DATA CURATION IN THE WORLD DATA SYSTEM: PROPOSED FRAMEWORK

P Laughton *** and T du Plessis***

Department of Information and Knowledge Management, University of Johannesburg, Cnr University and Kingsway Road, Auckland Park, 2000, South Africa
**Email:* paull@uj.ac.za
Email: tduplessis@uj.ac.za

ABSTRACT

The value of data in society is increasing rapidly. Organisations that work with data should have standard practices in place to ensure successful curation of data. The World Data System (WDS) consists of a number of data centres responsible for curating research data sets for the scientific community. The WDS has no formal data curation framework or model in place to act as a guideline for member data centres. The objective of this research was to develop a framework for the curation of data in the WDS. A multiple-case case study was conducted. Interviews were used to gather qualitative data and analysis of the data, which led to the development of this framework. The proposed framework is largely based on the Open Archival Information System (OAIS) functional model and caters for the curation of both analogue and digital data.

Keywords: Data curation, World Data System, Framework

1 INTRODUCTION

Society is becoming more and more dependent on data. Palmer (2006) credits Clive Humby for the saying that "data is the new oil". Unlike oil, data are abundant, but like oil, they are a very valuable resource. Data are a very useful resource too as through analysis, patterns emerge, leading to understanding. The understanding from data analysis often leads to new discoveries. Although data are such a valuable and useful resource, they are also a problematic resource.

Hey and Hey (2006) identify some of the problems with data in the 'data deluge'. The data deluge is a state where society is overwhelmed by the amount of data generated. The data deluge highlights the need for effective management of data as they are a valuable but difficult resource to manage. As with any non-renewable resource, the more effectively data are managed, the more value they can generate. New practices have been developed to enhance data management, such as data curation.

Data curation can be defined as the "activity of managing and promoting the use of data from its point of creation, to ensure it is fit for discovery and reuse" (Lord & McDonald, 2003, p. 12). Data curators are responsible for managing data stores; their role as data professionals includes the management of data, adding of value to data, data sharing, and data preservation for later use (Rusbridge, 2008). Data curators have an important role to play, especially in environments such as data centres where large quantities of data are dealt with.

The World Data System (WDS) is an example of an environment that relies heavily on the curation of data. The WDS consists of open access data centres that curate large quantities of research data. Many data centres have a unique approach to data curation, resulting in the use of a number of different models and frameworks for the curation of data. The objective of this paper is to determine whether it is possible to develop a standard framework for the curation of data within the WDS. To determine this, qualitative data were collected from online interviews conducted in research data centres that are now incorporated into the WDS.

2 DATA CURATION MODELS AND FRAMEWORKS

There are many different data curation models and frameworks used for the curation of data. Some of these models and frameworks are used by numerous organizations while others are used in unique environments based on the needs of a particular organization. Although there are numerous frameworks and models, there is only one International Organization for Standardization (ISO) approved data curation model. This ISO approved model is the Open Archival Information System (OAIS) Reference Model (ISO 14721, 2003).

Other data curation frameworks and models have been based on the OAIS Reference Model, such as the Tsinghua Digital Preservation Platform (THDP) (Ma, Li, Jiang, & Xing, 2008) and the JISC modified OAIS functional model (Beagrie, 2004). One of the popular data curation models that shares similarities with the OAIS functional model is the Digital Curation Centre (DCC) Lifecycle model (see Figure 2). There are, however, many data curation models and frameworks that are not based on the OAIS Reference Model but ultimately strive to ensure long-term sustainability of data.

The OAIS Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS), published in 1999 and later updated in 2001. The 2002 version of the OAIS Reference Model was accepted in 2003 as an international standard (Beedham, Missen, Palmer, & Ruusalepp, 2005, p. 6). The OAIS Reference Model consists of two models: the information model and the functional model. The information model addresses the information objects and metadata used to preserve and access items in an archive. The functional model defines the six functions that are necessary for data curation. This research focuses on the functional model.

The OAIS functional model (see Figure 1) begins with the Ingest function, which allows data to be accepted as Submission Information Packages (SIPs) and prepares data for storage and management within the archive. Following the Ingest function is the Archival Storage function, where SIPs are converted into Archival Information Packages (AIPs), which are necessary for the storage, maintenance, and retrieval of these data. The Data Management function within the OAIS functional model entails the service and function for populating, maintaining, and accessing Descriptive Information (DI) and administrative data used to manage the archive (see Figure 1). The Administration function provides services and functions to the overall operation of the archive. The Preservation Planning function is responsible for monitoring the environment to ensure data remain accessible to the user community through the interaction with producers and consumers. The final function is that of Access and requires supporting the consumers in determining the existence, description, location, and availability of data within the archive (CCSDS, 2002).

