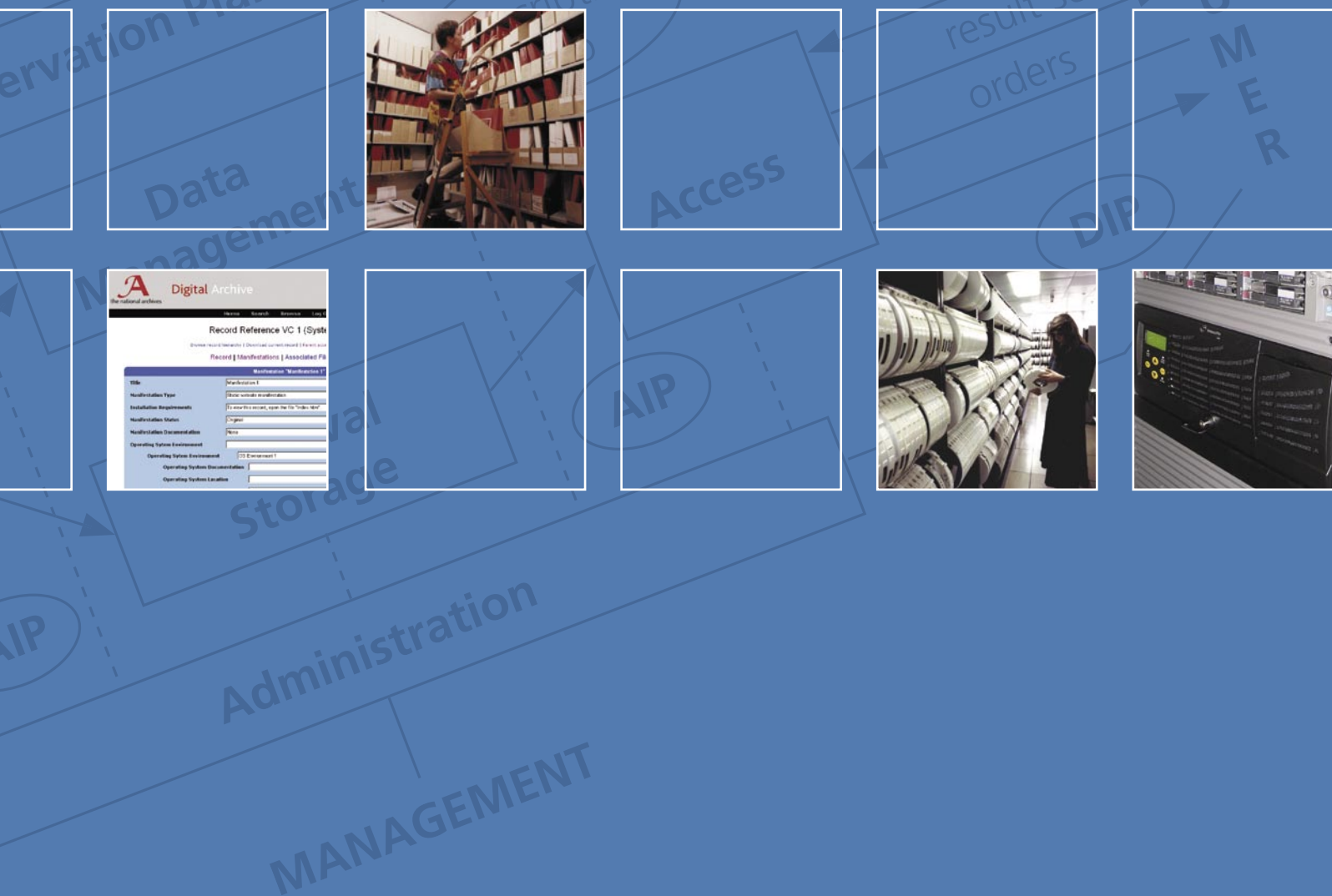




## ASSESSMENT OF UKDA AND TNA COMPLIANCE WITH OAIS AND METS STANDARDS

Hilary Beedham, Julie Missen (UK Data Archive)  
 Matt Palmer (The National Archives)  
 Raivo Ruusalepp (Estonian Business Archives Ltd.)



## EXECUTIVE SUMMARY

---

This report has been produced with funding from the JISC Institutional Digital Preservation and Asset Management Programme 2004 and contributions from The National Archives (TNA) and the UK Data Archive (UKDA) at the University of Essex.

The main focus of the work has been to compare the systems and processes that are currently in place at the UKDA and TNA with the OAIS reference model. Each of these organisations has a responsibility for the preservation and dissemination of electronic records of national importance but each implements systems which predate the development of the model. Consequently, the aims of the project were to determine whether each organisation is compliant with the model and to present the work in a report that can be used by other, similar, organisations to aid their own exploration into their compliance with the reference model.

The project report is structured in the following way. Chapters one to three provide background information and explore the compliance issues with the OAIS model and other relevant standards for archives. Chapters three and four describe the methodology applied and discuss how TNA and the UKDA comply with the OAIS mandatory requirements. Chapters five and six examine how the archival systems at these two institutions match with the OAIS functional entities and informational model and chapter seven considers the METS metadata standard, how it could be used further in a digital archive and its potential role in each archive.

Our aim has been to provide a 'use case' for organisations wishing to test their compliance with the model and to this end, the conclusions presented in chapter eight include observations and recommendations regarding the OAIS compliance testing. In addition we have included several appendices which, coupled with the detail of the report itself, will, we hope both encourage and facilitate the process of compliance testing. In particular, Appendix 1 contains a glossary of terms used in relation to the OAIS reference model and the METS standard, plus in-house terms referred to in the report whilst Appendix 5 offers a small set of questions for self-testing for OAIS standard compliance.

Kevin Schürer  
Director  
UK Data Archive

David Thomas  
Director of Government and Technology  
The National Archives



## ACKNOWLEDGEMENTS

---

The authors would like to thank the JISC for funding this work and colleagues at the UKDA and at TNA who have offered advice and comments on, and provided material for, this report.

Hilary Beedham > UK Data Archive

Julie Missen > UK Data Archive

Matt Palmer > The National Archives

Raivo Ruusalepp > Estonian Business Archives Ltd.



## TABLE OF CONTENTS

|     |                                                                               |     |
|-----|-------------------------------------------------------------------------------|-----|
| 1   | BACKGROUND .....                                                              | 4   |
| 1.1 | The Project and Participating Organisations .....                             | 4   |
| 1.2 | The Development of the OAIS Reference Model Standard .....                    | 6   |
| 1.3 | The Development of the Metadata Encoding and Transmission Standard .....      | 7   |
| 1.4 | Project Aims .....                                                            | 8   |
| 2   | WHAT DOES IT MEAN TO BE OAIS COMPLIANT? .....                                 | 9   |
| 3   | METHODOLOGY .....                                                             | 10  |
| 4   | COMPLIANCE WITH OAIS RESPONSIBILITIES .....                                   | 11  |
| 4.1 | Negotiates for and Accepts Information from Producers .....                   | 11  |
| 4.2 | Obtains Sufficient Control for Preservation .....                             | 15  |
| 4.3 | Determines Designated Consumer Community .....                                | 16  |
| 4.4 | Ensures Information is Independently Understandable .....                     | 18  |
| 4.5 | Follows Established Preservation Policies and Procedures .....                | 20  |
| 4.6 | Makes the Information Available .....                                         | 21  |
| 4.7 | Conclusions .....                                                             | 23  |
| 4.8 | Other Responsibilities and Standards .....                                    | 24  |
| 5   | OAIS FUNCTIONAL ENTITIES .....                                                | 26  |
| 5.1 | Ingest.....                                                                   | 26  |
| 5.2 | Archival Storage .....                                                        | 35  |
| 5.3 | Data Management .....                                                         | 39  |
| 5.4 | Administration .....                                                          | 42  |
| 5.5 | Preservation Planning .....                                                   | 49  |
| 5.6 | Access.....                                                                   | 54  |
| 5.7 | General Functionalities .....                                                 | 61  |
| 6   | OAIS INFORMATION MODEL .....                                                  | 62  |
| 7   | USING THE METS METADATA STANDARD IN A DIGITAL ARCHIVE .....                   | 71  |
| 7.1 | Introduction .....                                                            | 71  |
| 7.2 | The METS Document .....                                                       | 71  |
| 7.3 | Uses of METS .....                                                            | 73  |
| 7.4 | Strengths and Weaknesses of METS .....                                        | 77  |
| 7.5 | METS: Conclusions .....                                                       | 79  |
| 8   | FINAL CONCLUSIONS AND RECOMMENDATIONS .....                                   | 81  |
|     | BIBLIOGRAPHY .....                                                            | 85  |
|     | APPENDIX 1. GLOSSARY .....                                                    | 86  |
|     | APPENDIX 2. UKDA PREFERRED FILE FORMATS .....                                 | 89  |
|     | APPENDIX 3. TNA PRESERVATION METADATA MAPPING TO OAIS INFORMATION MODEL ..... | 90  |
|     | APPENDIX 4. EXTRACTS FROM TYPICAL UKDA READ AND NOTE FILES .....              | 101 |
|     | APPENDIX 5. A SET OF QUESTIONS FOR OAIS COMPLIANCE SELF-TESTING .....         | 105 |



## 1 BACKGROUND

---

### 1.1 THE PROJECT AND PARTICIPATING ORGANISATIONS

The participating organisations to this report are the UK Data Archive (UKDA) at the University of Essex and The National Archives (TNA), sited at Kew. Both organisations have long-established responsibilities for digital preservation of materials created on electronic media. Each organisation has developed similar but different systems for record keeping and the generation and storage of metadata relating to the files stored.

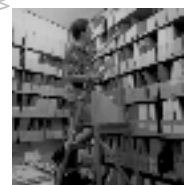
In late 2004, funding was made available from the JISC Institutional Digital Preservation and Asset Management Programme which supported projects, with a specific focus on strategies and procedures for long-term digital preservation and asset management. A wide range of activities are being undertaken within the programme, including institutional management support, the development of digital preservation assessment tools and institutional repository infrastructure. One particular area of interest was the role of standards for the development of reliable digital preservation and asset management procedures, in particular the potential of the OAIS Reference Model as a conceptual foundation for more focused work in digital preservation and to provide guidance for institutions in developing archival policy and procedures.

Thus the UKDA and TNA were able to test their compliance with both the reference model and the METS standard and, in so doing, were offered the opportunity to place a recognised 'seal of approval' on practices and systems that have evolved in response to an ever-changing technical environment. At the same time, it enabled the two organisations to compare their preservation practices within a common framework. This opportunity was particularly timely because, in January 2005, the UKDA was appointed as a legal place of deposit for TNA, meaning that the UKDA can now legally hold 'certain public records of strong local or specialist interest', and moreover will be recognised as the official repository or storage place for certain key government data. A further effect of this designation is that the UKDA and TNA now have common goals in relation to government datasets.

The specific aims of this short project were to map the systems and metadata currently in use by the two organisations against those in the OAIS Reference Model and the METS standard; to practically test the theoretical argument that the two partners comply with OAIS; to assess how the two institutions' operational structures can be informed by OAIS (and vice versa); and to explore the potential for interaction between existing metadata standards utilised by the two institutions and METS. It was expected that, as a result of this work, each organisation would be able to assess the relevance of the Reference Model and the metadata standard to their work and determine whether or not the assumption that each is compliant, is in fact correct.

A further aim was the production of this report outlining the experiences of each organisation in undertaking the mapping with the expectation that the record of the methodology and results would be of use to other organisations that might want to undertake a similar exercise.

In particular, both the UKDA and TNA have service functions as well as preservation functions. Consequently, a further aim is to contribute to the knowledge base necessary for the any future adaptation of the OAIS model to the JISC Information Environment (see section 7, JISC Continuing Access and Digital Preservation Strategy).



**The UK Data Archive** (UKDA) is now funded jointly by the University of Essex, the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC). It has been the primary repository for digitised social science research data in the UK since 1967. As a 'national data collection service' the UKDA, originally called the Data Bank, was created by the forward-thinking Social Science Research Council, now the Economic and Social Research Council (ESRC), to bring together 'social survey research materials for storage, retrieval and secondary analysis of the information in them'. For over three decades, preservation of these collections has been a core function of this enterprise. Over recent years the remit of the UKDA has been extended with the addition of new services such as the AHDS History Service, the Census Registration Service and an extensive programme of research and development projects in relevant areas such as multi-lingual thesaurus development, software for data publishing and browsing, research into the preservation and grid-enabling of social science data and collaborations with research councils, including the Medical Research Council (MRC) and the Natural Environment Research Council (NERC).

The UKDA continues to facilitate secondary analysis in the scholarly community by contributing to projects to produce teaching and learning material for students and by preserving and sharing research material that may have served its immediate purpose but has continuing value for re-use. By taking a strategic approach to long-term digital preservation, the UKDA ensures that it is at the leading edge of technical advances by monitoring hardware and software developments and migrating its collections accordingly. The UKDA is committed to using its resources wisely, and adding value to data collections where it will most benefit the user community.

Since January 2003, the UKDA has managed core activities and provided dedicated services for the ESRC under the banner of the Economic and Social Data Service (ESDS). The ESDS is dedicated to supporting users of social and economic datasets for secondary analysis for research and teaching, from the novice researcher to the experienced data analyst. ESDS provides preservation, dissemination and user training for an extensive range of key economic and social data, both quantitative and qualitative, spanning many disciplines and themes. ESDS provides an integrated service offering enhanced support for the secondary use of data across the research, learning and teaching communities, covering a collection of several thousand datasets. Examples of data acquisitioned by the ESDS include the General Household Survey, the Labour Force Survey, National Statistics Time Series Data, British Household Panel Survey (BHPS) and the National Child Development Survey (NCDS).

Under the ESDS Qualidata Service, the acquisition of qualitative data is encouraged and the UKDA has a policy of identifying and ensuring that large paper collections of qualitative material are archived in suitable repositories.

AHDS History is also based at the UKDA. The AHDS History (formerly the History Data Service) is one of five Subject Centres of the Arts and Humanities Data Service (AHDS) and is a national data archiving service jointly funded by the Joint Information Systems Committee and the Arts and Humanities Research Board.

The Census Registration Service, also sited at the UKDA, was established to facilitate access to the four Census Data Support Units for UK higher and further education users (see below). These four units have all been funded by the ESRC and JISC to supply value-added census data.

**The National Archives** (TNA), which covers England, Wales and the United Kingdom, was formed in April 2003 by bringing together the Public Record Office and the Historical Manuscripts Commission. It is responsible for preserving the records of central government and the courts of law, and giving public access to open records, and to the appropriate government departments where they are closed. The collection is one of the largest in the world and spans an unbroken period from the 11th century to the present day.



TNA operates the UK public records system and lies at the centre of the national archival network, which covers archive services and other institutions holding official and private archival material of public interest. TNA acts as the custodian of the national memory as revealed in the records of central government and the courts. Its work begins with overseeing the creation and management of active records in government departments, continues with the selection and permanent preservation of public records of enduring historical value in whatever format, and culminates in making those records available online and onsite to an increasing number of people worldwide in the ways which are most convenient to them.

Following the Modernising Government white paper and the e-Government initiative, TNA created the Digital Preservation Department (DPD) to investigate and operate solutions to the problems of preserving very long-term access to digital government records. DPD has designed, built and are currently operating four major systems – the Web Archive, which archives government web sites, the Digital Archive system containing born digital government records, PRONOM, an online registry of technical file format information, and the Electronic Records Online (EROL) system, now released on the Internet<sup>1</sup>.

## 1.2 THE DEVELOPMENT OF THE OAIS REFERENCE MODEL STANDARD

For the first thirty years of digital preservation, archives managed their digital collections with the help of a repository for storage of offline media, or a simple storage system and a catalogue box or a catalogue database. Although the fundamental design of a digital archive system has remained the same – data storage plus metadata database – a contemporary digital archive needs more than a storage area for magnetic tapes and a spreadsheet for the catalogue. The rapid growth of digital material in both volume and complexity, the rising expectations of archives' users for access services and the emerging digital preservation strategies, have all contributed to the re-definition of digital archive functions. The functionalities and procedures of a digital archive have now been collected into a reference model that has become an ISO standard (ISO 14721:2003). The standard, first developed by the Consultative Committee for Space Data Systems (CCSDS), establishes a common framework of terms and concepts which comprise an Open Archival Information System (OAIS). It allows existing and future archives to be more meaningfully compared and contrasted and it provides a basis for further standardisation within an archival context. It should promote greater vendor awareness of, and support for, archival requirements.

The CCSDS was established in 1982 to provide an international forum for space agencies interested in the collaborative development of standards for data handling in support of space research. In 1990 the CCSDS entered into a co-operative agreement with Subcommittee 13 (Space data and information transfer systems) of the Technical Committee 20 (Aircraft and space vehicles) of the ISO. At the request of the ISO, the CCSDS assumed the task of co-ordinating the development of archive standards for the long-term storage of archival data in 1995. Although the CCSDS was initially to address the problems of archiving data obtained from observations of the terrestrial and space environments and used in conjunction with space missions, it soon took an intentionally interdisciplinary view and ensured broad participation in the discussion of a reference model for the long term storage requirements of this digital information<sup>2</sup>. The very first draft of the digital archive model was released after a year of work<sup>3</sup>; the draft was then discussed by international and national working groups and at workshops<sup>4</sup>, resulting in the publication of the first version of the OAIS model in 1999 and its update in 2001. Work had also begun on an additional standard guideline detailing the acquisition process: *Producer-Archive Interface Methodology Abstract Standard*<sup>5</sup>.

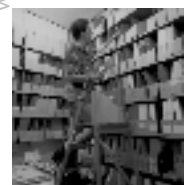
1 <http://www.nationalarchives.gov.uk/ero/>

2 Lavoie, 2000, p. 26

3 <http://ssdoo.gsfc.nasa.gov/nost/isoas/us01/p004.html>

4 cf. <http://ssdoo.gsfc.nasa.gov/nost/isoas/dads> and <http://ssdoo.gsfc.nasa.gov/nost/isoas/awiics>

5 <http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft.pdf>



Development of the reference model began with the premise that one of the greatest challenges in accepting preservation responsibility within an organisation is finding a shared vocabulary for stakeholders with a variety of backgrounds to use for productive discussion of the issues. Thus, the model was first developed to establish common terms and concepts, to provide a framework for elucidating the significant entities and relationships among entities in an archive environment, and to serve as the foundation for the development of standards supporting the archive environment. A broader task for the OAIS development has been defined as articulating the functionality and components of any system responsible for preserving any type of information over any length of time. The terminology used to describe the OAIS is often not the traditional archival or recordkeeping terminology since it is intended as a common language within which a diversity of communities can continue to implement and develop the OAIS model. The model has been very successful in one of its main goals – to spur further interest and discussion of digital preservation and archiving issues and standards. The 2002 CCSDS version of the OAIS reference model<sup>6</sup> was proposed and was accepted as an international standard in 2003: ISO 14721:2003 *Space data and information transfer systems – Open archival information system – Reference model*.

### 1.3 THE DEVELOPMENT OF THE METADATA ENCODING AND TRANSMISSION STANDARD

The Metadata Encoding and Transmission Standard (METS) is a recent standard designed to encode all varieties of metadata necessary for a complete description of digital objects within a digital library environment. Such objects may take the form of electronic texts, still images, digitised video, sound files or more interactive material such as VRML virtual environments. Until recently, no standardised method for encoding metadata on these objects has been available and as a consequence, digital library projects have tended to follow their own practice, often making use of whatever software package and data format the project team had become familiar with<sup>7</sup>.

METS is a community-based development, led by the Digital Library Federation and involving institutions such as UC Berkeley, Harvard University, the Library of Congress, Michigan State University, METAe, the Australian National Library, the RLG (Research Libraries Group), the California Digital Library, Cornell University and the University of Virginia<sup>8</sup>. The Library of Congress also hosts a METS web site for developing the standard and documentation.<sup>9</sup>

METS developed from the University of California at Berkeley's MOA2 (Making of America II) concept; a common object format which allowed for the sharing of effort of developing tools/services. MOA2 is a common object format which ensures interoperability of digital library materials as they are exchanged between institutions.

METS was created due to the continuing need to share, archive and display digital objects. It provides more flexibility for varying descriptive and administrative metadata than MOA2. METS was primarily intended for use within the digital library environment and was originally limited to objects comprising text, image, audio and video files. The METS format attempts to provide a standard format to hold metadata associated with a digital object, in a form which can easily be shared, cross-searched, exchanged and rendered for browsing and display purposes. METS is intended to be a flexible, yet tightly structured, container for all metadata necessary to describe, navigate and maintain a digital object (descriptive, administrative and structural metadata). METS is written in XML, a generic language designed for marking up electronic text.

6 <http://www.ccsds.org/documents/650x0b1.pdf>

7 Gartner, 2002, p. 3

8 McDonough, 2004

9 <http://www.loc.gov/standards/mets>



## 1.4 PROJECT AIMS

The aims of this project were:

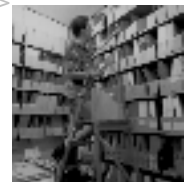
- to compare the preservation systems and metadata currently in use by the UKDA and TNA against those in the OAIS Reference Model and the METS standard;
- to practically test the theoretical argument that the two institutions comply with OAIS;
- to assess how the two institutions' operational structure can be informed by OAIS (and vice versa) and;
- to explore the potential for interaction between existing metadata standards utilised within the two institutions and METS.

