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).

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

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).

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).

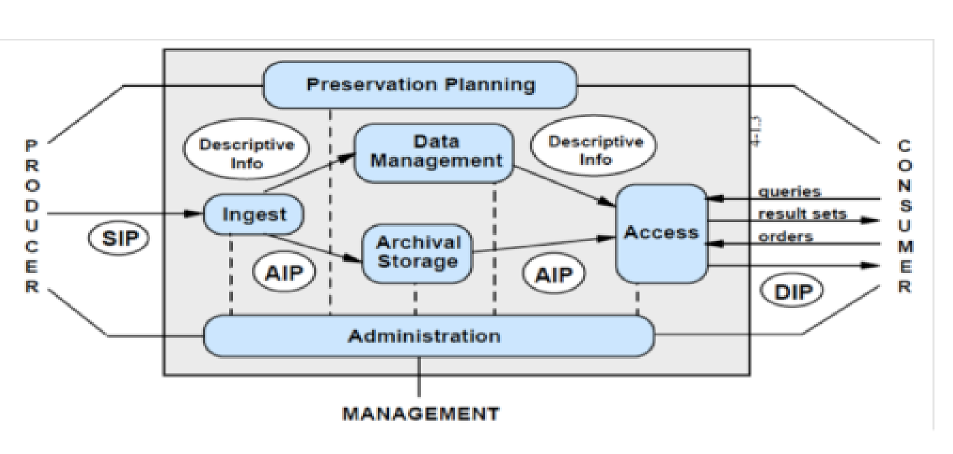

Figure 1. OAIS Functional Entities (CCSDS, 2002, p. 4-1)

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.socsciresearch.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