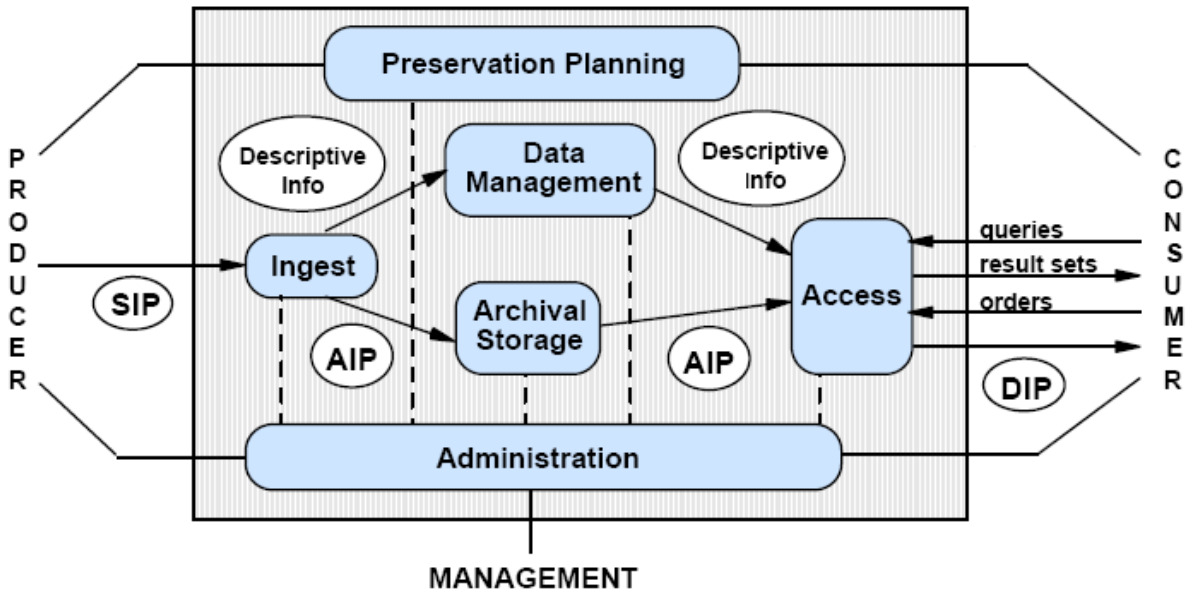

Figure 1. OAIS functional model (McCory, Connell, & Black, 2008, p. 2)

The OAIS functional model (Figure 1) is a well-designed approach that takes many factors into consideration. This comprehensive approach to data curation does however have shortcomings, expressed by critical reviewers and summarized in this section. Some of these critical viewpoints may not apply to all archives or data storage facilities using the OAIS functional model. Firstly, there is no functional entity assigned to the phase prior to Ingest. A Pre Ingest phase is crucial and the need for this function is highlighted by Karasti, Baker, and Halkola (2006, p. 344). The methods used for data collection could be addressed in the Pre Ingest function and have an impact on the development of metadata and the storing of these data. Data can be collected automatically or manually. Manual data collection affords freedom and flexibility, which should be addressed in Pre Ingest. Beedham et al. (2005) believe that the OAIS reference model would be more effective if it included the Pre Ingest phase rather than referring to this in a separate model. Pre Ingest is very important because it ensures quality, understanding, and accessibility to data. Nicholson and Dobreva (2009) suggest defining details (metadata) are needed in the grey areas prior to Ingest (production) and after Access (reuse). This critique of the OAIS functional model suggests a need for a framework that goes beyond the OAIS and addresses implementation issues regarding the specification of minimum requirements of policies, process, and metadata. OAIS fails to specify specific interfaces or protocols to support design and implementation of entities responsible for encoding and processing.

Despite receiving valid criticism, the OAIS functional model has effectively focused the attention of the e-science community on the development of a globally accepted standard for the curation of data. The OAIS functional model was systematically developed with good effect and through a universal approach though it may not be a best fit for all data repositories or archives.

The DCC Lifecycle model is comprised of full lifecycle actions, sequential actions, and occasional actions. Some of the lifecycle, occasional, and sequential actions of this model, such as community watch, conceptualize, appraise and select, and participation, can be seen as action that fall outside of the OAIS functional model and can be likened to a Pre Ingest function, which the OAIS functional model does not cater for. Furthermore, there are a number of actions in the DCC Lifecycle model similar to the functions of the OAIS functional model, such as Preservation Planning, Preservation Action, Ingest and Action, and Use and Reuse.