It was expected that the results of this work would enable each organisation to assess the relevance of the OAIS reference model and the METS metadata standard to their work. A further aim was to report on the experiences of each organisation in undertaking the mapping. This is expected to be of use to other organisations that might want to undertake comparison to the OAIS standard.

This work explicitly excluded a lengthy description of the conceptual framework or component elements of the OAIS reference model. However, it has highlighted one aspect of the OAIS initiative in relation to the work of the UKDA and TNA, and one that potentially has wider implications for other HE/FE institutional repositories. The environment, information model and functional entities of an OAIS-type archive are intended to interact to form a broad conceptual framework characterising the primary entities, relationships and processes of that archive. However, for the framework to really work, it requires acceptance and integration of standards.

One particular approach of the research was to explore the METS metadata standard in relation to the work of the UKDA and TNA and the standards already employed at these institutions. One of the main challenges facing all digital repositories is the provision of seamless access to the assets within the repository. Access is partially dependent on the provision of different levels of metadata to describe the assets. The METS system would appear to be a good overall metadata solution as it fulfils all the criteria within the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) framework. Given that the OAIS Reference Model allows the conceptual mapping between heterogeneous systems, METS is one method of implementing this concept. The three main elements of this protocol are, to a certain extent, implemented in theory within METS. For both the Submission Information Packages (SIP) and Dissemination Information Packages (DIP) METS can be used as the syntax for the transfer. In DIPs METS can also be used to display data and associated applications; and for AIPs (Archival Information Packages) METS can be stored internally within the repository. In many respects METS can act as the 'glue' which holds together the different elements that make up a practical implementation of the OAI-PMH.

A number of concerns exist with METS in the preservation element of the repository process. The main concern relates to the use of namespaces which identify different DTDs used by different standards within the METS wrapper. The contents of these namespaces are likely to change, as for example, when a standard is updated, and while it is to be supposed that continued access will be possible, this may prevent long-term legacy material from being accessible.



## 2 WHAT DOES IT MEAN TO BE OAIS COMPLIANT?

It is not uncommon to encounter the term 'OAIS-compliant' used in reference to a digital archiving system. For example, the University of Texas Digital Asset Management System,<sup>10</sup> the Digital Information Archiving System (DIAS) built by IBM for the National Library of the Netherlands,<sup>11</sup> and the Online Computer Library Center (OCLC) Digital Archive service are all positioned as conforming to the OAIS reference model.<sup>12</sup> The architects point to its potential application as an implementation of the OAIS Archival Information Package concept.

The OAIS standard states that an OAIS-compliant digital archive implementation "supports the OAIS information model" (OAIS Ch. 2.2). It is also committed to "fulfilling the responsibilities listed in chapter 3.1 of the reference model" (see also chapter 6 below). Finally, the reference model notes that standards and other documentation that purport to conform to the OAIS reference model must incorporate relevant OAIS terminology and concepts, applied according to the interpretation and context defined in the reference model.

The OAIS standard allows for a "conformant OAIS archive providing additional services to its users that are beyond those required of an OAIS". It also assumes that "implementers will use the OAIS reference model as a guide while developing a specific implementation to provide identified services and content" (OAIS Ch. 1.4). The OAIS standard does not, however, assume or endorse any specific computing platform, system environment, system design paradigm, system development methodology, database management system, database design paradigm, data definition language, command language, system interface, user interface, technology or media required for implementation.<sup>13</sup>

The meaning of 'OAIS-compliant' is necessarily vague because the reference model is a conceptual framework rather than a concrete implementation. Conformance to the reference model can imply an explicit application of OAIS concepts, terminology, and the functional and information models in the course of developing a digital repository's system architecture and data model but it can also mean that the OAIS concepts and models are 'recoverable' from the implementation; in other words, it is possible to map, at least from a high-level perspective, the various components in the archival system to the corresponding features of the reference model. Further ambiguity is introduced when institutions and organisations claim OAIS compliance without defining or clarifying what this means in regard to their particular implementation.

The RLG and the OCLC funded an initiative to define the attributes of a trusted digital repository. A working group of international experts translated the OAIS models and concepts into a consensus statement on the responsibilities and characteristics of a digital repository housing a large-scale, heterogeneous collection of culturally significant materials. A key objective of this effort was to enumerate attributes of a digital repository which, taken together, serve to inspire trust within the archive's designated user community that the repository is indeed capable of preserving and making available the portion of the scholarly and cultural record in its custody.

First on the list of attributes of a trusted digital archive is:<sup>14</sup>

"A trusted digital repository will make sure the overall repository system conforms to the OAIS Reference Model. Effective digital archiving services will rely on a shared understanding across the necessary range of stakeholders of what is to be achieved and how it will be done.

10 <http://www.lib.utexas.edu/dams/development/system/index.html>

11 <http://www-5.ibm.com/nl/dias/index.html>

12 <http://www.oclc.org/digitalarchive/about/works/features/default.htm>

13 Lavoie, 2004, p. 18

14 RLG/OCLC, 2002, p. 13



The OAIS provides both a functional model – the specific tasks performed by the repository such as storage or access – and a corresponding information model that includes a model for the creation of metadata to support long-term maintenance and access. Organizations and institutions building digital repositories should commit to understanding these models and make sure all aspects of the overall system conform.”

The OAIS compliance testing in the current project is not solely based on the OAIS model, but also considers recommendations of the RLG/OCLC work on attributes of trusted digital repositories.

As a continuation of work on the definition of attributes for trust, RLG has set up a joint task force with the US National Archives and Records Administration (NARA) to develop a process of digital repository certification. This process should address the range of functions associated with repositories whilst providing layers of trust for all parties involved. It should yield a high degree of confidence that the information a repository disseminates is the same information that was ingested and preserved. The certification process must also address the consequences of failure, including fail-safe mechanisms that would enable a certified archival repository to perform the rescue of endangered digital information.<sup>15</sup> The task force is currently working on identifying and describing the elements of a digital repository that can be assessed and certified, the actual certification is planned for the future. The current project has, however, not yet taken into account the digital repository certification criteria being developed by the RLG and NARA working group.

### 3 METHODOLOGY

---

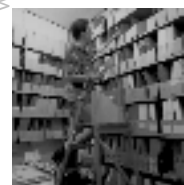
Since there is, as yet, no formal OAIS standard certification process in place, the UKDA and TNA had to develop their own methodology for testing their OAIS compliance and mapping the concepts, processes and responsibilities. The underlying idea was that the concepts, terminology and models of the OAIS model represent a common point of reference around which comparisons and interoperability could be built.

Having studied the mandatory responsibilities listed in the OAIS standard itself, it quickly became clear that it would be difficult for any functioning archive not to comply with these criteria. It was therefore decided to conduct a more detailed-level study and map the OAIS functional entities to the workflow processes at both archives. The aim of this exercise was twofold: first, to discover any gaps in the current workflow and procedures of the archives compared with the OAIS recommendations; and, second, discovering if any assumptions have been made in the OAIS standard, that make it impossible for these archives to comply.

Both the UKDA and TNA, chose a multi-pronged approach to the methodology for compliance testing. As a first step, all relevant documents were identified from the internal controlled document list. The documents selected as relevant to this project included a wide range of materials such as procedural documents, licences and depositor forms. Concurrent with this work, other documents were identified and in some cases created or enhanced. In the case of the UKDA, other documents were also considered but not used, such as detailed job particulars for staff working in relevant areas. In addition, a number of information documents, available from the organisations’ web sites and intended for use by the designated communities, the consumers and producers were also considered relevant. Other information was gathered from face to face meetings and discussions with staff who have key responsibilities, for example, the Systems and Preservation Manager.

Having completed the information gathering and generation stage of the project, the material was collated and ordered to map to, or match, as far as possible, the responsibilities set out in the OAIS reference model.

<sup>15</sup> [http://www.rlg.org/en/page.php?Page\\_ID=7783](http://www.rlg.org/en/page.php?Page_ID=7783)



## 4 COMPLIANCE WITH OAIS RESPONSIBILITIES

The OAIS standard establishes mandatory responsibilities (in Chapter 3.1) that an organisation must discharge in order to operate an OAIS archive. In order to fulfil these relatively broad requirements, the OAIS must:

- *negotiate for and accept appropriate information from information producers;*
- *obtain sufficient control of the information provided to the level needed to ensure long-term preservation;*
- *determine, either by itself or in conjunction with other parties, which communities should become the designated community and, therefore, should be able to understand the information provided;*
- *ensure that the information to be preserved is independently understandable to the designated community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information;*
- *follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original;*
- *make the preserved information available to the designated community.*

Both TNA and UKDA found that these responsibilities are generally carried out by almost any archive and the compliance with them is, therefore, not difficult to meet. The requirement that raised further questions is the “dissemination of information as authenticated copies of the original”. Since the OAIS standard does not explain what is meant by ‘authenticated’, further investigation is necessary to ascertain what it means in this context, and whether digital records are covered adequately by existing legislation.

Both TNA and UKDA currently provide their users with digital material that can be traced back to the original deposited version using extensive metadata kept by the archive and its policies.

Chapter 3.2 of the OAIS standard provides some examples of how the mandatory responsibilities of an OAIS archive can be discharged. Details of the compliance to these individual requirements follow below.

### 4.1 NEGOTIATES FOR AND ACCEPTS INFORMATION FROM PRODUCERS

“An organisation operating an OAIS will have established some criteria that aid in determining the types of information that it is willing to, or it is required to, accept.” For the archives participating in the compliance testing, the selection criteria are determined by legislation, appraisal policy, operational selection policies and acquisition policy.

The OAIS standard envisages that the OAIS archive should extract, or otherwise obtain, sufficient descriptive information from the data depositors to assist the designated user community in finding the digital objects of interest from the archive. It should also ensure that the information meets all OAIS internal standards.

It is customary for archives to set certain criteria on quality of the material they accept from depositors. This practice has become a standard for digital archives that take on the responsibility for long-term preservation of the deposited material and can fulfil this task only on the condition that the deposited material meets certain criteria for preservation.



### Scope of Collections and Selection Criteria: UKDA

The UKDA collects information, data and other electronic resources of long-term interest and use across the range of social science and historical disciplines. They are acquired to support research and teaching activities in the UK and elsewhere. The studies acquired contain a mixture of textual and numeric data as well as other less-used formats such as image and audio files. The majority of the data result from survey materials but also include administrative, business and aggregate statistical information. New data collections include both quantitative and qualitative surveys. The UKDA collection development policies are implemented in line with the acquisitions policies of the different services so the collection content varies depending on the service.<sup>16</sup> For example, some government surveys are designated as part of TNA's collection; the AHDS History service has its own collections development policy and acquisition strategy; and the UKDA undertakes the collection of material for dissemination and preservation via its own projects. A typical example of the latter is the census digitisation project Online Historical Population Reports. This is a project run by AHDS History working closely with TNA. It is a JISC-funded project to digitise all UK census reports 1801-1937, Registrar General(s) reports 1801-1921 and ancillary material. The project will result in a web-based user interface for browsing, searching, viewing and downloading images of historical population reports. The interface will also allow the viewing and downloading of machine-readable versions of a number of statistical tables contained within the reports.

In general, the selection of materials falls into three key areas:

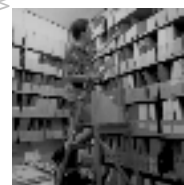
- data and electronic resources for research, for example, data that are suitable for informed use in a variety of research purposes;
- data and electronic resources for teaching and learning;
- replication data and electronic resources, the material (data, computer programs and instructions, and related outputs) necessary for the replication of published or unpublished research.

The UKDA will seek to acquire material:

- at the specific request or recommendation of a user or group of users;
- on the recommendation of the relevant service Advisory Committee;
- when the data collection has been fully or partially funded by organisations whose area of interest and expertise matches a particular service.

Occasionally, the UKDA accepts material for preservation only, for example the JISC New Opportunities Fund (NOF) projects. This represents a preservation function which falls outside the scope of the collections policy and is something that does not fall into the OAIS model. Factors affecting all the service policies include data usage by users and whether or not the data being offered are an update of an existing collection. The number of new data collections which are acquired each year is restricted by available resources. However, for exceptional collections, additional resources would be sought.

<sup>16</sup> cf. <http://www.esds.ac.uk/aandp/create/policy.asp#scope>



To ensure that the UKDA continues to build a collection of value, clear criteria are applied to assess their content, long-term value and the level of potential interest in their re-use. Factors influencing this assessment include:

- the geographic and/or temporal scope is significant;
- the subject coverage of the data is broad and may be of interest across the relevant disciplines;
- the data are not available in any other form, e.g., paper;
- accession into the UKDA makes the resource more accessible;
- a dataset adds to or is made more valuable by existing holdings, in particular where it fits into an existing series;
- a dataset fills a gap in the existing holdings;
- there is research and/or teaching activity in the subject area covered by the data;
- data for which longevity and access would otherwise be threatened.

Criteria are also applied to assess whether material may be viably managed, preserved and distributed to potential secondary users. Factors considered include checking that:

- the data are of a type with which the UKDA has expertise or may easily obtain expertise or expert advice;
- the data format can be converted to suitable dissemination and preservation formats;
- the level and quality of documentation reaches an appropriate standard to enable a secondary analyst to make informed use of the data. Ideally, datasets would be documented to UKDA standards as outlined in the *Research management and documentation guidelines*.

General guidance on creating and depositing both qualitative and quantitative data at the UKDA are published on its web site.<sup>17</sup>

AHDS History Data collections are accessioned for all periods, from ancient history through to 1945, and although the primary focus is on the UK, cross-national data collections are regularly accessioned. Data are accepted in a variety of formats, including ASCII files, database files, spreadsheet files, image files and SGML marked-up texts, and on a variety of media including CD-ROMs and disks, as well as via FTP. Academic projects funded by the AHRC are required to offer digital resources for deposit with the AHDS. Projects funded by the JISC, the ESRC, the British Academy, the Leverhulme Trust and other funding bodies that support Higher Education can also normally deposit.

When a data collection is deposited with the AHDS History service, it is first validated and then archival copies are made to ensure its long-term preservation. A full catalogue record is created which describes the data and is included in the UKDA catalogue. AHDS History also manages the distribution of the data collection to users in the research and teaching community, and regulates access in accordance with the terms and conditions chosen by the depositor.

Similarly to the UKDA, depositors are asked to complete a Data and Documentation Transfer Form,<sup>18</sup> a catalogue form<sup>19</sup> and a licence.<sup>20</sup> The two former are equivalent to the UKDA's Data Collection Form.

<sup>17</sup> <http://www.esds.ac.uk/aandp/create/research.asp>

<sup>18</sup> <http://www.ahds.ac.uk/documents/transfer-form.doc>

<sup>19</sup> <http://www.ahds.ac.uk/documents/ahds-catalogue-form.doc>

<sup>20</sup> <http://ahds.ac.uk/documents/ahds-history-licence-form.doc>



## Scope of Collections and Selection Criteria: TNA

TNA collects records defined under The Public Records Act 1958,<sup>21</sup> and amended by various statutory instruments and further Acts including the Freedom of Information Act; all of these are publicly available on government web sites. Without quoting the full legislation, it covers records of central government, the law courts and public enquiries. All material held is at a security classification of Restricted or below.

Examples of TNA (paper) material include the Domesday Book, Cabinet Office minutes, census data, and records of war medals issued, to name but a few. Many items are of particular interest to the genealogical community who form one of TNA's biggest user communities. Occasionally private material is donated to, and accepted by, TNA if it has special historical interest and relevance to the collection.

TNA has three main archives for born digital material:

- The Web Archive.<sup>22</sup> Web sites from central government are archived in collaboration with the Internet Archive.
- TNA holds other digital records in its Digital Archive.<sup>23</sup> This material includes digital objects of a wide range of types, as working file formats to other government departments cannot be prescribed. These include documents, images, databases, emails, spreadsheets, public enquiry web sites, video, audio and virtual reality models.
- The National Digital Archive of Datasets<sup>24</sup> holds datasets from central government departments. These are typically survey or census type materials, which appear in tabular form and require extensive supporting contextual documentation to make sense of the information.

For the purposes of this report, NDAD is considered to be out-of-scope, as it is effectively run as a separate archive by ULCC under contract with TNA.

TNA defines thematic Operational Selection Policies,<sup>25</sup> which outline the process by which material is selected for permanent preservation for each type of material, so there is no single agreement. Since TNA collects material from a wide variety of sources, agreements controlling quality and acceptance criteria are formed by agreement between TNA and the departments and agencies concerned. Where a specific Operational Selection Policy is not applicable, an Acquisition Policy<sup>26</sup> applies.

TNA's Digital Archive does not operate a prescribed file format list, as the archive cannot dictate the form in which records are electronically created by all central government departments and agencies, thus there is a very wide potential range of digital object types. Records are transferred to TNA in the form in which they were created by the government department or agency. However, TNA does provide guidance on format selection for long term preservation.<sup>27</sup>

Metadata that accompany a submission are strictly controlled and validated through the use of a Java applet that manipulates and validates an XML file that must accompany the submission and which conforms to the Digital Archive schema and to the e-GMS (Government Metadata Standard). All digital objects must be assigned to one or more record references, fixity information is appended, and sufficient descriptive information provided. This is done in collaboration between the Departmental Records Officers in the government department and client managers from TNA using the applet. Descriptive information is further validated once it arrives at TNA, and in

21 cf. <http://www.nationalarchives.gov.uk/policy/act/act.htm>

22 <http://www.nationalarchives.gov.uk/preservation/webarchive/>

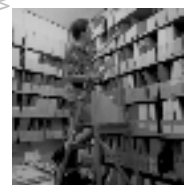
23 <http://www.nationalarchives.gov.uk/preservation/digitalarchive/>

24 <http://ndad.ulcc.ac.uk/>

25 <http://www.nationalarchives.gov.uk/recordsmanagement/selection/ospintro.htm>

26 <http://www.nationalarchives.gov.uk/recordsmanagement/selection/acquisition.htm>

27 [http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting\\_file\\_formats.pdf](http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_file_formats.pdf)



some cases enhanced, by the Catalogue Unit. Technical information is likewise validated and enhanced by digital preservation staff, although increasingly it is becoming possible to automate the production of much of the technical metadata.

Selection of which records to take is covered by TNA's appraisal policy<sup>28</sup> and TNA is currently developing a custodial policy for electronic records.<sup>29</sup>

## 4.2 OBTAINS SUFFICIENT CONTROL FOR PRESERVATION

The OAIS standard recommends that when acquiring the content from a producer, the OAIS archive must ensure that there is a legally valid transfer agreement that either transfers intellectual property rights (IPR) to the archive, or clearly specifies the rights granted to the OAIS and any limitations imposed by the rights holder(s). The OAIS must ensure that its subsequent actions to preserve the information and make it available conform with these rights and limitations. At the same time, the OAIS archive must assume sufficient control over the objects and their metadata so that it is able to preserve them for the long term.

### Depositor Agreement: UKDA

The UKDA Depositor Licence<sup>30</sup> allows the University of Essex (the host institution for the UKDA and the formal legal entity) to:

- distribute archived data collections to registered users in a variety of formats;
- to catalogue, enhance, validate and document the data collection;
- to store, translate, copy or re-format the data collection in any way to ensure its future preservation and accessibility;
- incorporate metadata or documentation in the data collection into public access catalogues.

The Depositor Licence is designed to preserve IPR and ownership of data and copyright of the original data remains with the depositor. The licence refers to re-use of the data collection, e.g., for educational and research purposes and/or commercial purposes. Royalty payments may be collected on behalf of the depositor.

The Depositor Licence works in conjunction with the End User Licence (EUL) to pass on the responsibility of IPR, respondent confidentiality and other conditions of use agreed in the Depositor Licence. The sharing of data with other researchers or students and the re-use of data for a new purpose is restricted by the terms and conditions outlined in the EUL that all users agree to when registering with the Economic and Social Data Service (ESDS).<sup>31</sup> On occasion, a depositor requests that special conditions are attached to a deposit form. These will be reflected in the EUL which obliges users to observe these extra conditions. For example, permission must be obtained from the Home Office before users can access certain parts of the British Crime Survey datasets, such as data about stalkers.

### Depositor Agreement: TNA

TNA does not require a depositor to sign a licence to transfer records each time as this is covered by legislation (Public Records Act 1958 and subsequent acts). However, TNA does require that its depositors sign a transfer form (AA2) giving authority to the specific records transfer.<sup>32</sup>

28 <http://www.nationalarchives.gov.uk/recordsmanagement/selection/appraisal.htm>

29 [http://www.nationalarchives.gov.uk/recordsmanagement/custody/pdf/custodial\\_pol\\_draft.pdf](http://www.nationalarchives.gov.uk/recordsmanagement/custody/pdf/custodial_pol_draft.pdf)

30 Available from <http://www.esds.ac.uk/aandp/create/depproc.asp>

31 <http://www.data-archive.ac.uk/orderingData/sharingData.asp>

32 [http://www.nationalarchives.gov.uk/recordsmanagement/advice/pdf/cat\\_aa2.rtf](http://www.nationalarchives.gov.uk/recordsmanagement/advice/pdf/cat_aa2.rtf)



Rights over material deposited at TNA are described in detail in the Public Records Act.<sup>33</sup> Largely, material will be Crown Copyright, although provision is made for TNA to legally archive and make available material with other copyrights attached. TNA has a legal responsibility to preserve records selected for preservation by agreement with the department or agency concerned, and to provide access to those records to the public where they are open, and additionally to the originating department or agency if they are closed. TNA is responsible for answering Freedom of Information requests on material held by TNA, including closed records.

