Archive (2010) "Preparing Qualitative Data for Archiving" http://www.iqda.ie/content/preparing-qualitative-data-archiving
Joint Information Systems Commission (2009) "Managing Research Data (JISCMRD)" http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx
Kahneman, D. (2003) "A perspective on judgment and choice: Mapping bounded rationality" American Psychologist, 58(9), pp.697-720. doi: 10.1037/0003-066X.58.9.697
King, G. (1995) "Replication, Replication" PS: Political Science & Politics, 28, pp.444-452
Massachusetts Institute of Technology, "DSpace@MIT" http://dspace.mit.edu/
National Institutes of Health (2003) "Final NIH Statement on Sharing Research Data" http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html
National Science Foundation (2011) "NSF Data Management Plan Requirements" http://www.nsf.gov/eng/general/dmp.jsp
Natural Environment Research Council (2011) "Data Centres" http://www.nerc.ac.uk/research/sites/data/
Organisation for Economic Co-operation and Development (2007) "OECD Principles and Guidelines for Access to Research Data from Public Funding" http://www.oecd.org/dataoecd/9/61/38500813.pdf
Research Councils UK (2011) "RCUK Common Principles on Data Policy" http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
The Engineering and Physical Sciences Research Council (2011) "EPSRC Policy Framework on Research Data" http://www.epsrc.ac.uk/about/standards/researchdata/Pages/impact.aspx
UK Data Archive (2010) "Data Management Practices in the Social Sciences" http://www.data-archive.ac.uk/media/203597/datamanagement_socialsciences.pdf pp.17-21
UK Data Archive (2011) "Create and Manage Data" http://www.data-archive.ac.uk/create-manage
University of California-San Diego (2010) "Digital Library Program" http://libraries.ucsd.edu/about/digital-library/index.html
University of Edinburgh (2010) "Research Data Management Policy" http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy
University of Hertfordshire (2011), "Data Management Policy" http://sitem.herts.ac.uk/secreg/upr/IM12.htm
University of Illinois at Urbana-Champaign (2005) "IDEALS" http://www.ideals.illinois.edu/
University of Rochester (2008) "UR Research" https://urresearch.rochester.edu/home.action

Notes

1. GESIS-Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany, Tel. +49-221-47694 494. Email: laurence.horton@gesis.org and alexia.katsanidou@gesis.org
2. All online literature references available 29 March 2012.
3. A persuasive case for qualitative data reuse is made by Bishop (2009)

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Service and Technology** (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: *1 year, on: July 1st*
Automatic renewal: *no*

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page