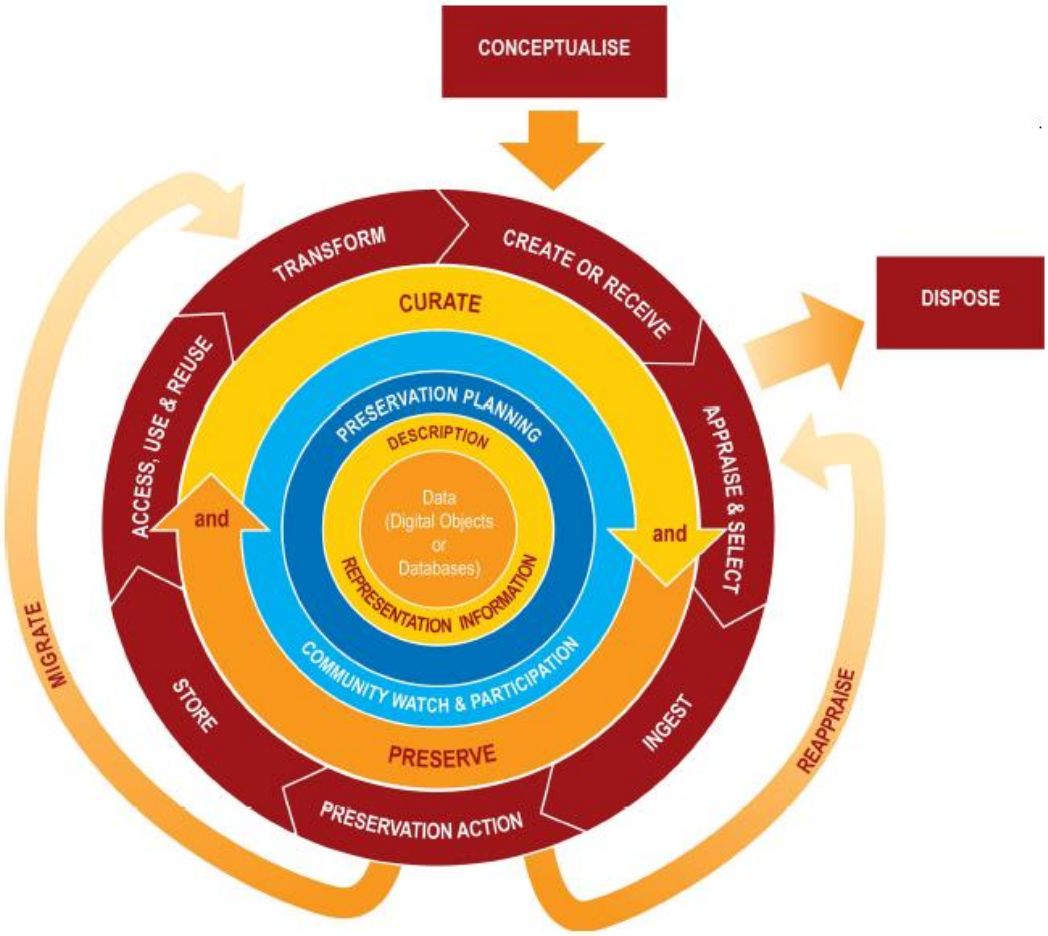

Figure 2. The Digital Curation Centre Lifecycle model (Higgins, 2008, p. 136)

The OAIS functional model and the DCC Lifecycle model are two popularly adopted ways of curating data (see Figure 1 and Figure 2). This functional model is part of the comprehensive OAIS Reference Model. All data curation models and frameworks strive to ensure that the data under current curatorship are available for future use. Models and frameworks (such as the OAIS functional model) are adopted by data centres and archives, such as the WDS, as a strategic plan on how to address the curation of data.

3 WORLD DATA SYSTEM

The WDS is a newly formed International Council for Science (ICSU) interdisciplinary body created to replace two outgoing organizations, the World Data Centre (WDC) and the Federation of Astronomical Geophysical Data Analysis and Services (FAGS). The WDS was officially launched at the 29th ICSU General Assembly in 2008, but applications to join the WDS were only opened later in February 2011. The WDS is planned for an initial period of 10 years; thereafter it will be reviewed to determine its effectiveness (ICSU, 2012; Rickards, 2011).

The WDS was formed because of a perceived inability of the WDC and FAGS to meet the needs of the scientific community and the insufficient flexibility of their interdisciplinary studies. The WDS is designed to co-ordinate a global approach to scientific data by planning to guarantee universal access to data for education and informed decision-making. Challenges for the WDS include the unification of scientific data transmission formats, information protocols, and data curation standards (Zgurovsky, Gvishiani, Yefremov, & Pasichny, 2010, p. 211; Rickards, 2011).

The WDS attempts to transform the separate member data centres of the WDC and FAGS into a globally interoperable and distributed data system incorporating emerging technologies and new scientific data activities. According to ICSU (2012), the goals of the WDS strive to:

- Enable universal and equitable access to scientific data, data services, products, and information
- Ensure long-term data stewardship
- Foster compliance to agreed upon data standards and conventions
- Provide mechanisms for ease of use and improved access to data
- Ensure provision of quality assured data, data products, and information

The WDS has an open access policy and gathers scientific data sets from different communities of interest, allowing the data to be accessed by others and managed appropriately. The member data centres play an important role in the scientific community by allowing scientists access to research data where they may not have had the funds or capacity to conducts the research to generate such data sets.

4 METHODOLOGY AND RESEARCH DESIGN

A multiple-case case study was conducted to determine whether it was possible to develop a data curation framework for the WDS. An interview approach was used to gather data for the multiple-case case study, which was conducted online via email. The WDC was selected as the population from which the sample was drawn as the research was conducted prior to the approval of any member data centres for the WDS.

The four cases selected for this qualitative inquiry were identified using a maximum variation sampling technique based on a quantitative score. When selecting a small sample of significant diversity, high quality, detailed descriptions of each case and a shared pattern that cuts across cases can be identified (Patton, 2002, p. 235). Qualitative inquiry usually focuses on small samples selected purposefully, with a focus on data quality to uncover multiple realities. According to Crabtree and Miller (1992, p. 233) "sample size is not the determinant of research significance in qualitative studies, the major concern is with information richness".

Maximum variation sampling allows a researcher to study variation in a phenomenon while focusing on a wide range of characteristics to build a conceptual understanding. Maximum variation sampling helps to identify emerging themes amongst different participants (Crabtree & Miller, 1992, p. 37; Lindof & Taylor, 2002, p. 123). This method of sampling allows researchers to address variability characteristics of random selection. It applies the logic that "any common patterns may emerge from great variations of a particular interest and value in capturing core experiences and central, shared dimensions of logic" (Patton, 2002, p. 235). A sample of significant diversity selected through the maximum variation sampling technique is popular in qualitative research.