### 4.3 DETERMINES DESIGNATED CONSUMER COMMUNITY

The OAIS standard requires that the designated user community is identified when material is submitted. This is necessary in order to determine whether the information, as represented, will be understandable to that community. In the OAIS thinking, the determination of the designated user or consumer community is crucial to the selection of preservation methods and metadata.

#### User Community: UKDA

The UKDA user community is largely determined by its contractual obligations with the organisations which fund the services under the umbrella organisation. By far the greatest number of users are users of ESDS and are drawn from the Higher and Further Education communities. Some 7-8% of users are from the public sector, with most of the rest from HE/FE. The recent collaboration with TNA may result in a new community of general public users but this is not anticipated in the near future due to the small amount of material that will be deposited under this agreement. Further constraints depend on technological progress: in the past when the UKDA computer room was housed in an entire top floor, data could only be transferred on 12" reels and users had to write their own command files to produce tables or statistics. Consequently the designated community was extremely limited and specialist. Nowadays, with online browsing via Nesstar,<sup>34</sup> any user can select a couple of variables and produce a table with only basic computer literacy skills.

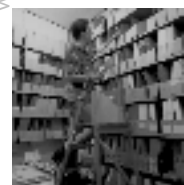
This has been one of the biggest challenges facing the UKDA over recent years: the specialist user of the past usually had programming skills and could understand a dataset at its fundamental file level. Such users gained an understanding of a file and its content by reading technical information such as column and row information and coding details. Many users are now unable to understand data at this level and require new tools to enable them to produce the tables they want. In order to serve this community and provide the level of functionality it expects, the UKDA now adds significantly more value to its collections by the capture or transfer of information into a structured form to make it computer understandable. This approach has enabled the UKDA to remain at the forefront of technological developments for searching and browsing datasets and in the development of tools for both users and data archivists. For example, access to resources has been made easier by the introduction of Nesstar and of the Integrated Data Catalogue (IDC),<sup>35</sup> which permits simultaneous searching of the catalogues of many of the European data archives.

Resource discovery and the developments surrounding the catalogue provide a typical example of how changes to users' levels of skill have impacted directly on the work of UKDA staff. In its first guise, the catalogue was available only for staff internally and skilled staff were employed specifically to interrogate it to provide information for users about dataset content. This work preceded email communication and the requests were mainly by phone, with some by letter. It is now much less usual for researchers to seek help in using the catalogue from staff at the UKDA, because so many users are familiar with the web and the use of search and browsing software to discover the information for themselves.

<sup>33</sup> <http://www.nationalarchives.gov.uk/policy/act/act.htm>

<sup>34</sup> Nesstar is an infrastructure for data dissemination via the Internet. The Nesstar Explorer provides an end user interface for searching, analysing and downloading data and documentation and the Nesstar Publisher provides the tools and resources for making the data and documentation available via the Internet. See: <http://www.nesstar.org/>

<sup>35</sup> <http://dasun3.essex.ac.uk/Cessda/IDC/>. The IDC will shortly be completely updated. The new version will have a different underlying architecture and functionality to permit simultaneous searching of the catalogues of many of the European data archives. It will be re-named C-CAT)



The UKDA has also increased the availability of data formats for users and has modernised its data distribution system in response to user requirements. Specific projects have also enabled UKDA to place more archived material on the web via user-friendly sites with new dissemination mechanisms. One such example is Edwardians Online, a project which has improved access to archived, qualitative data offering content-based access to a collection of oral history interviews with people who lived in Edwardian Britain.

Similarly, the CHCC project offered the opportunity to develop the Collection of Historical and Contemporary Census data and related materials (CHCC) into a major Distributed National Electronic Resource (DNER) for learning and teaching. It has successfully promoted increased and more effective use of network-based data services for problem-based learning and student project work and has developed an integrated web-based learning and teaching system linking data extraction and visualisation/exploration tools with comprehensive learning and teaching resources.

These developments have necessarily resulted in enhancements to the preservation system to accommodate new levels and types of information and additional formats.

These improvements in access have undoubtedly contributed to an expansion in the number of users of the UKDA over recent years. Moreover, as the richness of the UKDA's collections has been more widely promoted and appreciated, demand to access them has also increased.

There has to remain an assumption that the majority of users have good computer skills and a certain level of understanding of social science data. The UKDA adds value to the data by the addition of metadata and changing the formats of data to aid access. These assumptions about the designated community are applied during the acquisition stage when data are acquired which are assumed to be of interest to users.

### **User Community: TNA**

The Public Records Act defines TNA's user community as government departments and agencies and the public. This designated user community is extremely broad and informs the appraisal, selection, preservation and access policies.

Records of historical interest are appraised and selected for permanent preservation in collaboration with the submitting government agency, and through the use of Operational Selection Policies that are tailored to the agencies concerned.

TNA must provide access to as wide a community as possible. Viewing technology must be accessible to the greatest number of people possible. TNA has, therefore, mandated the use of open standards in information presentation as far as possible. Access to digital records is both via the Internet and in dedicated reading rooms.

Within this very wide designated community however, TNA is cognisant of the fact that it does serve particular special interest groups, including schools, family historians, academics, archivists and journalists, and that they each have particular levels of experience with digital technology which again informs TNA's advice and means of presentation. Specific user groups have been formed to represent different sectors of the community.

Significant efforts are being made to place increasing amounts of material online, including the digitisation of paper records of interest to specific communities, and the recent release of the EROL system.

Work is ongoing to better define and serve the particular needs of the designated community, including analysis of queries and records access, usability studies, and the production of specialist guidance and advice on conducting research at TNA.



#### 4.4 ENSURES INFORMATION IS INDEPENDENTLY UNDERSTANDABLE

An OAIS archive must determine what it can and has to do in order to preserve the usability and maintain the understandability of its collections. Since the usability requirements of users change over time, the archive must choose a preservation strategy that reduces the risk of non-usability of its collections. According to the OAIS standard (OAIS Ch. 3.2.4): “Even when a set of information has been determined to be understandable to a particular designated community, over time the Knowledge Base of this community may evolve to the point that important aspects of the information may no longer be readily understandable. At this point it may be necessary for the OAIS to enhance the associated representation information so that it is again readily understandable to the designated community.”

##### Digital Preservation Strategy: UKDA

The UKDA takes a practical and pragmatic, migration driven, approach to preservation, which has evolved since the establishment of the Data Bank in 1967. It is worthy of note that when the UKDA began life as a databank, preservation was not an issue and is not even mentioned in early material. It only became clear, years later and especially as storage moved from cards to magnetic tape that migration of data types and formats needed to be considered and acted upon if the data held were to be kept alive. This is an important point in that an organisation may not start out thinking that preservation is an important element of their work but over time priorities will change.

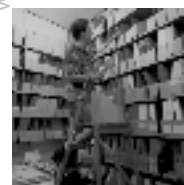
Subsequently, the UKDA developed a Preservation Policy document which is part of a defined and stated policy. The strategy is based upon open and standardised file formats, data migration and media refreshment. Preservation decisions are made within the context of the Collections Development Policy, balancing the constraints of cost, scholarly and historical value, and user accessibility. Different preservation techniques may be required for material with different levels of quality and significance.

The UKDA recognises that in principle no file format or physical storage media is going to last forever. Indeed, it has seen movement from punched cards and paper tape, through 7 and 9 track tapes to optical media and high capacity magnetic tape cartridges. As a consequence, a strategy has been adopted to store data on at least two and often three different storage media. These are reviewed regularly and data are copied onto new media when appropriate.

The minimum number of preservation formats that are necessary to manage the full range of data types in the UKDA's collections through time has been identified; migration paths for these are carefully chosen. Wherever possible these are standard formats that require little or no migration. The ASCII format is used as a lowest common denominator to facilitate the reading of the data by any program. In addition, the data are stored in the format as received from the depositor, typically SAS, SIR or SPSS. SPSS portable format is especially desirable as it is an ASCII-based format that is platform independent.

However, the advent of long variable names (up to 64 characters) introduced from SPSS version 12 onwards has caused the UKDA to review this position and, during 2005, new preservation formats are being introduced:

- fixed width text of specified character set with accompanying SPSS, STATA and SAS command files and variable level DDI XML file;
- tab-delimited text with UKDA data dictionary.



This plurality of formats for encoding the variable level metadata is a pragmatic decision to maximise ease of data migration as the future need arises. Ideally, a fully comprehensive and open XML standard for describing statistical datasets will emerge in the next couple of years to provide the UKDA with a single definitive preservation format.

The UKDA approach is aimed at facilitating translation of the data into a format specified by the user and more importantly ensures the preservation of the maximum amount of metadata. The UKDA endeavours to follow international best practice in its choice of preservation formats and data migration procedures.

Defining, timing, testing and implementing migration pathways are the responsibility of the Systems and Preservation group. When new formats are created from data files either through migration into new file formats or through the creation of new file formats for dissemination, the old files are retained alongside.

The preservation strategy of the UKDA aims to maintain a flexible preservation system that evolves to meet the demands of changing technology and new and increasing user expectations. The preservation policy covers preserving data collections for which the UKDA is a custodian and does not consider preservation of other materials such as the UKDA web pages, internal administrative documents and correspondence and the UKDA's intranet. These materials are governed by the UKDA's records management programme.

In fulfilling this mission, the UKDA strives to ensure that the:

- materials it acquires and accessions are suitable for scholarly use;
- data are accompanied by adequate documentation to enable their use for secondary analysis;
- data are checked and validated according to strict data processing procedures;
- data are professionally catalogued and indexed according to appropriate standards;
- data are effectively preserved for future use by converting them to several standardised formats and retaining multiple copies on different storage media;
- format of materials is changed as necessary to preserve access to their intellectual content, reducing the risk of losing access to them over time;
- materials are kept in conditions suitable for long-term archival storage.

### **Preservation Strategy: TNA**

TNA is adopting a migration-driven strategy for digital preservation, but is not ruling out the application of other techniques such as emulation if appropriate and available. While the Digital Preservation Department cannot control the formats in which it receives information, it intends to select open target standards for information representation of common record types, ensuring that the records remain manageable over time and tend to converge on common solutions. Extensive technical and archival metadata are captured, both during ingest and any subsequent migrations, informed by the Technology Watch function. Additionally, presentation technology requirements are fed back to Technology Watch by the Online Presentation Department at TNA.

Original bit streams transferred to the archives are held in perpetuity, along with all previous preservation manifestations and current presentation surrogates. Full metadata histories are maintained, even for obsolete manifestations and surrogates in support of the presumption of authenticity.



The system in which these records are stored and managed is called the Digital Archive. Work on the design of the system began in March 2002, and the system was operational by March 2003. It comprises an HSM (Hierarchical Storage Management) system for scalable information management, a relational database holding records metadata, and a web based J2EE Java application server providing a management front end. Records metadata are stored both relationally and as XML.

The HSM system in the Digital Archive currently holds information on tapes, and provides for bit level integrity checking, media refreshment and multiple copies of material within a single management system. Additionally, TNA preserves information on two archive systems and through backups. Presentation surrogates of digital records are managed via TNA's resilience infrastructure, which includes clustered servers and replicating network file stores held at different locations.

Aside from the Digital Archive, other systems required for migration of records have been built and are still in active development, including PRONOM. The first version of PRONOM was developed by The National Archives Digital Preservation Department in March 2002. Its genesis lies in the need to have immediate access to reliable technical information about the nature of the electronic records now being stored in the Digital Archive. By definition, electronic records are not inherently human-readable - file formats encode information into a form that can only be processed and rendered comprehensible by very specific combinations of technical components, such as hardware, software and operating systems. The accessibility of that information is therefore highly vulnerable in today's rapidly evolving technological environment. This issue is not solely the concern of digital archivists but of all those responsible for managing and sustaining access to electronic records over even relatively short timescales.

Technical information about these technical components is therefore a prerequisite for any digital preservation regime.

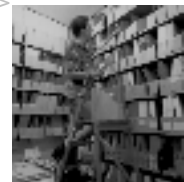
#### **4.5 FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES**

It is essential for an OAIS archive to have documented policies and procedures for preserving its collections, and it should follow those procedures. The producer and consumer communities should be provided with submission and dissemination standards, policies, and procedures to support the preservation objectives of the OAIS.

##### **Preservation Policy: UKDA**

Digital preservation is the main crux of the UKDA and is covered by the preservation policy document, which is concerned with the preservation of information on optical and magnetic media. It can be defined as the actions needed to ensure enduring access to the full content of digital resources over time. Data and documentation will be converted to and held in stable formats which are considered to be as software and hardware independent as possible. The UKDA will monitor its preservation policy as necessary to account for technological shifts, changes in perceived best practice and the nature of the UKDA holdings.

The document is available on request and, via a protected web site, to interested readers. It is not made widely available as its detail may compromise system security. Consideration is currently being given to the production of summary version of the document which would be made publicly available.



## Preservation Policy: TNA

The current TNA preservation policy outlines TNA's preservation strategy, for records on all media, including born digital records. The policy is currently being revised to take account of recent developments.<sup>36</sup>

TNA's chosen primary digital preservation strategy, migration, is intended to ensure the continued meaningful existence of electronic records by replacing the obsolete archival record with a new digital version. Migration also entails managing the risk of information loss by assessing technological change and preserving contextual information or metadata that might be lost during data migration. Migration must also include continuous maintenance of the history of migrations as part of the metadata associated with the record, which will be made available to the user. Each migration is termed a manifestation, and multiple manifestations will be produced and preserved for a single record.

## 4.6 MAKES THE INFORMATION AVAILABLE

The expectations of OAIS users (consumers) regarding access services will vary widely among archives and over time as technology evolves. Pressures for more effective access must be balanced with the requirements for preservation under the available resource constraints. Multiple views of OAIS holdings, supported by various search aids that may cut across collections, may be provided.

Some collections may have restricted access and therefore may only be disseminated to consumers who meet access requirements. The OAIS should have published policies on access and restrictions so that the rights of all parties are protected.

## Dissemination Policy: UKDA

The UKDA operates under a system of Depositor and End User Licences which protect against the transfer of any interest in intellectual property from the data collection funders, service funders, the data service providers, the original data creators, producers, depositors, copyright or other right holders (including without limitation the Office for National Statistics or the Crown). Users of data held at the UKDA have to agree to acknowledge in any publication, whether printed, electronic or broadcast, the original data creators, depositors or copyright holders, the service funders and the data service provider, as specified in the data distribution notes or in accompanying metadata. Consequently, access restrictions may apply to some users/usages.

Nevertheless, the UKDA strives to make data available for secondary analysis and ordinarily the concept of 'closed records' does not apply to UKDA collections. Moreover, as a publicly funded organisation, the UKDA has responsibilities under the Freedom of Information Act (FoI). As an agent of TNA, the UKDA has the same responsibilities as TNA for material deposited under this service. Insofar as the UKDA Depositor Licence does not transfer any interest in intellectual property from the depositors, requests for information for datasets deposited to the non-TNA services must be answered but responsibility for providing the information remains with the depositor. The assumption however, has always been that data deposited at the UKDA are available to any person who can meet the requirements of the EUL as defined by the conditions set by depositors in the Depositor Licence. Exceptionally, a dataset may be held for preservation only and occasionally, a temporary embargo can be placed on a dataset, for example, for reasons of confidentiality or as a requirement from the depositing research team.

<sup>36</sup> [http://www.nationalarchives.gov.uk/about/pdf/preservation\\_policy.pdf](http://www.nationalarchives.gov.uk/about/pdf/preservation_policy.pdf)



The UKDA has no formalised policy on dissemination of its holdings, but guidelines have been created and some of the principles governing dissemination are included in the service contracts. Some information on dissemination is also embedded in other documents (e.g., the procedures for deciding the processing level and therefore whether a dataset will be made available in Nesstar). In providing access to its collections, the UKDA is regulated by the deposit agreements that establish use conditions for every data collection and by the access agreements that its users have to accept.<sup>37</sup>

One aim of the UKDA is to develop interface and analysis tools appropriate to differing levels of expertise amongst users. The UKDA recognises that the use of its collections is a prime motive for its existence. The UKDA engages with a wide range of stakeholders, including data suppliers, data funders and end users.

All users can access the catalogue, including study descriptions and online documentation, such as questionnaires, free of charge and without registering. Registered users can also download and explore, or analyse online, a large and growing number of datasets. Registration requires that users accept standard conditions of use for all datasets and additional conditions for certain datasets. The registration process is a de facto legal agreement with the user that they will act by certain terms and conditions. It is not simply a process by which UKDA collects user contact details. Registered users can also request data on CD or other media but charges may apply for this service. The UKDA distributes and provides access to data from its collections via:

- HTTP download;
- online access;
- guest FTP;
- CD-R and DVD-R;
- other media by special request, e.g., DAT, Exabyte, Zip disc;
- specialist services such as Edwardians Online.

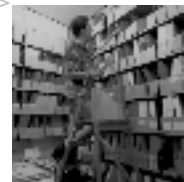
The UKDA's HTTP-based download service provides a quick and reliable means of gaining access to the most heavily used collections held at the UKDA. The UKDA also provides online access to data that have been enhanced and published via the Nesstar system. A minority of data are mounted in the Nesstar system for online browsing and visualisation of the data, including tabulation, graphing, book marking, sub-setting, filtering and downloading. The system is based on the DDI standard.

Historical data collections can also be downloaded from the AHDS History web site, once the user has agreed to the terms and conditions. The terms and conditions are similar to the EUL in that they require the user to adhere to confidentiality agreements and agree that the data will only be used by registered users for non-commercial purposes. Those data that are not available via download can be accessed through the UKDA online ordering system.

Access restrictions may apply to some users/usages – details can be found in the relevant dataset online catalogue record, e.g., not-for-profit community only.<sup>38</sup>

<sup>37</sup> <http://www.esds.ac.uk/aandp/access/licence.asp>

<sup>38</sup> <http://www.esds.ac.uk/aandp/access/introduction.asp>



## Dissemination Policy: TNA

TNA access policy is governed by the Public Records Act and the Freedom of Information Act. TNA has a duty to make records available to the public and to originating government departments and agencies.

Until the Freedom of Information Act, records were presumed to be closed for general access for 30 years. Some records could be closed for longer than this if required. Additionally, records at Restricted level would be reviewed to see if they could be declassified. Currently, records are presumed to be open, if not explicitly restricted, but can have Freedom of Information exemptions applied to them to limit access. These exemptions are detailed in part II and particularly part VI of the Act.<sup>39</sup> TNA must track any exemptions that apply and limit access to records accordingly. Naturally, over time these decisions can be reviewed, for example if a Freedom of Information request is received, and amendments made.

Although TNA would consult with the originating department before declassifying a record, ultimately the decision lies with TNA and is based on the following criteria from the Freedom of Information Act:

- (a) that, in all the circumstances of the case, the public interest in maintaining the exclusion of the duty to confirm or deny outweighs the public interest in disclosing whether the authority holds the information, or*
- (b) that, in all the circumstances of the case, the public interest in maintaining the exemption outweighs the public interest in disclosing the information.*

Aside from this, TNA can place some restrictions on the publication of material that is regarded as sensitive, such as photographs of a disaster taken from a public enquiry. Nevertheless, if a special request were made; such material would be available to view in the reading rooms.

The main access channels to the TNA holdings are the reading rooms at Kew and via the Internet<sup>40</sup> but TNA can also arrange to supply copies of records on other media, or in alternative formats. Charges may apply where costs are incurred by TNA, such as for digitisation or copying onto physical media (users can set up online accounts to pay any charges). Access to records on site is free. All charges for records are set out in the various Statutory Instruments.

## 4.7 CONCLUSIONS

The compliance testing of UKDA and TNA against the OAIS mandatory responsibilities resulted in a good match between the expectations and the practice at the two institutions. Since both UKDA and TNA have functioned as dedicated archives for a long time, and it is their mission to collect, preserve and disseminate data and records, it is to be expected that they have clear policies in place that govern their archival activities and that these meet the OAIS requirements. Indeed, when it comes to digital preservation strategies and practices for ensuring long-term preservation of their holdings, both UKDA and TNA have more exact regulation in place than required by the OAIS standard.

The biggest discrepancy with the OAIS's concept of an archive's responsibilities was revealed to be the understanding of the Producer and the Designated Consumer Community. In an ideal world the archive would exercise strong control over the producers of material that is later handed over to the archive – digital preservation is said to begin with the creation of the object that is to be preserved and archival requirements should be considered at the time of creation of a digital object. In the case of TNA, the archive has in reality very little control over the format in which material is created and later received from their donors. This situation has its roots in the legal framework, as well as in the way the material is created and its creation funded.

39 <http://www.hms.gov.uk/acts/acts2000/20000036.htm#aofs>

40 see also: <http://www.documentsonline.nationalarchives.gov.uk/help/default.asp>



The OAIS model frequently points to the strong link between the user community and the way the material in the archive should be described and preserved. Indeed, it is commonly understood that preservation only has a meaning if the preserved information is used. At the national level, for archives that have existed for decades and offer services nation-wide it is, however, difficult to limit the user groups or communities to as narrow groups as the OAIS standard examples. For example, TNA is collecting records not for a single user community but for many different (most of them as yet unknown) uses. The collections within the Digital Archive at the TNA are, as yet, not large enough to solicit different preservation strategies and description methods for different user communities. Therefore, this requirement is complied with at more general level than is described in the OAIS standard.

A general conclusion by both partners to this project is that the OAIS has an inbuilt limitation in that it overly assumes both an identifiable and relatively homogeneous consumer (user) community. This is not the case for either the UKDA or TNA, both of whose consumers spread across a broad spectrum of knowledge and skills from the highly specialist and statistically literate to the skilled amateur and interested but non-specialist members of the public.

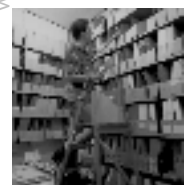
#### 4.8 OTHER RESPONSIBILITIES AND STANDARDS

When testing for compliance with the OAIS standard, one also has to consider other standards, legislation, regulations and guidelines that the institution has to follow. Both TNA and UKDA comply with an extensive range of international and national standards which, if necessary, would take preference over the OAIS standard, for example:

| STANDARD NO    | STANDARD                                                                                                                                                  | SUBJECT  |
|----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| BS 5454:2000   | Recommendations for storage and exhibition of archival documents                                                                                          | General  |
| ISO/IEC 17799  | A code of practice for information security management                                                                                                    | Security |
| BS 6266:2002   | Code of practice for fire protection for electronic data processing installations<br><br>Contingency plan for restoring a service after disaster recovery | Building |
| BS 7083:1996   | Computer rooms – Guide for accommodation and operating environment of information technology equipment                                                    | Building |
| ISO 15489:2001 | Standardisation of records management policies and procedures                                                                                             | Records  |

Both archives comply with all relevant UK legislation (Freedom of Information Act, Data Protection Act, Copyright, Design and Patents Act), and in the case of TNA also Public Records Act 1958 and the Lord Chancellor's Statutory Instruments. TNA is also compliant with the Government Secure Intranet (GSI) codes of connection (COCO), and other government regulation and accreditations covering information security and records management. TNA digital systems comply with the e-GIF (Government Interoperability Framework) and the e-GMS (Government Metadata Standard). Both organisations comply with SENDA, the legislation ensuring ease of use of web sites for the disabled. In addition, the UKDA operates within the JISC Information Environment requirements.<sup>41</sup>

41 [http://www.jisc.ac.uk/index.cfm?name=ie\\_home](http://www.jisc.ac.uk/index.cfm?name=ie_home)



The UKDA regulatory framework for managing its archival collection is composed of:

- a deposit agreement and licence form which confirms the rights and obligations of both parties and offers an opportunity for depositors to specify the conditions under which access may be given to third parties;
- assertion of copyright and IPR to ensure that the data creator has cleared all necessary copyright permissions;
- where necessary, negotiations for licence agreements with third parties to enable the UKDA to distribute the material for the purposes of research and teaching;
- access agreements and end user licences which specify the use conditions of the data collections.

One area of regulation which is of particular importance to both the UKDA and TNA but which is not discussed in the OAIS reference model, perhaps because it is considered out of scope, is the need to comply with legislation regarding web access for the disabled. This requirement is taken seriously by both organisations. At the UKDA, a dedicated web development team applies strict procedures for all web development, thus ensuring compliance with the legislation for all service sites and for the UKDA site.

The UKDA sites meet the following agreed best standards and good practice in web design and usability. The underlying code of the ESDS web site achieves compliance with level 3 (AAA) W3C guidelines for XHTML and Cascading Style Sheets, conforms to Web Content Accessibility Guidelines, and is SENDA compliant. The following XHTML and Accessibility Standards also apply:

- XHTML 1.0 Transitional;
- Cascading Style Sheets (CSS) Version 2;
- Web Content Accessibility Guidelines 1.0 (WCAG 1.0).

In addition, the UKDA service teams maintain a help desk which provides help and support to any users having difficulties using the web site.

TNA's new web site has been tested with text only browsers, screen readers and for keyboard-only navigation. Extensive usability testing took place before launch, and TNA is currently commissioning an independent accessibility audit. The web site:

- Is XHTML transitional;
- Uses CSS version 1 for old browsers and 2 for recent ones;
- Meets government guidelines on accessibility.

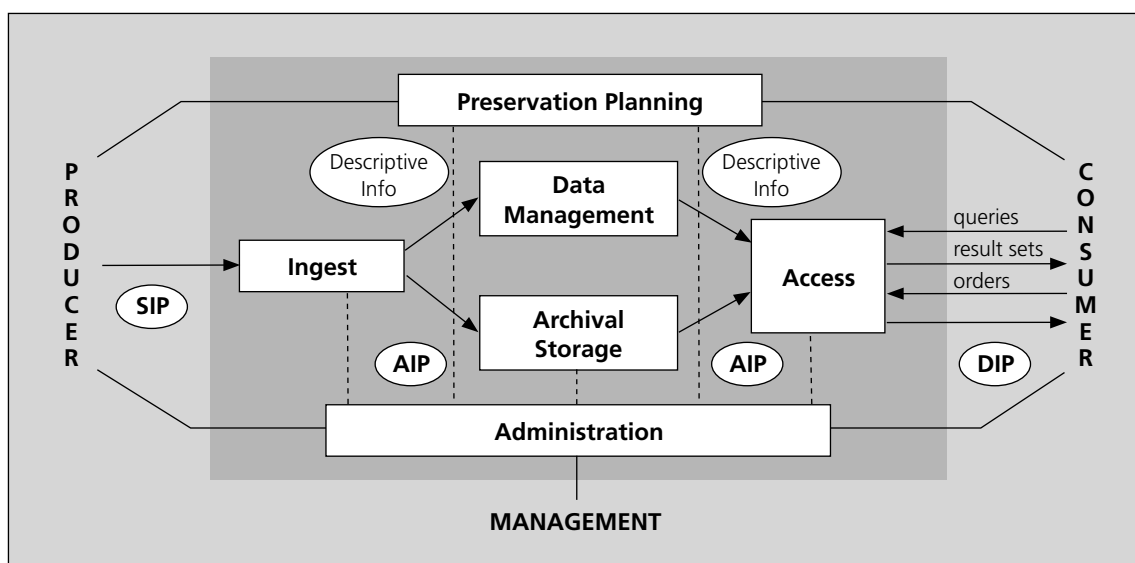
A further area of responsibility is the responsibility of the UKDA to the depositor of a dataset. Such responsibilities are set out in the Depositor Licence and relate mainly to Intellectual Property and statistical disclosure. This is managed through the EUL and through registration and authorisation processes. Further information about both the Depositor Licence and the EUL is included elsewhere in this document.



## 5 OAIS FUNCTIONAL ENTITIES

The OAIS reference model identifies and describes the core set of mechanisms with which an OAIS-type archive meets its primary mission of preserving information over the long-term and making it available to the designated community.<sup>42</sup> The OAIS functional model comprises six functional entities (represented as white boxes in Figure 1 below) and related interfaces, which together form a collection of six high-level services that fulfil the OAIS's dual role of preserving and providing access to the information in its custody. The information objects, or information packages (SIP, AIP and DIP) managed by these six services, are described below, in chapter 10.

It is worth repeating here, that the diagrams do not depict an architectural design of an archival system but represent a functional view of what an OAIS archive should do, regardless of the systems implemented and used. An OAIS archive will have to implement each of the six services or functional entities, in one form or another, in the course of developing its archival system.



**Figure 1. The OAIS Functional Model.**

*OAIS diagrams and textual descriptions are reproduced from the CCSDS 651.0-B-1*

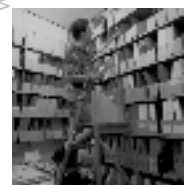
### 5.1 INGEST

The Ingest function, which in some domains is also known as acquisition, accession or authoring, is responsible for accepting the digital object submitted to the archive and preparing it for entry into the archive. The OAIS itself assumes that the traditional archival processes of appraisal and selection have been applied to digital objects prior to the decision to ingest them into the OAIS system and, also, that the negotiated agreements between content suppliers and the OAIS must be in place so that provenance, levels of control and collective description can be implemented on digital objects at the earliest possible time.<sup>43</sup> A separate standard *Producer-Archive Interface Methodology Abstract Standard* that accompanies the OAIS reference model details the process of negotiating for submissions and agreeing on ingest conditions.<sup>44</sup> As a consequence, the OAIS model does not cover pre-ingest activities and assumes that agreements are in place before SIPs are ready for transfer. In practice, elements of pre-ingest and ingest overlap at both the UKDA and TNA and are not easily separated. To ensure full comparison, the stages in workflow and functions described in this separate standard were also used for the OAIS compliance testing by each organisation.

<sup>42</sup> Lavoie, 2004, p. 7

<sup>43</sup> Lazorchak, 2004, p. 33

<sup>44</sup> CCSDS 651.0-B-1, 2004



## The pre-ingest activities

According to the standard, the pre-ingest phases for producer-archive interaction that lead to new material being accepted to the archive are:

- *the Preliminary Phase, also known as a pre-ingest or pre-accessioning phase, includes the initial contacts between the Producer and the Archive and any resulting feasibility studies, preliminary definition of the scope of the project, a draft of the submission information package (SIP) definition and finally a draft Submission Agreement;*
- *the Formal Definition Phase includes completing the SIP design with precise definitions of the digital objects to be delivered, completing the Submission Agreement with precise contractual transfer conditions such as restrictions on access and establishing the delivery schedule;*
- *the Transfer Phase performs the actual Transfer of the SIP from the Producer to the Archive and the preliminary processing of the SIP by the Archive, as it is defined in the agreement. The transfer and validation phases are often carried out partially in parallel, as there is iteration when all the information to be submitted is not submitted at once;*
- *the Validation Phase includes the actual validation processing of the SIP by the Archive and any required follow-up action with the Producer.*

## Pre-ingest activities: UKDA

The UKDA procedures for the workflow leading to ingest are as follows. The depositor completes a data submission form, which the UKDA Acquisitions Review Committee (ARC) then evaluates against stated acquisition policy criteria: usability of data; documentation for secondary use; and perceived user demand. The depositor will be sent an accept, reject or waived letter and the acquisitions database (MIRAGE) updated once a decision has been made. A work plan is created for each accepted resource that specifies, among other criteria, file formats of submission and the level of validation, cataloguing and indexing, and additional documentation that needs to be created, if any. Once the depositor has completed the deposit licence form and data documentation transfer form, the data can be sent to UKDA. Data are usually submitted via CD-ROM, floppy disk or e-mail, although submissions are accepted on a variety of storage media and through a number of transfer channels.

Only preliminary validation of the received data objects is conducted at this stage – an informal process that currently is not documented in a report although the Producer will be contacted if any gaps or omissions are discovered (e.g., in the documentation accompanying the data resource) and re-submission of the material requested. If no problems are found, the depositor is sent an acknowledgement for the materials by letter.

The material is then passed to the processing team where further checks are made and a thorough validation of the data occurs, before an archival information package (AIP) is created. The acquired data and documentation are thoroughly checked for anything that might cause problems for users (e.g., the variable lists are checked for completeness, descriptive statistics are produced and frequencies are checked). Validating the dataset consists both of checks on the data and internal metadata. For instance, one might check the data in terms of examining whether a variable that records people's ages contains values of less than zero or greater than, say, 100 years. In terms of internal metadata checks, this typically involves cross-referencing between the documentation provided by the depositor and the internal metadata in the data file, to check whether value labels and missing values in the data file accord with the depositor's documentation.



The UKDA currently uses four levels of processing to which slightly different validation requirements apply. Levels of processing are assigned by the UKDA's ARC in response to a number of criteria, by far the most important of which is the anticipated usage of the data collection over a foreseeable period. The level of processing assigned by the UKDA governs the thoroughness and complexity of the dataset validation procedures. It should be noted that processing levels can be up-graded at a future date, in response, for example, to changes in usage patterns.

A workflow diagram for UKDA pre-ingest and ingest processes is presented in Figure 2.

Although the OAIS and the Producer-Archive Interface standards are quite flexible in relation to precisely what type of validation of submissions the archive should perform and how it is to be done, the current UKDA practice seems to differ slightly from the workflow recommended by the OAIS.

While the OAIS reference model Quality Assurance function (see subsection *OAIS Ingest Function* below) recommends validating only the accurate transfer of a SIP, and a more thorough validation of the content of a SIP is required as part of the pre-ingest work,<sup>45</sup> the UKDA is using these processes in a "reverse order". When the OAIS standards describe the submission validation as discovering errors or non-conformities that would lead to rejection of the SIP, the UKDA data validation practice is directed towards more detailed documentation of the SIP and understanding its content and context. For the UKDA, this represents both a pragmatic and a more resource efficient approach. Although no separate formal audit report is produced as a result of the validation, when anomalies in data are discovered, the depositor is notified and when necessary asked to re-submit the data. An acknowledgment letter is sent to the depositor if no anomalies are found in the SIP.

### Pre-ingest activities: TNA

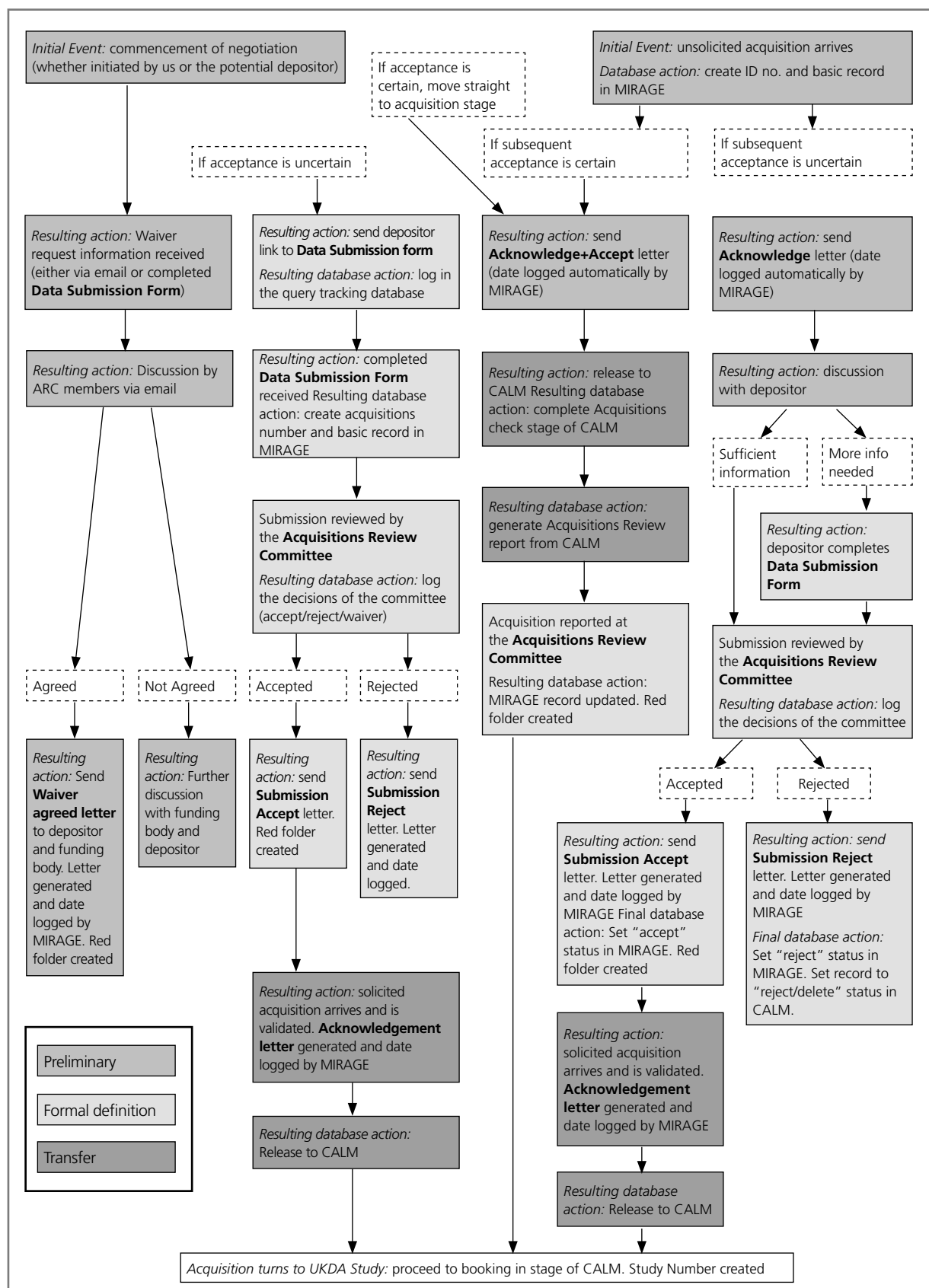
TNA has a team of client managers to manage the relationship with each department or agency through an appointed Departmental Records Officer, who are normally relatively senior managers in their department. Knowledge of potential archival material is regularly assessed. Records undergo an appraisal process, which may result in them being selected for permanent preservation at the archives. This is governed by formal appraisal and selection policies agreed in conjunction with each agency or department individually, as the nature and extent of the records of each are different.

Once a record has been selected, catalogue references are generated, in the context of previous accessions from that department, and catalogue descriptions are created. Records metadata and the extent and type of the digital objects will be assessed and entered into the transfer tool – a portable application to map digital objects and records metadata to catalogue levels, and to provide fixity and technical information. The transfer tool creates metadata in XML conforming to the e-GMS and allows the creation of these metadata in stages, as the process of assembling and verifying the record to the required standards may take some time to complete. Once all mandatory metadata are present and the department and client manager agree that the record is complete, this constitutes the SIP, which is normally burned onto TNA branded CD or DVDs.

Transfer involves sending the record by secure delivery to TNA, where it is loaded into a quarantine area and checked to ensure it is complete, usable and virus free. Management of digital records is governed and monitored by the Accessions group with representation of all parties involved in this process.

<sup>45</sup> cf. "Producer-Archive Interface Methodology Abstract Standard" (2004), ch. 3.2.2.4

Figure 2. Workflow diagram of the UKDA Acquisitions Section.





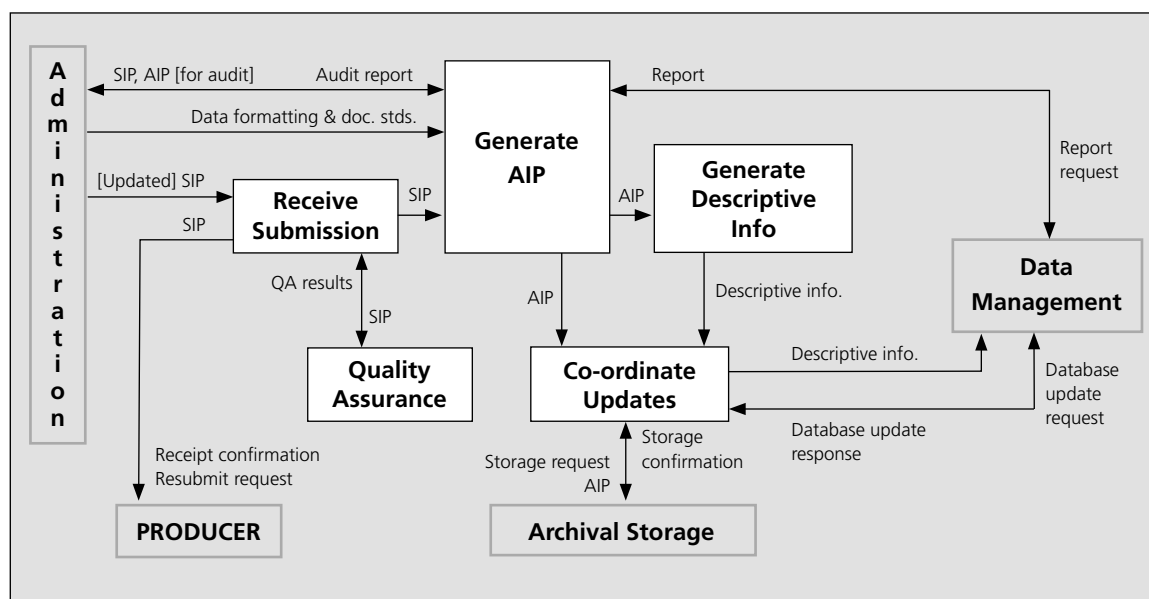
The pre-ingest workflow at TNA follows the *Producer-Archive Interface* standard methodology, with some minor discrepancies. For instance, TNA accepts digital submissions to the digital archive only on official TNA branded CDs and DVDs, although exceptions can be made by special agreement, for example where a high data volume might dictate transfer media. These media carry a serial number that is tracked against which Government Department it has been issued to and what records are expected to be on it. The initial validation provides checks against digital records that are not required for their paper counterparts, for example, file corruption and technical usability of the digital record. TNA Web Archive harvests its content straight from the public internet for Government departments with which an agreement exists to archive from their web site. The relevant Operational Selection Policy defines the update frequency.

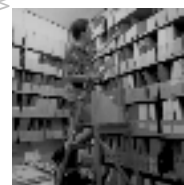
### OAIS Ingest Function

The functions of the OAIS standard Ingest entity that acts as the OAIS's external interface with information Producers are illustrated in Figure 3. The five services included in this entity manage: the receipt of digital objects transferred to the OAIS; validation that the information received is uncorrupted and complete; transformation of the submitted information into a form suitable for storage, preservation and management within the OAIS archival system; extraction and/or creation of descriptive metadata to support the OAIS's retrieval tools and finding aids; and transfer of the submitted information and its associated metadata to the archival store.