All 52 WDC member data centres were asked to complete an online survey, resulting in 26 complete surveys. The maximum variation sampling technique was applied to the scores obtained from the survey, which generated an OAIS functional model conformance score[1] (based on 36 closed ended questions from which scores were allocated). The two lowest scores were 27 and 30 (out of a maximum of 92) while the two highest scoring cases scored 80 and 90 (out of a maximum of 92). These four cases were selected for the online interviews.

To conduct the interviews, an interview schedule was sent to data curation experts at each of the cases. The questions were developed to gain insight into how each case curated data. The questions were derived from the

[1] More detailed information on the OAIS functional model conformance score is available from the following publication: Laughton, P. (2012) OAIS functional model conformance test: a proposed measurement. *Program: Electronic Library and Information Systems*, 46(3).

CCSDS (2002) recommendations on the OAIS functional model and dealt with a range of issues around the curation of data. Once the interview questions were returned, the answers were analyzed. Following the analysis, follow-up questions were sent to the data curation experts to clarify concepts that were not understood or to find out more on a particular aspect. The transcripts generated from the online email interviews were used to generate a framework.

The interview process took place over a period of 3 months. Prior to conducting the interviews, permission was granted and a letter of consent was accepted. The cases' identities were kept anonymous in an attempt not to disclose information that might negatively impact the data centre.

The reported data curation practices for each of the data centres were analyzed according to the OAIS functional model. The model was used as a basis for comparison as it is currently the only ISO approved standard for curating data. Some of the data centres referred to procedures outlined in policies and in documentation. These policies and documents were also analyzed and included in the relevant case summaries. Variations to the OAIS functional model were also recorded and were included in the case analysis.

5 FINDINGS

To put the selected cases into perspective, some characteristics of the cases based on the data in the online survey are conveyed. The data displayed in Table 1 show the physical storage space of the data centres and the number of employees as well as whether or not each data centre believed they had the capacity (in terms of staff, training, storage, hardware, and software) to effectively curate the data within their holdings. Three of the four WDC member data centres were of a similar size, smaller than 50 terabytes (TB), while one had a storage size of more than 3 petabytes (PB). The two highest scoring data centres in the OAIS functional model conformance test had a greater number of employees in comparison to the two lowest scoring data centres. Only one of the data centres believed they did not have the capacity to effectively curate the data.

Table 1. Quantitative data summary of cases selected for case study

	OAIS functional model conformance score	**Size**	**Number of employees**	**Have capacity required?**
Case 1	90	Smaller than 50 TB	More than 10	Yes
Case 2	80	More than 3 PB	More than 10	Yes
Case 3	35	Smaller than 50 TB	Two or less	No
Case 4	27	Smaller than 50 TB	Two or less	Yes

From the interview transcripts, case summaries were developed. Each case summary was incorporated into the proposed framework, taking into account a number of considerations. Each case's data holdings influenced the way in which they curated their data. Below are the case summaries.

5.1 Case 1 summary

Case 1 scored the highest in the OAIS functional model conformance test and further qualitative analysis reveals a similar picture. Almost all the data stored in this data centre were digital, with a few analogue records that consisted of supplementary documents and signed permissions. The analogue data were used predominantly for administrative purposes, and consumers or end-users would not be required to access this data.

Data curation at this data centre starts out with a Pre Ingest function (prior to Ingest) where a number of reviews are conducted to evaluate the data that are coming in for storage. These reviews include an internal scientific review, a literature review, a community review, a data preparation review, and a deposit review. Following the Pre Ingest function, if data pass or receive favorable reviews they will be ingested into the data centre. For the Ingest function, documentation from the Pre Ingest reviews is submitted along with the data. An AIP is created and data are assigned. The data are further reviewed by a User Working Group before they are sent to the next function, namely Archival Storage. The AIP is received, and it is added to permanent storage. These data are duplicated, creating security copies, which are sent to an offsite location for back-up.

The Data Management function for Case 1 is carried out through the testing and improvement of alpha and beta reviews of new products and services proposed. The alpha reviews are conducted by internal staff, and the beta reviews are conducted by volunteers from the user community. The findings from these reviews, along with the relevant documentation, are reviewed by the Configuration Management Board, allowing a final decision to be made on whether or not to adopt new product(s) or practice(s). This test bed is designed to improve the experience for those internally at this data centre as well as the end consumers.

The Preservation Planning function at Case 1 makes up a large part of the data curation activities. Systems and forms of media are routinely reviewed and upgraded, based on community reviews of technology and standards. The internal staff at this data centre work closely with scientific communities from various disciplines through the attendance of conferences, workshops, and scientific meetings to ensure they are familiar with what is happening in the respective communities. New data sets intended for storage are reviewed and scrutinized to ensure consideration for current and future data curation practices.

At Case 1 the Administration function is conducted by checking that a security copy of the data has been created and is stored in an offsite location as part of the disaster and recovery planning. The Access function is regulated through an online website and is available to the public.

5.2 Case 2 summary

For Case 2 a combination of digital and analogue data is stored in the data centre. The older analogue data date back to the 1930s while the digital data have only been received since the 1970s. The stored data consist of mostly images and raw data. This data centre scored well in the OAIS functional model conformance test.