- the *Receive Submission* function provides the appropriate storage capability or devices to receive a SIP from the Producer (or from Administration). This function provides a confirmation of receipt of a SIP to the Producer, which may include a request to resubmit a SIP in the case of errors resulting from the SIP submission;
- the *Quality Assurance* function validates the successful transfer of the SIP to the staging area;
- the *Generate AIP* function transforms one or more SIPs into one or more AIPs that conform to the archive's data formatting and documentation standards. This function sends SIPs or AIPs for audit to the Audit Submission function in Administration, and receives back an audit report;

**Figure 3. The OAIS Ingest entity**





- the *Generate Descriptive Information* function extracts descriptive information from the AIPs and collects descriptive information from other sources to provide to *Coordinate Updates*, and ultimately *Data Management*;
- the *Co-ordinate Updates* function is responsible for transferring the AIPs to *Archival Storage* and the descriptive information to *Data Management*. Transfer of the AIP includes a storage request and may represent an electronic, physical, or a virtual transfer. After the transfer is completed and verified, *Archival Storage* returns a storage confirmation indicating (or verifying) the storage identification information for the AIP.

The materials that an OAIS archive receives already have some metadata associated with them and the archive will usually negotiate with depositors to receive the digital objects in a specific format and with a particular set of metadata attached. The quality of metadata and format of these deposited objects may, however, be out of the OAIS control – as in the case of TNA which can prescribe the required metadata but not the file formats it receives from its depositors. The OAIS archive may have to make changes to the digital objects once they have been ingested into the archive, to bring them under the archive's intellectual control and ensure their subsequent preservation and use. Once the archive has ensured that the submitted digital object satisfies all of the criteria for inclusion in the archive and the final AIP has been created from the originally submitted SIP, the processing workflow is handed over to *Archival Storage* and *Data Management* entities.

## Ingest: UKDA

The transfer of submission data to UKDA can happen via several channels. The UKDA accepts data and documentation on a variety of media, but the most popular are CD-ROM, DVD, 3½ inch floppy disc and email attachment. Other media can be and are occasionally accepted, including: ZIP/JAZ cartridge; Exabyte; 9-track tapes; and punched cards.

The received submissions and their documentation are equipped with checksums which are preserved alongside the preservation metadata files. A more thorough validation of the content, as described above, also happens at this stage. The UKDA does not only generate a preservation version from the SIP during Ingest, but also a dissemination version. See sample ingest processing workflow diagram, Figure 4.

The UKDA accepts data and documentation files in a variety of formats. Preferred file formats are those that are not platform or software dependent or that can be easily transferred to a suitable preservation format; however, most formats can be handled by the UKDA. The Data Services section works with depositors to recommend a suitable ingest format and to assist with the transfer of data and documentation to the UKDA. If necessary, it provides data conversion support for depositors.

The UKDA processes during Ingest and creates two main types of format for any given submitted dataset: preservation format(s) and dissemination format(s). All deposited data and documentation files are also retained in their original formats. Original copies of the data and documentation are kept so that the archive can refer back to them and also send them back to the depositor if requested. Preservation formats are formats designed to be 'open' and as platform, operating system and software independent as possible, so that the UKDA creates a permanent machine-readable version of the dataset, which will still be readable in future generations, whatever the future changes to computer hardware, operating systems and software might be. In contrast, dissemination formats are geared to the UKDA's function as a data disseminator. They are designed to allow users to have access to data in a format that is convenient for their types of use and expertise. Whereas preservation formats have been selected (as far as achievable) to minimise the need for migration of formats to maintain access and usability, dissemination formats will require periodic migration to match changes in software and user preferences over time.

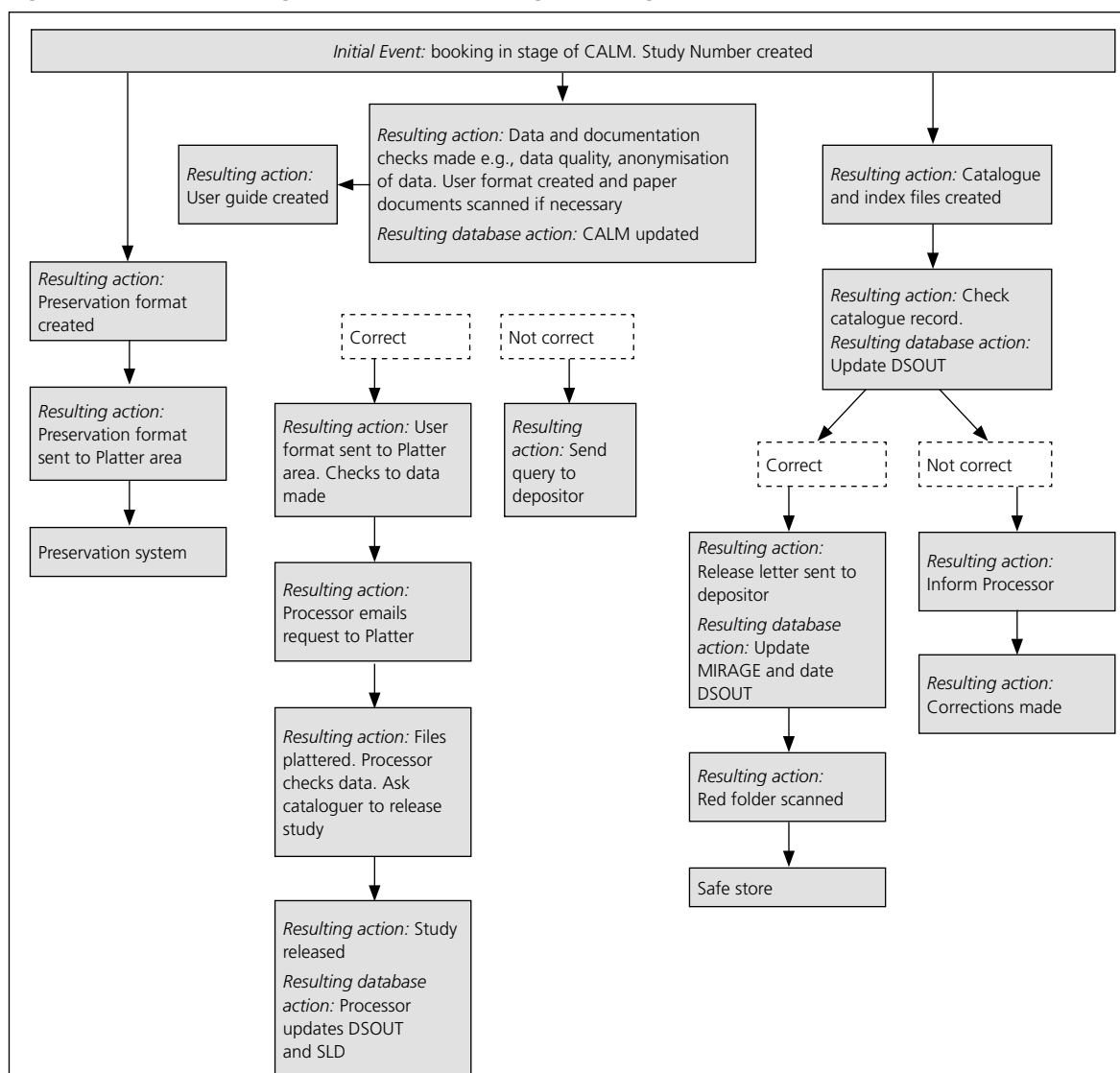


The two categories of preservation and dissemination format are not entirely mutually exclusive. For example, SPSS portable is a format that satisfies many users' needs but, in its current form, also comes close to being a pure preservation standard.

The UKDA's current preferred ingest, preservation and dissemination file formats are listed in appendix 2 of this report.

All of these versions, including the original SIP, are documented and a catalogue record is created for them. The UKDA currently publishes key data from its collection with their associated documentation and study description into the Nesstar system. An XML Generator is used to create an XML file that contains the study description and documentation for the study. These files are also archived in the preservation system. The OAIS Co-ordinate updates function is currently divided between several sections that are responsible for producing the AIP and DIP, and storing them, alongside their metadata, in the storage and catalogue systems.

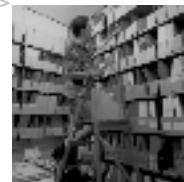
**Figure 4. UKDA Processing Section workflow diagram at Ingest**



#### Definitions

Note file: for internal UKDA use only. Contains a detailed history of the main processing events and outcomes.

Read file: intended for users and is supplied with each order. Includes UKDA processing level, notable features and outstanding problems.



## Ingest: TNA

TNA receives submissions on TNA branded storage media (CD-ROMs), each of which carries a serial number. After producing a security copy, the received submission is loaded into a pre-accession quarantine area, which exists on an isolated network, where it is virus checked and checked by the records management and digital preservation department to ensure that the record is complete. The record will remain in quarantine for four weeks. No alterations to the record are made at this stage; the original bit-stream is regarded as an essential piece of evidence supporting the presumption of authenticity. However, record metadata may be updated to reflect the fact that it has been received and that the checksums of the digital objects match the transfer manifest.

TNA cannot control the file formats in which data are delivered to them but it does control the quality of the associated metadata. Transfers without valid TNA metadata, represented as an XML file or a spreadsheet, which must include pre-assigned TNA catalogue references, are not accepted.

TNA Ingest validation phase checks that:

- the record is complete and valid;
- all objects meet the descriptions and checksums in the transfer manifest;
- TNA Records metadata are present and complete to the level required at ingest;
- all objects are certified as virus free;
- objects have been delivered on approved media.

Descriptive information (cataloguing) metadata are manually checked and amended if required. Once cataloguing is complete, the record is handed to the Digital Preservation department. The record is copied from the pre-accession server with its enhanced metadata and placed onto a loading client on another isolated network. The loading client once again validates and verifies that the data object it has received is correct, then will initiate an upload to the storage function (the Digital Archive system). The server performs additional validation checks and virus scans and, if the entire record is loaded successfully, a message will be returned signalling success. The record will appear in the archive under the catalogue references assigned in the metadata. It is not possible to overwrite material held by a current catalogue reference. All loaded material must have a unique reference which did not exist previously in the archive.

The Web Archive issues reports listing those web sites that it has crawled successfully and are catalogued in PROCAT under the reference ZWEB.

With the exception of the built-in reporting function at present, TNA ingest workflow matches the OAIS recommendation. The submission validation tools in some cases may be hand-crafted queries and scripts that extract the information required for validation. Most of the co-ordinate updates, data management and the archival storage functions are represented by a single application: the Digital Archive system.

Similar to UKDA, TNA creates dissemination versions of the SIPs at the Ingest stage and stores them in a separate archive system. Initial migration is limited to the production of presentation surrogates (e.g., for easier delivery within a web browser without requiring proprietary software to view the record).



## Ingest: Conclusions

A key conclusion is that, for both organisations, the pre-ingest functions are considered to be essential for efficient and effective archiving. The OAIS reference model would serve the community better if it included this function rather than relying on the existence of a separate model. The strongest argument for this is that pre-ingest activities are not without cost and they ensure quality, understandability and accessibility of information packages. It is a stated aim of this report that it will be of value to others who might want to undergo the process of validating their work against the model or to use the model in conjunction with the setting up of a digital archive. As it stands, the model does not, in the view of either the UKDA or TNA, incorporate the essential functions associated with pre-ingest activity and, as such, is leaving open the potential for problems later in the process.

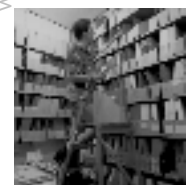
Both institutions testing for OAIS compliance are institutions that have been functioning as archives for a long time and, therefore, have a detailed set of policies and operating procedural manuals that cover their work with depositors. Due to the nature of the information producer community and the nature of digital objects submitted to the archives, there are slight differences in the way the submissions are negotiated, received and processed by either institution. The main groups of functions identified in the OAIS and Producer-Archive Interface standards are all demonstrably present but, again, with some discrepancies in the order and level of detail of services. By far the most time-consuming stages in the ingest process are the negotiation phase with the depositor, the quality assurance (with in-depth validation of data and metadata), generate AIP (and DIP, in the case of these institutions), and generate descriptive information.

One possible reason for the expenditure of time at the ingest stage is that both organisations have very clearly defined dissemination functions. The skilled researchers who form the majority of the UKDA user community expect meticulously detailed information about a DIP and thus the UKDA imposes stringent internal definitions on what is understood as 'independently understandable'.

From our reading of the reference model, it appears that OAIS assumes that DIPs are created on demand from AIPs. This is not the case for either the UKDA or TNA. For a short period in its history, the UKDA took this approach and preserved the DIP as a separate file, created on demand after the AIP. It soon became evident that this approach has serious drawbacks. This is because many of the UKDA AIPs are created by ephemeral research teams, making it difficult, as time passes, to contact members of these teams to resolve substantive and detailed queries identified during the creation of the DIP. It is also worth noting that in the past there was an assumption on the part of both the depositor and the UKDA that neither a DIP nor an AIP had to be completely understandable. This was evidenced in early versions of the Depositor Licence and the user agreement in which the user was required to notify the archive of errors in the data and, in turn, the archive was to inform the depositors of such notifications.

In the case of TNA, digital records are frequently between five and ten years old by the time they are selected for archiving, and will normally be in old data formats. This makes it extremely important to validate that the records are in a technically usable state.

By creating the DIP as part of the ingest process, it is possible to eliminate many such errors in co-operation with the depositor and thereby improve the understanding of the AIP. This is because the process of creating the DIP involves processes not essential to the production of the AIP but which are essential to ensuring that the AIP is independently understandable in the future. Essentially, the creator of the DIP is the first user of the AIP and only by using the AIP can one identify missing information at the detailed level required to meet the organisations' high standard for the definition of independently understandable.



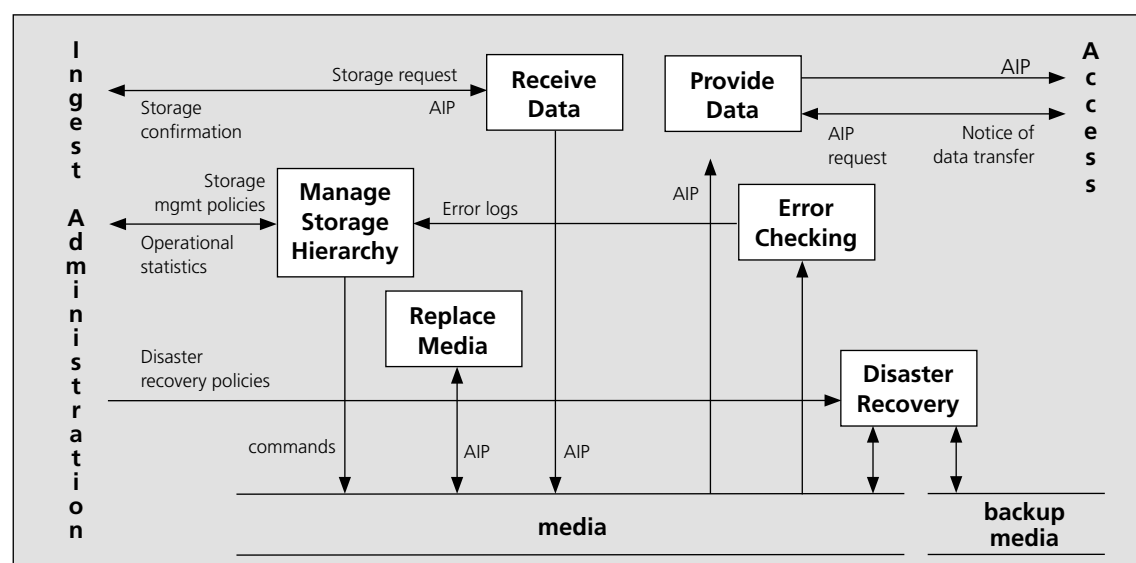
## 5.2 ARCHIVAL STORAGE

The second functional component of an OAIS archive is Archival Storage. This represents the part of the archival system that manages the storage and maintenance of digital objects entrusted to the archive. The Archival Storage function is responsible for ensuring appropriate types of storage, the appropriate structure of the file system, the necessary amount of storage available and other issues related to the physical management of data storage.

The first function of Archival Storage is to receive the AIP from the Ingest entity and to add it to the permanent storage facility. The Archival Storage entity oversees the management of this storage, including media refreshment, monitoring statistics of use and error logs to ensure all necessary levels of protection for the archived items. The Archival Storage is responsible for ensuring that the AIPs can be retrieved, as well as evaluating the outcome of preservation processes and disaster recovery policies to mitigate the effects of catastrophic events.

The final responsibility of the Archival Storage function is to provide copies of the stored AIPs to the Access entity and oversee data transfer operations to users of the archive.

The functions of this entity are presented on the following diagram.



**Figure 5. The OAIS Archival Storage entity functions.**

- the *Receive Data* function receives a storage request and an AIP from Ingest and moves the AIP to permanent storage within the archive. This function will select the media type, prepare the devices or volumes, and perform the physical transfer to the Archival Storage volumes. Upon completion of the transfer, this function sends a storage confirmation message to Ingest, including the storage identification of the AIPs;
- the *Manage Storage Hierarchy* function positions the contents of the AIPs on the appropriate media based on storage management policies, operational statistic, or directions from Ingest via the storage request. It will also conform to any special levels of service required for the AIP, or any special security measures that are required, and ensures the appropriate level of protection for the AIP. It monitors error logs to ensure AIPs are not corrupted during transfers. This function also provides operational statistics to Administration summarizing the inventory of media on-hand, available storage capacity in the various tiers of the storage hierarchy, and usage statistics;



- *the Replace Media function provides the capability to reproduce the AIPs over time;*
- *the Error Checking function provides statistically acceptable assurance that no components of the AIP are corrupted during any internal Archival Storage data transfer. This function requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard error logs that are checked by the Archival Storage staff;*
- *the Disaster Recovery function provides a mechanism for duplicating the digital contents of the archive collection and storing the duplicate in a physically separate facility;*
- *the Provide Data function provides copies of stored AIPs to Access entity. This function receives an AIP request that identifies the requested AIP(s) and provides them on the requested media type or transfers them to a staging area. It also sends a notice of data transfer to Access upon completion of an order.*

### Archival Storage: UKDA

The UKDA has a number of preservation systems in place that provide the storage facilities for all SIPs, AIPs and DIPs. The main preservation file system is based on an AMASS® HSM system where the files appear to be local to the user but are mainly based on tape. The system provides 3.3 TB of native storage capacity with an expected capacity of 8 TB using lossless drive compression. The HSM is using a 30 slot SDLT220 library with 3 SDLT drives. Additional, unlimited capacity can be gained from external media which can be operator loaded on system request.

The current UKDA storage solution keeps up to six copies of the same data file on at least four separate preservation servers:

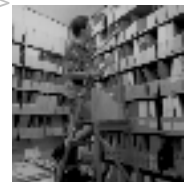
- main nearline copy (on main preservation server);
- shadow copy (on main preservation server);
- access online copy (on mirror preservation server);
- nearsite online copy (nearsite mirror preservation server);
- offsite online copy;
- CD-ROM offline copy.

A CD-ROM copy is created for each submission as part of the preservation procedure. This allows convenient access to an alternative local copy in the case of downtime of the main preservation system and serves as an alternative short to medium-term storage medium. The UKDA is currently testing selectively, the use of DVD-Rs as replacement storage media for the CD-ROM copy. Whilst DVD-R is no longer a 'new' medium, it is still considered experimental due to its possible fragile nature. UKDA is also looking at protective shields for CD's/DVD's<sup>46</sup> and highest grade, ruggedised media.

Error checking is internal to the preservation system and a range of recovery measures, designed to meet consequences of disasters that result with a loss of stored data files have been put in place. The UKDA's IT architecture for safe-guarding and managing its data collections is built on these basic preservation principles:

- physical reliability of storage media and servers;
- security of the data from unauthorised users;
- usability of data formats in the future.

<sup>46</sup> A transparent plastic sheet to provide additional protection.



The UKDA IT systems employ various disaster prevention measures, including restricting preservation server access to key personnel, routinely monitoring storage media and implementing a high degree of server redundancy to ensure the isolation of server malfunction or loss of data should they occur. The UKDA's disaster recovery procedures predominantly focus on reacting to scenarios of hardware and software failure, corruption or degradation of storage media and potential problems caused by human error. A general overview of responsibilities for disaster recovery and business continuity planning is outlined in the UKDA Information Security Management Policy.

In the OAIS model the Provide Data function responds to requests from the Access entity by providing AIPs from the storage system. At the UKDA this is managed differently. Users do not gain access to AIPs, nor do user requests address the AIPs in the storage system. Instead, the access requests retrieve a previously created and stored DIP from the storage system, which is currently residing on a separate server from the main preservation system. The AIPs are considered to be for preservation purposes only and are kept mainly as archival masters, with as little use as possible. For some of the older datasets held in the UKDA, AIPs were not created on deposit as they would be now. In the instance of a request for such a study, an AIP will be created by conversion from the SIP, which is also stored in the main preservation system. The newly created AIP will also be lodged in an area and added to the automated download system to enable immediate future access for registered users.

### Archival Storage: TNA

TNA's Digital Archive system is served by a FileTek HSM system on which the AIPs are loaded. The HSM duplicates the data internally on to two separate tapes in the library. The entire archive is backed up each day onto a third set of tapes which is rotating with the cycle of five and the backups being stored in a fire-safe in another building. The HSM system manages the file storage and error messaging internally by conducting bit-level integrity checking. Integrity checks are made by the FileTek system whenever a record is accessed, which at present is every night due to the backup regime in place. If the integrity check fails, the system can be recovered from the two copies of data that reside within the system, or from the daily offline backup copies.

Additionally, there are two archive systems – a closed archive and an open archive. The closed archive system contains all the archive's holdings and is kept for preservation purposes on a strictly isolated network and the open archive system only keeps the records opened for public access (currently on TNA's internal network). In total, there is a minimum of seven copies for each AIP if the record is closed, and fourteen if the record is open, due to the duplication between the closed and open systems. This ratio may shift in future, as TNA moves towards an incremental backup regime as the volume of data in the archive increases. The Web Archive stores its information on hard disks that are mirrored and backed up to tape, with three copies of the AIP stored in total.

Disaster recovery and business continuity planning are required for each TNA department and must conform to National Archives plans and standards. Disaster recovery plans cover the complete rebuilding of digital archiving facilities and data recovery from backup. They do not cover actions for faulty or lost storage media as this is not regarded as being at a disaster level: the backup and media checking regime, combined with a high number of redundant media copies, provide operational assurance of data integrity and availability and are covered by standard operating procedures.

As at the UKDA, TNA access system is entirely isolated from the storage function. The DIPs are created and stored separately, on separate systems. The preservation system does not handle user access requests at all – this is handled by a separate presentation system. If a copy of a data object within the archive is required, the Digital Archive system can supply the files of the original record, or any other manifestation, as a zip archive for download via a built-in management web server but this function is not accessible to the public. Digital



preservation staff alone have access to the storage function. The files in the archive are copied via removable storage devices to public-facing systems and presentation surrogates are produced and loaded into the presentation system.

### Archival Storage: Conclusions

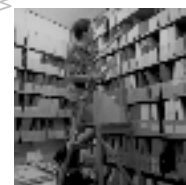
As described in the OAIS reference model, the Archival Storage function is a base-level service with a limited number of responsibilities that can often be resolved with the help of technology. The tasks this entity has to perform are either internal to its functioning (e.g., media refreshment) or related to input/output with other entities. The Archival Storage function has no direct external interface and interaction with Archival Storage is confined to the OAIS's internal high-level services. The OAIS Archival Storage layer has, thus, very little knowledge built into it and this knowledge only takes place at the interface of Archival Storage with other services.

Both UKDA and TNA rely on advanced technology for their storage management and media checking regime but also use a policy of multiple copies for every AIP to reduce the risk of data loss. However, the storage technology used incorporates some elements of the OAIS Administration entity when it comes to managing the storage and disaster recovery. The function of managing the storage hierarchy, too, has some 'knowledge' built into it, in order to separate the preservation and dissemination storage areas and to keep the integrity of the AIP and DIP versions of the same SIP together.

The strict separation of the AIP storage from the DIP storage is the biggest discrepancy here from the Reference Model, with the Access entity only interacting with the DIP storage, rather than the AIP storage, as recommended by the OAIS standard. In practice the access to AIPs and their storage system is strictly controlled in both the UKDA and TNA. The reason for this is that both TNA and UKDA are following archival tradition in that the approach is in line with standard archival practices in which original document records are often copied (usually to microfilm) and only the copies are made available to users, unless there are exceptional circumstances. This important mechanism to protect the archival version of the original record seems to be lacking in the OAIS reference model.

One of the functions in the Archival Storage entity that both institutions found might be made more use of, is the advance storage space planning. The nature of the UKDA ingest process generally allows for sufficient time for planning of storage space for electronic media but sometimes metadata may be accepted in hard copy, thus requiring digitisation. In such cases a check on physical storage facilities would have to be made in advance. The main UKDA systems are funded under the ESDS and AHDS History service contracts. The former include estimates of numbers of SIPs and their sizes but this is not the case for the latter. Consequently, for the ESDS, storage space is guaranteed over five year periods, after which it is renewed with a one year overlap of the old and new systems. If an exceptionally large SIP was to be accepted, the storage requirements could be accommodated easily although a problem might arise if the SIP required a specialised DIP. In this case, the production of the DIP would probably be treated as a special project with a request for a dedicated server.

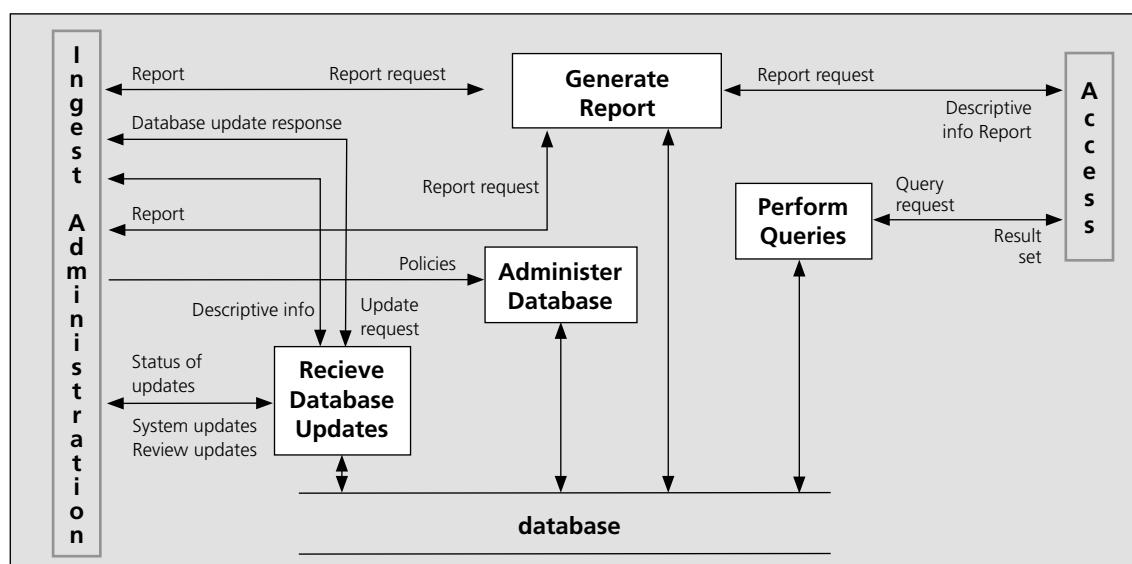
TNA is also constantly monitoring its capacity planning. Normally, TNA are aware of the type of material that will arrive in any one year, as they are always in negotiation with content producers. However, sizes of digital records vary considerably. Many records exist as simple sets of documents or emails but some individual records, for example a public enquiry incorporating over one hundred hours of digital video, can be individually larger than the combined accessions to date. Therefore the requirement is for a high degree of scalability within reasonable but quite large limits.



### 5.3 DATA MANAGEMENT

The third functional component of an OAIS archive is Data Management. It operates in conjunction with the Archival Storage entity and maintains databases of descriptive metadata identifying and describing the archive's holdings. It supports the OAIS's external finding aids and also manages the administrative metadata that supports the OAIS's internal system operations. Data Management functions include administering and maintaining the archive database; performing database updates as new information arrives or existing information is modified or deleted; performing queries on the databases to generate result sets; and producing reports in response to requests from other functional components within the OAIS.

The functions of this entity are presented in the following diagram.



**Figure 6. The OAIS Data Management functions.**

- the Administer Database function is responsible for maintaining the integrity of the Data Management database, which contains both Descriptive Information and system information. Descriptive Information identifies and describes the archive holdings and system information is used to support archive operations. The Administer Database function is responsible for creating any schema or table definitions required to support Data Management functions; for providing the capability to create, maintain and access customized user views of the contents of this storage; and for providing internal validation of the contents of the database;
- the Perform Queries function receives a query request from Access and executes the query to generate a result set that is transmitted to the request;
- the Generate Report function receives a report request from Ingest, Access or Administration and executes any queries or other processes necessary to generate the report that it supplies to the requester;
- the Receive Database Updates function adds, modifies or deletes information in the Data Management persistent storage. The main sources of updates are Ingest, which provides Descriptive Information for the new AIPs, and Administration, which provides system updates and review updates. The Receive Database Updates function provides regular reports to Administration summarizing the status of updates to the database and also sends a database update response to Ingest.



## Data Management: UKDA

The archival information systems within the UKDA have been developed over time and were built in a modular fashion where each component was developed and used in isolation. The process of integrating these into a centralised management system has been a major priority in recent years and now allows once-only input of common data about the archive's holdings for use in several systems. The decision to develop the system in-house was made because, at the time, there were no packages available to fulfil all the needs of a digital archive. The decision also offered the flexibility to control and quickly update the system to specific and developing requirements.

Currently, metadata management is split into three different parts. Basic information about each SIP is captured at ingest in a database known as CALM. Catalogue metadata are also produced during ingest, in a separate database and the process continues as the DIP and AIP are created. The information systems section of the archive has strategic responsibility for this work and contributes heavily to the development of new programs and tools to assist the process. CALM and the catalogue are currently not integrated but work is underway to remedy this. Links between the preservation metadata management system and the catalogue metadata system permit daily updates in a two-way, synchronised process.

As part of the value-added work of the UKDA, some SIPs will be enhanced with extremely detailed metadata using tools that have been developed in-house. For example, programmes have been written to extract variable and value information from several commercial statistical software packages, outputting the results into standard XML files. These can then be used within the Integrated Data Catalogue and the Nesstar system to facilitate resource discovery and data browsing. The database can be, and is, used to produce detailed regular reports, for example on the number of new datasets from particular depositors and the time lapse, per dataset, from acceptance into the collection to release.

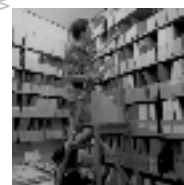
In addition, the catalogue database is linked to the user database which permits the output of statistics on, for example, the use of individual datasets, the use of DIPs at a particular institution or organisation, or the number of datasets deposited by organisation. Such reports are used as performance indicators to funding bodies or as a means of meeting contractual requirements to depositors where they require information on usage of their material.

Regular performance indicators, known as SLD's (Service Level Definitions) are created from the CALM database and include usage figures, response times for queries, the number of data collections acquired and numbers of user registrations. Data are also collected from MIRAGE to indicate how many acquisitions have been accepted, rejected or waived. These figures may be used to review the acquisitions and data collection policies.

Any paper documents relating to a SIP are digitised and attached to the AIP. The paper originals are held in a storeroom meeting the standards required by TNA (BS 5454).

The updates to existing holdings and their metadata are handled by the system in two forms:

- complete replacement – the earlier version of the data object is archived under the date of the time of the update; the newer addition is uploaded and becomes the master data; preservation and metadata information is updated as part of this process;
- amendment to a specific part – almost the same as above but with the exception that the earlier edition is not kept, only the required file(s) and metadata are updated.



Very exceptionally, a depositor will request that a dataset is removed from circulation. In such instances the study would be de-catalogued so that the public view of the catalogue will reflect the fact that the study is no longer available. Nevertheless, the study will remain visible to internal users and the files will remain in the preservation system as it is considered too expensive to remove them and physical removal would present unacceptable risks to other parts of the collection.

### **Data Management: TNA**

At TNA the catalogue system PROCAT is the authoritative source for descriptive information – catalogue data.

Since the preservation and presentation systems are disconnected, only the dissemination copies are accessed via the online catalogue system (PROCAT) or more directly via Electronic Records Online, which also contains a duplicate of the digital records cataloguing information. The descriptive information is updated rarely and the updates are managed in the catalogue management system PROCAT. The Digital Archive receives descriptive information cataloguing updates from PROCAT via removable media.

Records metadata in the Digital Archive are described in an XML file for each record, which includes all categories of OAIS metadata, although the mapping is not perfect (see UML diagram and metadata discussion later). The XML file and associated digital objects making up the AIP are loaded into the Digital Archive through a Java loading client, only accessible to approved Digital Preservation staff via a dedicated client in a physically secured area. The loading process is atomic; records will only be committed to storage if the entire loading and validation process is successful. It is stored directly as XML within an Oracle 9i database with XML extensions, although parts of the XML are also explicitly copied into relational tables for faster access by the web application. Bit level integrity checking is performed directly via the HSM system whenever a record is accessed.

Reports from the Digital Archive can be generated to show what records have been accessioned, their closure status, the XML metadata for each record and the operational status of the archive systems. More specific reports are hand-crafted through direct SQL queries and scripts to read the XML, for example to analyse the distribution of file formats throughout the accessioned material.

As can be seen, there is no single unified metadata database from which to report. Much of the Digital Archive system must co-operate with other archival systems at TNA, for managing Freedom of Information access restrictions (SAR), the unified catalogue (including physical records) PROCAT, the EROL system and the storage systems (Digital Archive, Web Archive).

Broadly speaking, the Administer Database function maps to the administration of the PROCAT system, with Descriptive Information updates feeding into the Digital Archive which maps into the Receive Database Update function. Although the Generate Report function is clearly functionally mappable to TNA, it is highly distributed around the various systems and functions of the archive.

The Perform Queries function maps to searching the PROCAT catalogue, although unlike the OAIS model, Access function does not directly query the archival storage, and neither are result sets or user views of the data generated on demand by such a process.



## Data Management: Conclusions

It is clear that an OAIS requires several layers of rich metadata to assure the usability and understandability of the digital objects it keeps over the long-term and a flexible database system is essential for ensuring that those complex metadata architectures can be implemented and maintained. While some of the metadata management tools required for archivists can be built into the technological database solutions, the access to the database for external users will have to be customised according to user community needs.

Both participating institutions are archives with long histories and, therefore, have a long-standing tradition of describing their holdings. In the case of TNA, the description of the new digital material has been integrated into the archive's main catalogue system PROCAT; the UKDA has used an electronic catalogue for many years. The UKDA catalogue is now both Dublin Core and Data Documentation Initiative (DDI)<sup>47</sup> compliant. The latter is used to describe the context, content and structure of the collections. Database systems developed for managing these metadata are quite specific but do provide management tools for both users of the archive and archivists who are responsible for the holdings.

It is also worth noting, for the benefit of others who might wish to do so, that mapping the OAIS data management functions results in an 'explosion' of mappings to all the different systems and processes that an archive performs. This is one of the ideas in OAIS that looks simple but has proven to be highly distributed in practice.

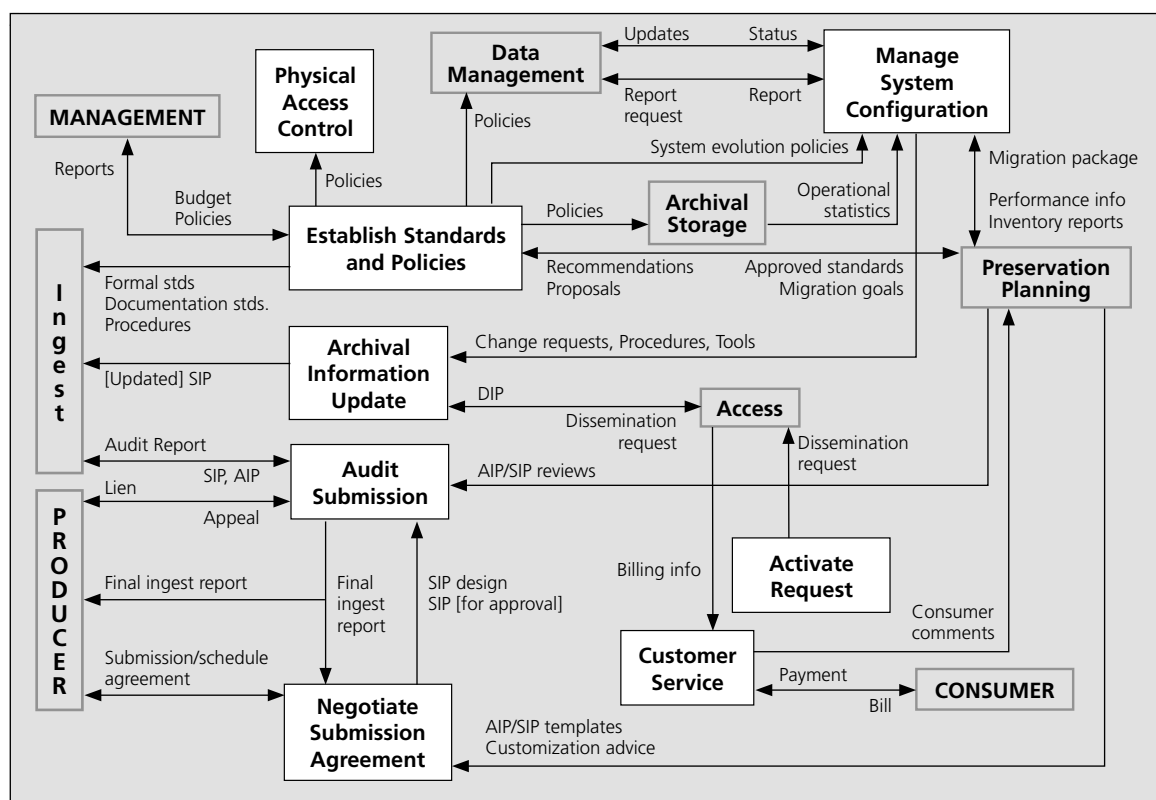
## 5.4 ADMINISTRATION

The fourth functional component of an OAIS archive is Administration. Every archival institution has a management that is responsible for establishing standards and policies, controls the physical access to the premises, provides customer service, negotiates with depositors for materials to be submitted and performs other management functions. These housekeeping functions are essential to the operation of any archival institution but the OAIS Administration entity is not only a co-ordinating entity, it has direct contact and oversight over all the other functional entities. Thus, Administration serves as the central hub for the OAIS's internal and external interactions: it communicates directly with the five other OAIS services – Ingest, Archival Storage, Data Management and Access, as well as the OAIS's external stakeholders – Producers, Consumers and Management.

Administration functions include soliciting and negotiating submission agreements with Producers, auditing submissions to ensure that they meet archive standards and maintaining configuration management of system hardware and software. It also provides system engineering functions to monitor and improve archive operations and to inventory, report on, and update the contents of the archive. It is responsible for establishing and maintaining archive standards and policies, providing customer support and activating stored requests. The Administration function is also responsible for overseeing the operation of the archiving and access systems and monitoring system performance. The Administration entity also has the customer service function, which creates, maintains and deletes customer accounts, as well as providing billing services.

The functions of this entity are presented in Figure 7.

47 Data Documentation Initiative (DDI) An international standard for metadata for data. See: <http://www.icpsr.umich.edu/DDI/>



**Figure 7. The OAIS Administration functions.**

- the *Negotiate Submission Agreement* function solicits desirable archival information for the OAIS and negotiates Submission Agreements with Producers. This function also negotiates a data submission schedule with the Producer. This function receives AIP/SIP templates and customisation advice from Preservation Planning and sends SIP designs and SIPs to the Audit Submission function as part of the submission approval process;
- the *Manage System Configuration* function provides system engineering for the archive system to continuously monitor the functionality of the entire archive system and systematically control changes to the configuration. This function maintains integrity and tractability of the configuration during all phases of the system life cycle. It also audits system operations, system performance, and system usage. It sends report requests for system information to Data Management and receives reports, it summarises those reports and periodically provides OAIS performance information and archive holding inventory reports to Preservation Planning;
- the *Archival Information Update* function provides a mechanism for updating the contents of the archive. It receives change requests, procedures and tools from Manage System Configuration. It provides updates by sending a dissemination request to Access, updating the contents of the resulting DIPs and resubmitting them as SIPs to Ingest;
- the *Physical Access Control* function provides mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the archive, as determined by archive policies;



- *the Establish Standards and Policies function is responsible for establishing and maintaining the archive system standards and policies. It receives budget information and policies such as the OAIS charter, scope, resource utilization guidelines, and pricing policies from Data Management. It provides Data Management with periodic reports. It receives recommendations for archive system enhancement, and proposals for new archive data standards from Preservation Planning. It also receives performance information and archive holding inventories from Manage System Configuration. Based on these inputs, archive standards and policies are established and sent to other Administration functions and the other functional entities for implementation;*
- *the Audit Submission function will verify that submissions meet the specifications of the Submission Agreement. This function receives AIP/SIP reviews from Preservation Planning and may also involve an outside committee;*
- *the Activate Requests function maintains a record of event-driven requests and periodically compares it to the contents of the archive to determine if all needed data is available. If needed data is available, this function generates a dissemination request that is sent to Access;*
- *the Customer Service function will create, maintain and delete Consumer accounts. It will collect billing information from Access and will send bills and collect payment from Consumers for the utilization of archive system resources. It will respond to general information requests. This function will also collect and respond to feedback on Access services and products.*

### **Administration: UKDA**

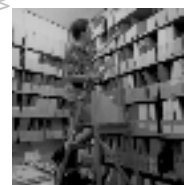
Currently, the major services provided by the UKDA are the ESDS, the AHDS History and TNA services which have been described at section 1 above.

Within the UKDA, work is guided by a number of standards and policies that have a bearing on the planning and management of the preservation function. The overall strategic planning and general policy issues are the responsibility of the ESDS and AHDS Management and the Director of the UKDA, via the Management Services Section and the management teams within the UKDA.