Case 2 has a well defined appraisal process, which is used for Pre Ingest. Of all the cases that were interviewed, Case 2 had the most comprehensive processes set up for the Pre Ingest function. For data to be considered for storage, a number of reviews of the data need to be conducted. A team of scientists is selected to review the data. Following this scientific review, an archivist is then required to document the data, using an in-house online appraisal tool. This archivist later briefs a project manager on the data and supporting metadata. From these reviews, recommendations are sent to senior managers who make a decision on whether to accept the data or not, based on the feedback obtained. During this review process, cost and potential impact or benefits, which have a strong influence on the decisions made by senior management, are assessed.

The Ingest function follows, and the accepted data are taken into the data centre where preparation for storage begins. Agreements between the data centre and the producer of the data are established. Some data, such as metadata population, that are accepted may require little modification while other data may require media migration before storage.

The ingested data are ready for storage. The Archival Storage function is divided into two parts, digital and analogue. Digital data are added to permanent storage; back-ups and duplications are created. Case 2 has a three copy policy, in which a primary copy is made accessible online, a secondary copy is stored onsite (off line), and a third copy is stored off site. A total of 50% of all their digital data adhere to this three copy policy. Analogue data are stored in a location that is physically and environmentally controlled to prolong the life of this data.

The Preservation Planning function is conducted through a range of activities and tasks. Case 2 works closely with organizations to ensure up to date practices are maintained. The preservation policy is regularly reviewed. A migration goal of every three to five years is maintained to prevent any digital data from becoming so outdated that it may not be possible to migrate them to a newer format. This migration includes the migration of hardware, software, firmware, and media obsolescence. In accordance with this goal, a trade study is conducted on a regular basis to determine technology that can improve storage and the curation of data, based on the future needs of the user. Plans are also in place to migrate older analogue data to a digital format through a scanning process.

The Administration function is addressed by ensuring that an offsite back-up is created as part of the disaster recovering planning. Their Customer Service unit deals with client (producer and consumer) queries and requests as part of the Administration function.

At Case 2 the Access function is managed through a website that is made available to the public, ensuring free access to data. The respondent did not address the topic of access to analogue data as only digital data can be made available through the website.

5.3 Case 3 summary

Case 3 stores a combination of digital and analogue data. Most of the data stored at this data centre are images and raw data. The majority of the images received are in analogue format. For data to be accepted for inclusion in this data centre, there is no formal policy, and data that are relevant to the subject field are provisionally accepted, which differs from Case 1 and Case 2 that have well establish Pre Ingest processes. The data sets' Submission Information Packages (SIPs) are prepared for storage. All the necessary metadata are added before they are ready for long-term storage.

Once the data are ready and all associated metadata are attached, they are sent to the Archival Storage function as an AIP for digital storage. A copy of the digital data is made for safekeeping. The analogue data that are received as paper records are kept onsite in a marked location for later retrieval.

As some of the data are not made available to consumers through an online website, access requests for these data are attended to by the Data Management function. The data are sourced under this function so they can be transferred to the requester.

The Preservation Planning function takes place as necessary content is identified for later migration. Case 3 is constantly looking out for new technology and procedures in an attempt to stay up to date and informed. There are projects currently running that convert analogue data to digital data. Part of this is the conversion of paper records to PDF and converting microfilm to Digital Video Disk (DVD) or Hard Drive Disk (HDD). Case 3 makes an effort to keep informed and is attempting to upgrade the infrastructure to keep up with the demand.

The Access function in Case 3 is dealt with separately for digital and analogue data. The Access function for digital data makes them available through either a self-service online website (this is only for some data), and other digital data are made available through a File Transfer Protocol (FTP) file server. However, the Access function for analogue data requires requests for data to be posted to the requester.

5.4 Case 4 summary

For this particular case, one of the defining characteristics is the need for a curation process that works for both digital and analogue data simultaneously. The data curator at Case 4 indicated that a considerable amount of analogue data was under curatorship at this data centre. The data gathered from the interviews on Case 4 indicate a duplication of functions for digital and analogue data. The data curator on more than one occasion stated that analogue data and digital data were dealt with differently at this data centre. As with Case 3, there was no mention of any formal Pre Ingest function when it came to determining whether the data were fit for curation at this particular data centre.

During the Ingest function for digital data, following the acceptance of the data, digital data are checked multiple times by an automated computer program to determine fitness for storage. The accepted digital data are labeled as preliminary data, and only after they has been checked for accuracy (a different check to the automated computer program mentioned earlier), are these data classified as definitive. The process to gain definitive status can take up to one year after acceptance for storage, but during this time the data are made available as a preliminary version.

Upon completion of the Ingest function, data (both digital and analogue) are ready for the Archival Storage function. Digital data are assigned the necessary metadata, stored, and made accessible through a Database Management System (DBMS). These digital data are checked regularly to ensure accessibility to the public; this is carried out both manually and digitally through computer programmes. The analogue data are stored in the archive and are catalogued, giving them an identification number and physical location to assist in future retrieval.

The Data Management function follows the Archival Storage function, which seems only applicable to the analogue data. Queries for analogue data are processed. Once a query is processed, the data will either be prepared for the Access function or requesters will be informed of where they could possibly find the digital version at another data centre if they are available in this format.

The Preservation Planning function occurs for both the digital and analogue data. For the digital data, Preservation Planning is used to stay up to date with known and accepted formats and conventions while for analogue data, the physical data are stored in a secure archive that is ordered and catalogued.

The Administration function occurs for both the digital and analogue data. For the digital data, back-ups are made and stored at other data centres for safekeeping, forming part of their disaster and recovery planning. For both the digital and analogue data, submissions to this data centre are negotiated by the director of the data centre.

Finally, the Access function is duplicated for digital and analogue data. Digital data are accessible via an online website, and users can use a self-service menu to search and download the data they require from a FTP file server. For analogue data, requests are indicated usually through email, and a copy will be made, or requesters will be informed at which data centre they may find the digital version of this data (saving time and money for the distribution of the data).

5.5 Summary of the case findings

The cases that scored higher in the OAIS conformance test (Case 1 & Case 2) indicated that they had less analogue data to curate compared to the lower scoring cases (Case 3 & Case 4). The functions involved in the data curation of both digital and analogue data work differently. This highlights a need for digital and analogue data to be addressed separately in a framework or model. Each function for digital and analogue data should

work independently as they require different resources and care. Currently the OAIS functional model does not cater effectively for analogue data.

Case 1 and Case 2 were more in line with the CCSDS (2002) recommendations and guidelines than Case 3 and Case 4. This correlates with the findings from the OAIS functional model conformance test, indicating the possible effectiveness of this test.

Case 1 and Case 2 indicated they had a Pre Ingest function. This allows for better screening of the data before they are sent to the Ingest function. Case 1 and Case 2 have policies in place to regulate the data considered for long-term storage. The Pre Ingest function is not part of the OAIS functional model and should be included in the proposed framework for the WDS. There are a number of scholars who believe that the Pre Ingest function should be incorporated into the OAIS functional model (Beedham et al., 2005; Karasti, et al., 2006, p. 344; Nicholson & Dobreva, 2009) (see Section 2).

6 PROPOSED FRAMEWORK FOR THE CURATION OF DATA IN THE WORLD DATA SYSTEM

Based on the analysis of findings from the multiple-case case study, a proposed framework was developed. This framework is largely based on the OAIS functional model. The best practices from each case have been incorporated into this framework in an attempt to cater for a range of data curation needs. The proposed framework is divided into two sections, digital and analogue. Each function is duplicated under these sections (see Figure 3). Duplication of these functions is deemed necessary as functions may be the same; however the processes within the functions differ, and the framework caters for these differences in process. In Case 2 and Case 4, duplication was apparent where certain functions handled analogue and digital data differently.

The proposed framework for the curation of data within the WDS begins with a Pre Ingest function (see Figure 3). This function is responsible for reviewing data to ensure they are suitable and fitting to the research area(s) of the data centre. On the analogue side of the framework, data are reviewed to determine whether it is possible to migrate them to a digital format in the near future and what the potential cost of such a migration might be. On the digital side of the framework, data centres need to determine whether they have the technology and the capacity to curate the data.

The Ingest function follows the Pre Ingest function. Data that pass the reviews in the Pre Ingest function are accepted for storage. Agreements need to be made and confirmed between the producer or owner of these data and the data centre. One the analogue side of the framework, data are readied for Archival Storage, and the relevant metadata are captured and catalogued. The digital side of the framework SIPs, which later become AIPs and are assigned a unique number, are prepared for storage.

The Archival Storage function keeps all the data onsite in a secure location. Onsite physical and environmental monitoring is conducted to ensure the best possible conditions for preservation. The analogue data should be continually checked to ensure they are in the correct physical location. Digital data need to be backed-up with a minimum of two copies, one onsite and one offsite. Digital data need to be continuously checked for data corruption and prepared for online access.

The Data Management function deals with requests from consumers and those looking to access data. Once the data are found, the results are sent. This function is responsible for recording access requests for reporting purposes. The Data Management function is also responsible for receiving updates on deletions, modifications, or additions to the data in storage. On the analogue side of this framework, requests for data are readied for the consumer (either a digitized or a physical copy is made). On the digital side of the framework, this function is responsible for reviewing new products (hardware and software) and new technology prior to implementation.

The Administration function is responsible for conducting negotiations with producers or owners of the data. Inventory reports need to be created and handed to the Preservation Planning function as a responsibility of this function. The Administration function also deals with consumer inquiries or complaints and undertakes a customer service role. The Administration functions for analogue and digital data are almost identical while reporting for the Preservation Planning function is only necessary on the digital side of the framework.

The Preservation Planning function plays a vital role in the migration of both analogue and digital data (digitization of analogue data and the migration of digital data to new formats) while remaining knowledgeable of new technologies that can assist with such migrations. It is important that representatives of the data centres keep informed of what is happening in the relevant communities by attending conferences, symposiums, workshops, and scientific meetings. This function is also responsible for reviewing the Preservation Policy of the data centre. On the analogue side of the framework, the scrutinizing of the proposed migration of analogue data to digital data formats takes place. Communication with other data centres that have undertaken similar migration projects is important. The development and testing of migration prototypes is conducted under this function. On the digital side of the framework, new data that are being considered for storage are scrutinized to ensure consideration for future and current curation practices. The development and finalization of migration plans are entrusted to the Preservation Planning function.