For these services the Administration function lies, contractually, with the University of Essex. The University appoints, and delegates responsibility for the service administration to, the Director of the UKDA. The Director takes personal responsibility for TNA services and for the Management Services Section which manages the strategic, financial and personnel elements of the organisation. The Director is supported by the Senior Management Team (SMT) which comprises the Associate Directors and Heads of Service who have responsibility for: the AHDS History Service; Systems, Preservation and Information; Data and Support Services; and Outreach and Training (which includes responsibility for acquisitions). The Associate Directors and Heads of Service report to the Director on a regular basis and have specific responsibilities for functions affecting their sections. The Director also has the support of an internal management group, the Director's Advisory Group (DAG).

Externally, the Director is supported by a number of service specific Advisory Boards which comprise representatives of the research community, funders and major depositors. The Director has a proactive role in setting strategy and was, for example, consulted about the content of the Economic and Social Research Council (ESRC) Datasets Policy<sup>48</sup> because of its strategic impact on the acquisition policy of the UKDA.

48 [http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/Annex%20C\\_tcm6-9738.pdf](http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/Annex%20C_tcm6-9738.pdf)



Consequently, internal working policies and procedures, for example, the acquisitions, storage management and the preservation policies, are the responsibility of the relevant Associate Directors in agreement with the Director and DAG.

The choice or application of standards can either be determined by contractual obligations (e.g., BS 5454 for TNA), by legal obligations (e.g. SENDA requirements) or it can be advisory from within sections and adopted through the UKDA's decision-making structure. It should be noted that, as a matter of principle, the UKDA aims to apply open standards where these exist. Exceptionally, it will work in conjunction with other archives or professional bodies to develop its own standards but only if the requirement is specifically focussed and directly relevant to the UKDA. This has been the case with the DDI. This standard, developed in conjunction with the network of data archives, is now available to the community as a whole for further development and wider adoption.

Thus, although the Administration function at the UKDA is very clearly vested in the Director who maintains direct financial control, the specific functions are spread across the organisation.

Submissions to the UKDA are reviewed and approved by the ARC which meets fortnightly and reviews all new acquisitions or submissions that have been received in the period since the last meeting. Membership of ARC comprises representatives from both the Outreach and Training, and Data and Support Services sections. The committee is chaired by the Associate Director, Outreach and Training. When the need arises, expertise is sought from the relevant Advisory Board or relevant subject experts. The ARC co-ordinates and develops all management policies and issues that relate to the Acquisitions functions of the UKDA. The Acquisitions Review Process has its own Processing Guide and a work plan is created for each 'accepted' submission. This plan specifies:

- that all files will be preserved in their original format;
- that all files will be converted to the appropriate preservation format, if necessary;
- the additional data formats or versions in which the data and documentation will be made available;
- the composition of the user guide for each resource;
- the level of validation, cataloguing and indexing, and additional documentation that needs to be created, if any.

At the submission stage, there is no formally signed Submission Agreement although depositors do supply a data submission form. Instead, on acceptance by the ARC, a legal agreement, the Licence form, is drawn up, by which the data are deposited formally whilst the depositor retains legal ownership of the dataset. This is an instance where the UKDA differs from the OAIS reference model. The licence gives the UKDA the right to process the data for preservation and dissemination whilst also keeping the original deposited dataset intact. Data transfer and submission are both handled by the acquisitions staff within the Outreach and Training Section. This section also works closely with data depositors to provide guidance and advice with regards to data creation and deposit. See also the workflow diagrams at Ingest entity above.

The UKDA runs a series of service specific customer help desks which accept queries by email, post, telephone and fax and which provide guidance to customers on finding, accessing and using data, as well as user support, registration support and user training. The help desks also manage the receipt and throughput of orders.



Requests for the ESDS and AHDS History services are entered into a database via the UKDA online access and ordering system. The database includes name, address and user type and also keeps a record of which studies the user has downloaded or requested and for which purpose (users have to register usages indicating how they intend to use the data). Other information recorded includes the usage number, the study number, format and date of download/request. This information is available to staff via a web interface to the database and queries can be run on the database to produce usage reports. For requests that require manual intervention, (including requests from commercial users or for commercial purposes, requests to prepare data for download or requests for data on CD-ROM), the system adds them to a list of outstanding orders viewable through the web interface. This list is checked daily and the orders processed as necessary. The orders can be assigned to different members of staff who can assign a new status to the order (e.g. 'with data delivery team' or 'in progress') which is reflected in the user's order tracking facility in their online account. Once the order is ready, it is marked as sent, the sent date is recorded in the database, the user is advised by email and the CD-ROM is despatched.

As noted previously, the UKDA does not provide access to stored AIPs to the general public. Users receive DIPs that have been prepared at ingest or are created for dissemination on request. This aspect of the workflow differs from the general OAIS functional model.

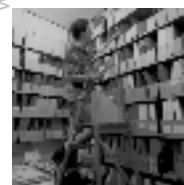
The majority of the UKDA users are academics, teachers and students in the UK HE/FE sector and can obtain the data from the UKDA at no cost, from the download system. Charges do apply for data requested for commercial purposes but such requests are infrequent. If users request data on CD-ROM or other media, in preference to automated downloading, a small handling charge is made. Accounting and record keeping fall under the remit of the Management Services Section.

The Physical Access Control function provides mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the archive, as determined by archive policies based on the most stringent requirements of the various services managed by the UKDA. Consequently, access to the UKDA building is restricted in accordance with TNA requirements and responsibility for maintenance lies with Management Services working closely with systems and preservation, and acquisitions staff.

The Head of Systems and Preservation and the Systems Manager manage the use, resilience and configuration of the preservation system. Specifications for all systems within the archive are configured and designed with a specific life-cycle from the outset. These are based on details of expected growth and usage of the archive's collections. The normal life-cycle for preservation systems is based on a five year cycle. The first year is run in combination with the outgoing system and the last year is run in combination with the incoming system. The preservation system is built up from various systems which have been designed to have a specific purpose. Each part of the system is automatically monitored individually for operation, performance and usage, in line with what it has been planned for. If it is found to deviate outside these values then the issue is investigated and a decision is made on how to remedy it.

The UKDA is committed to taking all necessary precautions to ensure the physical safety and security of all data collections that it preserves. Access to areas where electronic material and paper collections are stored is restricted to key preservation staff only. Storage rooms are protected by internal alarm systems that have local audible alarms and are linked to the central security section of the University of Essex.

The repository rooms are also protected by Argon gas-based, fire-extinguishing systems. This includes a four-zone fire detection system with automatic release that is activated when fire or smoke is detected within two of the four zones. Rooms are protected by a one hour secure fire door. Repository staff are instructed in the location of fire extinguishing appliances and in their use.



As mentioned above, the implementation and maintenance of security and access requirements across the UKDA services, is an administrative responsibility, managed by the Director. Once in place, day-to-day responsibility lies with the relevant section, for example, systems staff monitor access to the machine rooms and the functioning of the alarm systems whilst acquisitions staff manage physical access to TNA store.

## Administration: TNA

Manage System Configuration is handled by the systems section of the Digital Preservation Department. The preservation system configuration consists of several systems. Each system has its own diagnostic capabilities which are monitored by the systems administrator. Currently these are not unified into a single reporting framework. The migration planning facility is handled by the Technology Watch function within the Digital Preservation department, which uses PRONOM (a file format database accessible to all on the web<sup>49</sup>) and other internal tools.

Negotiate Submission Agreement and Audit Submission is handled via Client Managers from the Records Management department. They work in conjunction with the depositor and technical staff from the Digital Preservation department to ensure that the submitted records are viable, in terms of integrity and long-term preservation and presentation viability. Records are only accepted on approved media and are listed in a transfer manifest, either captured on a spreadsheet, or in an XML document. All mandatory technical and archival metadata must be present for TNA to accept the submission. These metadata are captured by the transfer loader client.

Depositors must currently provide a copy of the AA2 form on transfer for both paper and digital records.

The Archival Information Update function is handled by a combination of the loader client, PROCAT editorial and the Digital Archive. All of the received digital records and metadata are validated for completeness, integrity, metadata standards and virus checking. This includes a review of cataloguing metadata. Once agreement is reached that the record is complete and documented to appropriate standards, the record will be submitted to the ingest function and loaded into the Digital Archive.

Once the records are in the Digital Archive system, their updates are handled through the PROCAT catalogue updates to the archive. This function will also control closure status of records if they were to change (upon a Freedom of Information request, for example). However, due to the separation of DIP and AIP management, this function does not interact with access directly, although if a copy of an AIP is required, this can be provided.

The Activate Requests function is via the reading rooms,<sup>50</sup> although, as indicated previously this does not result in a request to Archival Storage. The Customer Service helps users with locating records, training in TNA's systems, providing record copies, etc. If a user requests a record which is closed, this is treated as a Freedom of Information request and must be tracked and reviewed. Legislation sets strict limits on the nature and timing of the response to this kind of query. If the request is for material without access restrictions, but, for example, delivered on other media or transformed into other formats, this work would be reviewed to determine whether a charge would apply. Some records cannot be delivered via the Internet, and special arrangements may be made to view those records that require specialised software.

Some digitised records are only available by payment from TNA. Payment for some types of born digital record is under review. The system is managed by an e-Commerce system and is entirely online, although payment can be accepted manually if required.

49 <http://www.nationalarchives.gov.uk/pronom/>

50 <http://www.nationalarchives.gov.uk/visit/plan.htm>, <http://www.nationalarchives.gov.uk/visit/pdf/rules.pdf>



There is deliberately no centralised user accounts management system at TNA, and access to records by user is not tracked. User accounts can be created for payment for certain digitised collections. Closure status of the records is tracked and managed by the SAR system. SAR is the System for Access Regulation, which tracks and manages Freedom of Information exemptions applied to records to limit access. FoI enquiries are tracked by the FoI tracking system.

Access to TNA's systems is restricted and controlled. The systems are physically secured to government standards for managing records at the levels of security classification TNA holds. This includes electronic access passes on the doors, fire control, intrusion detection systems and closed circuit cameras to limit physical access, and systems and networks accredited to government information security standards for the nature of the records held.

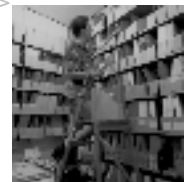
In terms of Establish Standards and Policies, TNA complies with government standards and legislation for records and information management. The establishment of particular standards and policies for digital preservation is largely the responsibility of the Digital Preservation department and the Digital Records Board, but must also include the Departmental Security Committee, the Records Management Department, the Advisory Service, the Seamless Flow programme, and others. TNA pays close attention to developing research and international standards and participates in the setting of some international standards.

### **Administration: Conclusions**

In practice, the disparate functions of the administration, as identified in the OAIS reference model, are usually spread across departments and units and the responsibility for them is shared amongst employees. The identification of all the listed OAIS administration functions at TNA and the UKDA proved to be a laborious task but the general conclusion is that all these duties are an organic part of every archive's way of functioning and will be present although only some are clearly visible in the day-to-day work of the archive.

The OAIS standard places a lot of emphasis on the administration, monitoring and management of decision processes. This reflects the situation in a larger institution with several departments and a large staff number. In a not insignificant number of European data archives, digital archiving is either a small element of a much wider concern or is a very small-scale operation in which decision-making and implementation are undertaken by the same individual. As examples of extreme instances of this, one Social Science Data Archive has just one employee who has responsibility for the entire system and all the associated processes, including the administration functions; another employs only two people, both of whom work part-time with one taking managerial and administrative responsibility and the other the full range of day to day responsibilities. The OAIS reference model, it seems, does not offer helpful guidance for such small-scale but real-life situations.

Similarly, the reference model places emphasis on the Activate Requests and Customer Service functions which are valuable in terms of maintaining records and monitoring use and services offered. However, it also seems to assume that such functions will be at least partially automated which, of itself requires financial investment to set up. For organisations of the size and stature of UKDA and TNA, these requirements have become functions of the working environment and are necessary to maintain effective management of extensive collections and numbers of users but such systems, either formal or informal, are much less of a priority for poorly resourced, small organisations with small collections, low numbers of users and limited technical infrastructures on which to build.

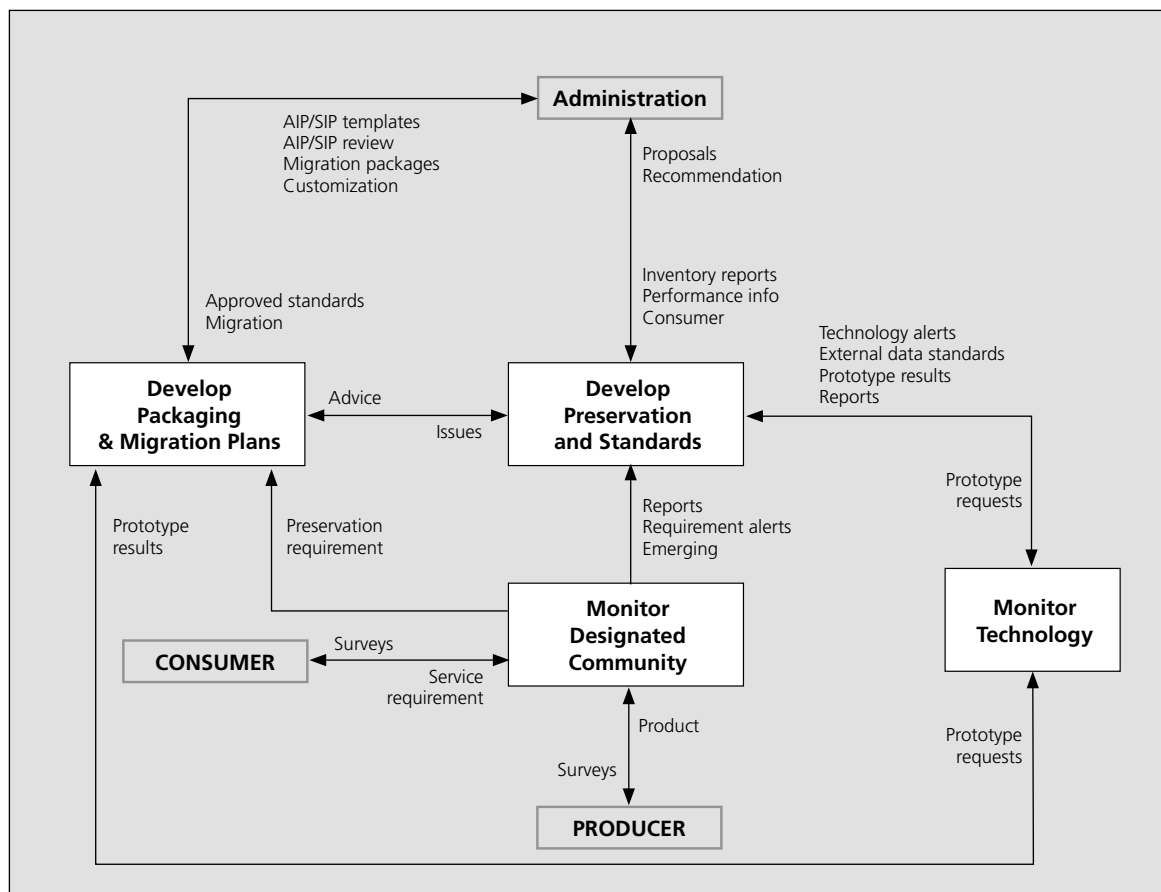


Administration is a vital activity that every archive must perform. While there is nothing much to disagree with in the OAIS model, every institution will have its own set of policies, procedures, requirements and responsibilities, which will inform the administration activities. OAIS, in this regard particularly, is only providing a very general functional subset of digital archive administration activities, where in other cases its functional model is more clearly complete. The biggest explicit divergence from the model is still the OAIS's assumption that DIPs are generated on demand from the AIP, but there are many other activities which simply do not appear. Both TNA and the UKDA have many administration functions surrounding DIP storage, DIP migration planning, information security management, access control, freedom of information tracking, resource discovery, catalogue enhancements, change management, accreditation procedures, etc., and the boundaries between these and the OAIS functional model are not always cleanly delineated. This makes it difficult to provide a mapping which is actually representative of the activities of each archive. The OAIS administration model in particular should be taken as necessary for digital archive administration, but in no way should it be regarded as sufficient.

## 5.5 PRESERVATION PLANNING

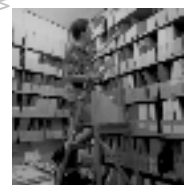
The fifth functional component of an OAIS archive is Preservation Planning. This service is responsible for developing the OAIS's preservation strategy, as well as recommending appropriate revisions to this strategy in response to evolving conditions in the OAIS environment. The Preservation Planning service monitors the external environment for changes that could impact the OAIS's ability to preserve and maintain access to the information in its custody, such as innovations in storage and access technologies, or shifts in the scope of requirements or expectations of the Designated Community. Preservation Planning develops recommendations for updating the OAIS's policies and procedures to accommodate these changes. The Preservation Planning function represents the OAIS's safeguard against a constantly evolving user and technology environment. It detects changes impacting the OAIS's ability to meet its responsibilities, designs strategies for addressing these changes, and assists in the implementation of these strategies within the archival system. Although the Preservation Planning entity operates under the Administration entity, it exercises some administration-like controls over preservation decision-making rather than general management functions.

The functions of this entity are presented in the following diagram.



**Figure 8. The OAIS Preservation Planning function.**

- the Monitor Designated Community function interacts with archive Consumers and Producers to track changes in their service requirements and available product technologies. It provides reports, requirements alerts and emerging standards to the Develop Preservation Strategies and Standards function. It sends preservation requirements to Develop Packaging Designs;
- the Monitor Technology function is responsible for tracking emerging digital technologies, information standards and computing platforms to identify technologies which could cause obsolescence in the archive's computing environment and prevent access to some of the archives current holdings. This function sends reports, external data standards, prototype results and technology alerts to Develop Preservation Strategies and Standards. It also sends prototype results to Develop Package Designs and Migration Plans;
- the Develop Preservation Strategies and Standards function is responsible for developing and recommending strategies and standards to enable the archive to better anticipate future changes in the Designated Community service requirements or technology trends that would require migration of some current archive holdings or new submissions. This function receives reports from the Monitor Designated Communities and Monitor Technology functions, and it receives performance information, inventory reports and summarized consumer comments from Administration. This function sends recommendations on system evolution to Administration. This function also receives external data standards from Monitor Technology and produces profiles of those standards that are sent to Administration as proposals on their potential usage. This function also receives issues from Develop Packaging Designs and Migration Plans in the case of unanticipated submission requirements, and responds with advice to handle the new requirements;



- the *Develop Packaging Designs and Migration Plans* function develops new IP designs and detailed migration plans and prototypes, to implement Administration policies and directives. This activity also provides advice on the application of these IP designs and Migration plans to specific archive holdings and submissions. This function receives archive approved standards and migration goals from Administration. It applies these standards to preservation requirements and provides AIP and SIP template designs to Administration. This function also provides customization advice and AIP/SIP review to Administration on the application of those designs. If this function encounters submissions that are not covered by existing standards and procedures, it can send issues to *Develop Preservation Strategies and Standards* and receive advice, including new standards, to assist in meeting the new submission requirements.

## Preservation planning: UKDA

The UKDA Outreach and Training and Data Support Services teams are responsible for monitoring the user community's needs and requirements. This is achieved by a number of means and can vary between services. Methods used include workshops, user surveys, opportunities to make requests and suggestions via web pages, user representation on advisory committees and monitoring of requests to the help desks. The relevant management or service team considers the results of these activities and, if resulting decisions are expected to impact across the entire organisation they will be taken to meetings of SMT and DAG for information or further discussion. The responsiveness of the UKDA to such information is evidenced by the development of the automated download system, the online registration system and the introduction of STATA files as DIPs.

The preservation strategy and policy are the responsibility of the Systems and Preservation Section which has produced the UKDA Preservation Policy document. The preservation policy of the UKDA is monitored and reviewed in the light of changing technologies on a biannual basis to ensure timely updates. The Head of Systems and Preservation group or the Director of the UKDA initiates the review process. Implementation of the preservation policy is monitored and there are regular planned audits to assess how the policy is executed, including an annual benchmarking of the UKDA preservation agenda.

The UKDA is committed to reviewing and remaining up-to-date with technological advances. It endeavours to be capable of adapting or initiating new preservation standards and procedures that are suitable for its collections. The UKDA supports and participates actively, where appropriate, in research, development and implementation of new practices for the preservation of digital resources.

The Systems and Preservation group of the UKDA is responsible for monitoring developments and advances in the digital preservation area through a technology watch scheme. Based on the technology watch, the Preservation section makes recommendations for migration plans that are then approved by the archive administration and management teams. Depending on the system involved, technology watch is performed by all members of the Preservation section. Any information gained, which is relevant to either current or new hardware or software, is drawn up into a report. Although the technology watch is an ongoing process, the reports are drawn up bi-annually and discussed in detail.

The AIPs and the DIPs are designed on two fundamental principles, respectively. For AIPs, the principle is the need to preserve the integrity, the use and the longevity of the object; for DIPs, it is to provide the user with an object that is understandable and useable.

These fundamentals have not altered over time but pragmatism and the responsiveness of the UKDA to changes in the technical environment and changes in its user base have resulted in modifications to the precise content of AIPs and DIPs. A typical example of such a change to both DIP and AIP is the shift from the preservation and supply of electronic data files and accompanying, explanatory paper documents, to the preservation and supply of data files and electronic copies of accompanying explanatory documents. In this respect, one might argue that



the actual nature of a DIP or an AIP has changed over time and may do again if, as is likely, life-cycle standards are developed which allow the holding of data and metadata in a single, software and system independent file.