The Access function for analogue data is considerably different from that for digital data. Requests for analogue data are placed by consumers, either automated through an online catalogue or through an email. A copy of the analogue data is made in the relevant format and placed in a staging area ready to be sent. With some requests, analogue data may be migrated to a digital format so they may be accessible through a website; this is dependent on the data centre and the complexity of such a migration. Digital data, on the other hand, are made accessible through a self-service website.

This proposed framework for the curation of data in WDS data centres is summarized in Figure 3. This framework has a two sided approach that deals with analogue and digital data differently and promotes the effective migration of analogue data to digital data to enhance future use.

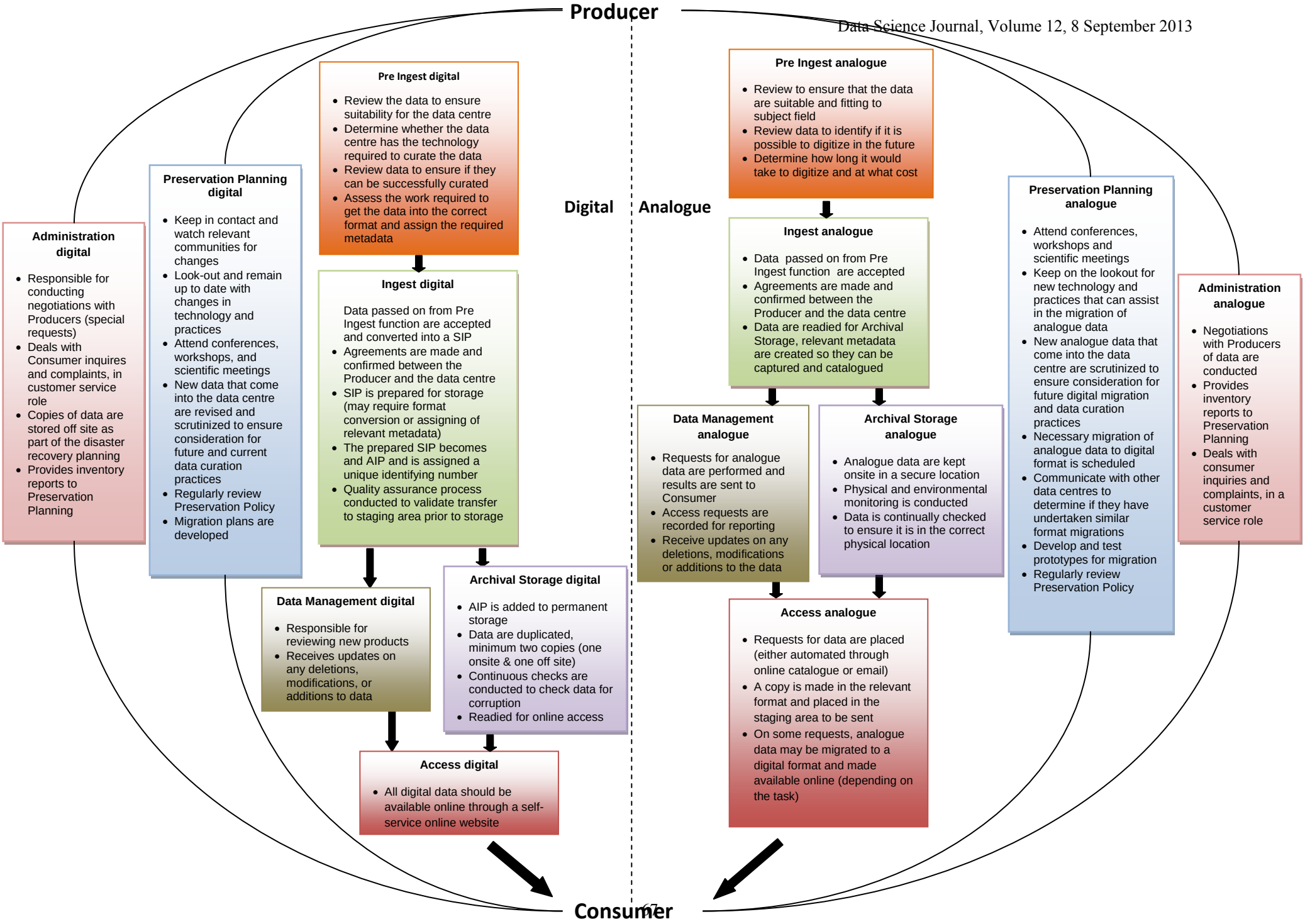

Figure 3. The proposed framework for the curation of data in the WDS

7 LIMITATIONS

Analysis of the data collected has led to the development of a proposed framework for the curation of data in the WDS. As with any research, it is not without limitations. Limitations are largely derived from the research design. Each design has its own set of challenges. Below is a list of the limitations of this research:

- The proposed framework is based on the OAIS functional model and does not take other models into consideration. This framework was developed in such a manner due to the ISO accreditation the OAIS functional model achieved, in an attempt to comply with these standards.
- The population from which the sample was selected was the WDC and not the WDS. The reason for this was at the time of conducting the research the WDS had no members. The WDC was selected, but the member data centres of FAGS were not included in the research. The proposed framework could act as a recommendation or guidelines for data centres looking to join the WDS.
- From the population of the WDC, only four cases were selected for the multiple-case case study. More cases could give more insightful data and analysis while possibly affecting the design of the proposed framework.
- The maximum variation sampling method used attempted to gather data on a broad spectrum of practices by selecting the two highest and two lowest scoring data centres in the OAIS functional model conformance test. The sampling method does, however, not investigate those with scores between these two extremes.
- Online interviews were conducted as opposed to face to face interviews due to the expansive distance between the physical locations of the four cases. Face to face interviews may have allowed for a more natural flow of communication, and respondents may have placed a higher level of trust in the interviewer, resulting in more descriptive answers to the questions.

These limitations have been considered by the researcher. Although the research design that was used may have had some limitations, these were outweighed by the inherent advantages.

8 CONCLUSION

A standard framework for the curation of data in the WDS was developed (see Figure 3). This framework is largely based on the OAIS functional model while changes were made to suit the WDS requirements. The proposed framework has seven functions (Pre Ingest, Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access). Each of these functions is duplicated for digital and analogue data while the processes within each function differ for digital and analogue data.

The analysis of the qualitative data gathered from the interviews reveals an inability of the OAIS functional model to sufficiently address the curation of analogue data. The proposed framework addresses this by duplicating each function for digital and analogue data and adjusting the recommendations to cater for each medium. Data centres and archives that work with large amounts of analogue data may find it difficult to follow the recommendations in the OAIS functional model, and a framework such as this proposed one may be an alternative.

The proposed framework should be tested in the WDS to see how effectively it allows data centres to curate data. Feedback from the implementation may allow the framework to be amended to better serve the WDS member data centres.

Data, like oil, are becoming more and more important resources in society, but unlike oil, data are abundant. This abundance poses problems. Proper management is necessary to ensure the sustainability of the data. There are many challenges that data professionals face; however, the need for a well structured approach to data curation is apparent.

9 REFERENCES

Beedham, H., Missen, J., Palmer, M., & Ruusalepp, R. (2005) *Assessment of the UKDA and TNA compliance with OAIS and METS standards*. Retrieved in February 17, 2012 from the World Wide Web: http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf

Beagrie, N. (2004) The continuing access and digital preservation strategy for UK Joint Information Systems Committee (JISC). *D-Lib Magazine, 10(7/8)*. Retrieved in February 17, 2012 from the World Wide Web: http://www.dlib.org/dlib/july04/beagrie/07beagrie.html

CCSDS (2002) *Reference model for an Open Archival Information System*. Retrieved in February 17, 2012 from the World Wide Web: http://public.ccsds.org/publications/archive/650x0b1.pdf

Crabtree, B.F. & Miller, W.L. (Eds) (1992) *Doing qualitative research: research methods for primary care*. Volume 3. Newbury Park: Sage Publications.

Hey, T. & Hey, J. (2006) e-Science and its implications for the library community. *Library Hi Tech 24*(4), pp 515-528.

Higgins, S. (2008) DCC curation lifecycle model. *The International Journal of Digital Curation 1*(3),pp 134-140.

ICSU (2012) Constitution of the International Council for Science World Data System. Retrieved in August 29, 2013from the World Wide Web: http://icsu-wds.org/images/files/WDS_Constitution_04_04_12.pdf

Karasti, H., Baker, K.S., & Halkola, E. (2006) Enriching the notion of data curation in e-science: data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work*,*15*(4), pp 321-358.

Lindlof, T.R. & Taylor, B.C. (2002) *Qualitative communication research methods* (second edition). Thousand Oaks: Sage Publications.

Lord, P. & McDonald, A. (2003) *e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*. Retrieved in February 17, 2012 from the World Wide Web: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Ma, N., Li, C. Jiang, A., & Xing, C. (2008) Design and implementation of open source based digital preservation experimental platform (THDP). *Conference proceedings of the 9th International Conference for Young Computing Societies*. Conducted by the China Computer Federation held in Hunan.

McCrory, A., Connell, T., & Black, B. (2008) *Detailed functional model of OAIS reference model: local implementation of functions*. Retrieved in May 25, 2009 from the World Wide Web: http://library.osu.edu/sites/dlib/OAIS_Report.pdf

Nicholson, D. & Dobreva, M. (2009) Beyond OAIS: towards a reliable and consistent digital preservation implementation framework. *Conference proceedings of the 16th International Conference on Digital Signal Processing*. Conducted by IEEE held in Santorini.

Palmer, M. (2006) *Data is the new oil*. Retrieved in February 16, 2012 from the World Wide Web: http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

Patton, M.Q. (2002) *Qualitative research and evaluation methods* (third edition). Thousand Oaks: Sage Publications.

Rickards, L. (2011) Developing the new ICSU World Data System (WDS). *Conference proceedings of the International Oceanographic Data and Information 50th Anniversary International Conference*. Conducted by the IODE and held in Liege, Belgium.

Rusbridge, C. (2008) *What makes up data curation?* Retrieved in February 17, 2012 from the World Wide Web: http://digitalcuration.blogspot.com/2008/12/what-makes-up-data-curation.html

Zgurovsky, M.Z., Gvishiani, A.D., Yefremov, K.M., & Pasichny, A.M. (2010) Integration of the Ukrainian Science into the World Data System. *Cybernetics and Systems Analysis 46*(2), pp 211-219.

(Article history: Received 3 June 2013, Accepted 19 August 2012, Available online 1 September 2013)