The decision-making structures of the UKDA have similarly changed over time, in response to the size and remit of the organisation. Nevertheless, fundamental decisions about changes to both AIPs and DIPs have been taken within a defined process as required. Changes to AIPs and to DIPs may, then, result from technological change or from user requirements. Consequently, requests for changes to AIPs or DIPs can come from a number of sources or sections within the organisation. A strength of the UKDA is that its decision-making structure is sufficiently flexible to manage requests for change from any direction or section so that if a request for change to the content of DIPs emanated from the Data Services section, it would be discussed at DAG where the implications on other areas of work could be assessed. Similarly, if a request for a change to AIPs came from a service advisory board, it would be discussed by the SMT and presented to DAG for discussion of its implications. If an occasion arose where a consensual decision could not be made, the Director would decide based on advice received from the different levels of management. Once a decision for change is made, it would then be presented to DAG, after which it would become part of the relevant standard or procedure and implemented as part of the daily work of the relevant section or sections.

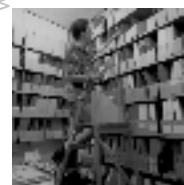
### **Preservation planning: TNA**

At the National Archives the user access departments feed back user and technology requirements to the Technology Watch function in the Digital Preservation department. Data producers interact with client managers from the Records Management department, who optionally involve technical staff in the Digital Preservation department as required. TNA's consumers interact with the Online Content and Presentation department and the Reader Services department. The Digital Preservation department gets information directly from client managers, as specified in the various TNA policies. Reader services provide useful information on record usage and any usability problems.

The Technology Watch function is carried out at the Digital Preservation department. Information from the departmental activities listed above, along with technical information generated from the Digital Archive is cross-checked with the file format and technical registry. PRONOM was developed to provide this function, initially as an internal resource for TNA staff, and now on the Internet.

PRONOM 2 was released in December 2002 and provided support for the development of multilingual versions of the system, through the replacement of field tags. Since that time, considerable effort has been devoted to the development of PRONOM's content. Digital Preservation staff have undertaken intensive research and liaison with major software developers in order to create an initial core data set of software product information. This work is ongoing, although TNA would also encourage software developers and others to be proactive in providing information.

The web enabling of PRONOM (PRONOM 3) marked the next stage in the evolution of this system. However, in many ways this represents the starting point for the development of PRONOM as a major online resource for the international digital preservation community. TNA has detailed plans for major enhancements over the next two to three years, including the development of a number of specific tools to support digital preservation activities.



Work is currently in progress on the next stage of development (PRONOM 4), which will deliver a number of significant enhancements. These include a major revision of the underlying database structure, to allow the recording of detailed technical information about file formats, and the development of a stand alone tool to perform automatic file format identification (DROID), using signature information stored in the database.

Future work on PRONOM is being taken forward as part of TNA's Seamless Flow programme. The PRONOM system will form the basis for developing the technology watch capability, providing a technical registry service to support a variety of preservation activities, including the technical characterisation of electronic records upon ingest to the digital archive (DROID), preservation risk assessment and planning, and automated format migration.

A programme of trials and pilots actively investigates new digital preservation and presentation technologies and frameworks to make best use of the capabilities of each under a single management system.

Results of the technology watch are reported via departmental team meetings, and presentations to the Digital Records Board which includes representation from across TNA. The Digital Records Board also provides a forum for the organisation to review and monitor proposals for activities such as preservation strategy updates, migration paths, and other activities relating to digital records.

It is important to note that DIP production and migration planning is currently a more significant activity in terms of technology watch than preservation migration. To date, most records accessioned are still readable using current technology, but the demands of presentation over the web are driving the need to migrate all these formats into web-deliverable information.

### **Preservation planning: Conclusions**

The planning of preservation activities in the UKDA and TNA has not been formalised to the full extent that the OAIS model would see it happening. Units responsible for preservation are not duplicating the work of user services units of the archive and hence receive information about changes in the OAIS external environment via these channels. Responsibility for developing new standards and policies for digital preservation is firmly rested on the preservation departments but their recommendations are, as a rule, approved by a corporate level decision-making body.

From the viewpoint of the UKDA this area of the OAIS standard can sometimes be overly bureaucratic and over-concerned with processes. Realistically organisations like UKDA have to be more pragmatic in their approach to decision making: decisions are often made out of necessity and are reactive rather than proactive and some decisions happen because of a need, rather than being planned. The OAIS reference model only provides a formalised view of the functions of digital archiving; it does not prescribe implementation strategy or management style. Nevertheless, a real archival organisation never operates quite as "cleanly" as the OAIS model envisages — the planning of preservation methods certainly happens at the UKDA but it is difficult to map the workflows quite as distinctly as the OAIS function suggests.

TNA is very active in setting preservation standards (e.g., PDF/A, eGMS, ERMS), building technologies to support migration planning and monitoring technology (PRONOM), and file format identification (DROID). TNA is also very active in monitoring its designated community (as is UKDA). This does in fact map well to the OAIS reference model functions.



The practical reality is that preservation planning is in heavy development and is probably not as mature as other aspects of digital archiving. So far, TNA's need to migrate has primarily been driven by the need to produce DIPs. The absence of any generally good technical information repositories or migration planning systems has forced TNA to develop its own and plan accordingly. This makes it sometimes difficult to map the activities to the OAIS reference model because those functions are not directly fulfilled in such a formalised way. The absence of tools, techniques and technologies to manage this, forces all archives to make pragmatic plans based on what is required, available and possible. This is one of the key drivers for TNA to invest in web-enabling PRONOM, as there is a real need for reusable tools to support general migration planning in digital archives around the world.

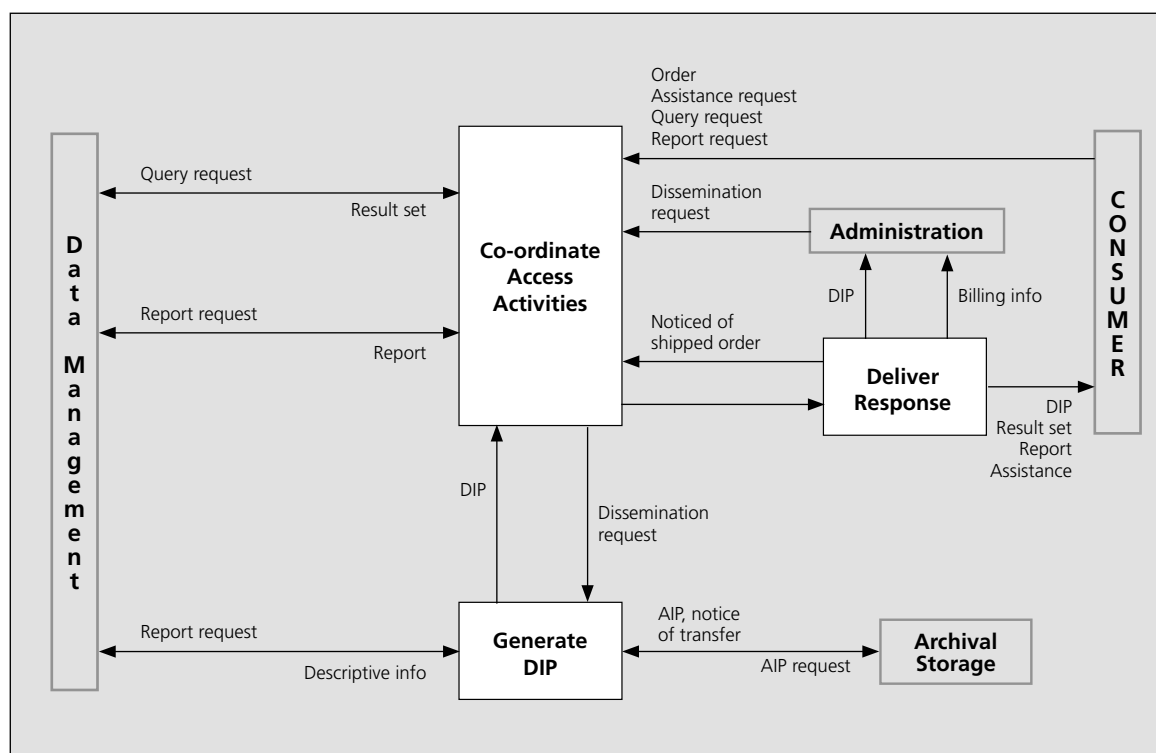
## 5.6 ACCESS

Access is the sixth and final functional component of an OAIS archive. The Access function represents the OAIS's interface with its Consumers: it is the primary mechanism by which the OAIS archive meets its responsibility to make its archived information available to the user community. It manages the processes and services by which Consumers – and especially the Designated Community – locate, request, and receive delivery of items residing in the OAIS's archival store. Access functions include communicating with Consumers to receive requests, co-ordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering the responses to Consumers. Access is also responsible for implementing any security or access control mechanisms associated with the archived content.

The Access entity in an OAIS has no particular dissemination or publication scheme in mind when it makes a digital object available, though particular publication destinations (such as web sites or advertising brochure) are amongst the many possible uses for disseminated objects.<sup>51</sup>

The functions of this entity are presented in figure 9.

51 William Lazorchak, "The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation" (2004), p. 47



**Figure 9. The OAIS Access functions**

- the Co-ordinate Access Activities function provides a single user interface to the information holdings of the archive. This interface will normally be via computer network link to an on-line service, but might also be implemented in the form of a walk-in facility, printed catalog ordering service, or fax-back type service. Three categories of Consumer requests are distinguished:
  - query requests, which are executed in Data Management and return immediate result sets for presentation to the user;
  - report requests, which may require a number of queries and produce formatted reports for delivery to the Consumer; and
  - orders, which may access either or both Data Management and Archival Storage to prepare a formal Dissemination Information Package (DIP) for on- or off-line delivery.

Other special request types are allowed, but are not detailed. This function will determine if resources are available to perform a request, assure that the user is authorized to access and receive the requested items, and notify the Consumer that a request has been accepted or rejected (possibly with an estimate of request cost and an option to cancel the request). It will then transfer the request to Data Management or to the Generate DIP function for execution. This function also provides assistance to OAIS Consumers including providing status of orders and other Consumer support activities in response to an assistance request.

- the Generate DIP function accepts a dissemination request, retrieves the AIP from Archival Storage, and moves a copy of the data to a staging area for further processing. This function also transmits a report request to Data Management to obtain Descriptive Information needed for the DIP. This function then places the completed DIP response in the staging area and notifies the Co-ordinate Access Activities function that the DIP is ready for delivery;



- *the Deliver Response function handles both on-line and off-line deliveries of responses (DIPs, result sets, reports and assistance) to Consumers. For on-line delivery, it accepts a response from Co-ordinate Access Activities and prepares it for on-line distribution in real time via communication links. It identifies the intended recipient, determines the transmission procedure requested, places the response in the staging area to be transmitted, and supports the on-line transmission of the response. For off-line delivery it retrieves the response from the Co-ordinate Access Activities function, prepares packing lists and other shipping records, and then ships the response. When the response has been shipped, a notice of shipped order is returned to the Co-ordinate Access Activities function and billing information is submitted to Administration.*

### Access: UKDA

The UKDA distributes and provides access to its data collections via:

- HTTP download;
- online access;
- guest FTP;
- CD-R and DVD-R;
- other media by special request (e.g., DAT, Exabyte, Zip disc).

The UKDA's HTTP-based download service provides a quick and reliable means of gaining access to the most heavily used collections held at the archive. The archive also provides online access to data that have been enhanced and published in the Nesstar system. Nesstar provides the capability for data discovery, browsing, subsetting, visualisation and downloading via the Internet. The system is based on the DDI metadata standard.

Users of the archive must be registered to order data but can browse the catalogue without being registered. All users can:

- view various levels of the metadata (bibliographic record, abstract, documentation, user guide, variable list, data dictionary);
- download or browse the documentation;
- access frequency counts for selected data (via Explore Online/Nesstar links).

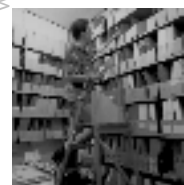
Registered users can:

- browse, analyse or download the data for a growing number of the UKDA's most popular data series;
- order the data and documentation in a variety of software formats and media.

For further information on access to UKDA data collections see the *Using data, Ordering/Downloading data* and Nesstar documents on the UKDA web site.<sup>52</sup>

The management of the UKDA's catalogue is shared between the Data Processing team which has day-to-day responsibility for content and the Information Development team which is responsible for the structure and for quality control. Catalogue records are created during ingest processing, when the main metadata record is generated. Further metadata are generated during processing to allow users to search through datasets and variable specific information within the catalogue.

<sup>52</sup> <http://www.data-archive.ac.uk/search/searchStart.asp>



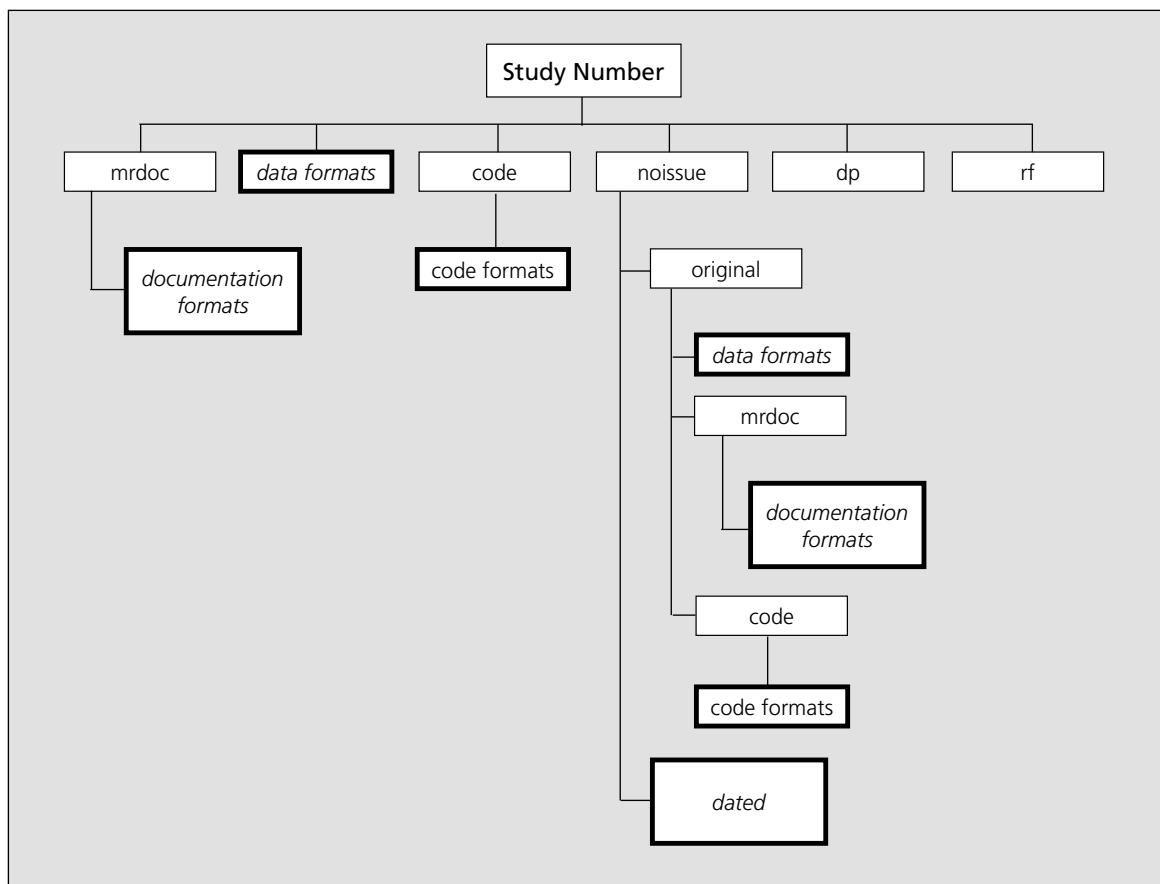
For studies processed in SPSS format (the vast majority) processing procedures consist of generating and archiving a UKDA data dictionary, a rich text format document created by UKDA software, giving more detail than the data dictionary generated by SPSS. A file logging any unavoidable inconsistencies between the data in SPSS and STATA formats is also generated and archived for all applicable studies. This is supplied to anyone ordering the data in STATA format, who can then locate the lost information in the data dictionary. For the small number of studies processed in other formats, equivalent information is recorded (e.g., MS Access data documenter output). Data are always preserved in a preservation format in addition to the proprietary formats used for dissemination. Preservation formats consist of tagged or delimited text of a given character set (ASCII or UNICODE) or eXtensible Markup Language (XML). Where necessary, data definition statements (in SPSS, STATA, SAS, SQL or Visual Basic command languages as appropriate) are also preserved, to preserve the full information of the dataset (variable formats, variable labels, code labels, missing value definitions, etc.).

Documentation required for resource discovery and resource use is collected, created and held in the UKDA as metadata attached to each archived study (data collection). The study read file, the user guide, Depositor Licence and the study description or catalogue entry are the main components of such metadata. The study description is based on the DDI metadata standard for data documentation. The study description is also mapped to the Dublin Core metadata standard, and the UKDA's online catalogue is Z39.50 compliant and compatible with the Open Archives Initiative (OAI) protocol for metadata harvesting. The catalogue is produced and maintained employing XML based on the DDI Data Type Definition (DTD). Resource discovery in the catalogue is further enhanced by the use of the Humanities And Social Science Electronic Thesaurus (HASSET) that was created and is maintained by the UKDA. The Information Development team interprets national and international standards for local implementation in its resource discovery metadata.

The workflow for generating DIPs in the UKDA does not follow the pattern suggested by the OAIS reference model. The DIPs are created at the ingest processing phase, together with AIPs. The DIPs and AIPs of each corresponding SIP are kept in different parts of the preservation system and when a user request is received, the Access function merely retrieves the DIP from the storage system and provides it to the user. Figure 10 below illustrates the preservation file structure in the UKDA.



**Figure 10. The file structure of the UKDA preservation system**



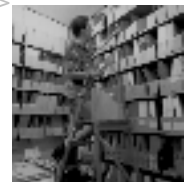
*Files on this side of the chart above  
can go to users*

*Files on this side of the chart above  
are only for internal archival purposes*

**Key to boxes in tree diagram above:**

actual  
directory  
names

names(s) is  
software and/or  
format specific



**Key to directory types shown left:**

| STUDY DIRECTORY                        | PURPOSE                                                                                                                                     |
|----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| [study]                                | The top level (i.e. root) directory.                                                                                                        |
| [study]/'data format'                  | Directories/ parent directories for all machine readable data, both quantitative and qualitative, that are to be issued to users.           |
| [study]/mrdoc                          | Parent directory for all machine readable documentation that is to be issued to users.                                                      |
| [study]/code                           | Parent directory for all machine readable code (command files) that is to be issued to users (code to run on quantitative data files only). |
| [study]/noissue                        | Parent directory for all data and doc not to be disseminated.                                                                               |
| study]/noissue/dated                   | Parent directory for files updated or altered by a new edition.                                                                             |
| study]/noissue/original                | Parent directory for original files – i.e. files sent by the depositor and not altered in any way other than unzipping/decompressing.       |
| [study]/noissue/original/mrdoc         | Parent directory for original doc files (i.e. all machine readable doc files sent by the depositor).                                        |
| [study]/noissue/original/'data format' | Directories/ parent directories for all original machine readable data files sent by the depositor.                                         |
| [study]/noissue/original/code          | Directories/ parent directories for all original machine readable code files sent by the depositor (code to run on quantitative data).      |
| [study]/dp                             | Parent directory for internal use only files created by the Archive during processing.                                                      |
| [study]/rf                             | Scanned page images of material from the study's folder (licence forms, deposit forms, correspondence and miscellany).                      |



In providing access to its collections, the UKDA is regulated by the deposit agreements that establish use conditions for every data collection and the associated access agreements that its users have to accept. Users must be registered to order data from the archive but can browse the catalogue without registering.

The UKDA utilises the Athens Access Management system for authentication and authorisation. This is a national, centralised service which provides, in effect, a gateway connecting service providers and users.

The UKDA owns Athens resources which may be made available, by the Archive's permission, to any organisations that supply Athens usernames to their communities. The UKDA also acts as a username supplier for those individuals whose organisations do not have this facility. Under an agreement signed with Athens, the UKDA may issue Athens usernames to up to 15 people from any one organisation at any one time. In this way, all the UKDA's users may acquire a username and log in to the resources.

All authentication takes place centrally via Athens. Fundamental authorisation also occurs at Athens, granting the user permission to access the resource or not. In the case of the main ESDS resource, all users may enter; however, for ESDS International only those within UK HE/FE are permitted. Further access control then takes place at the UKDA for both of these resources, once the user has authenticated, as some datasets within the collections require further terms and conditions to be accepted before they may be viewed.

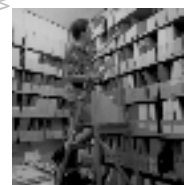
### **Access: TNA**

The users of the National Archives digital collections can gain access to the Digital Archive records (using a web browser) via reading rooms or from the Internet using the EROL system. The Digital Preservation Department produces the DIPs for each AIP that are loaded into the presentation system, and users interact with the Digital Archive Presentation system directly in the reading rooms. If there are special viewing requirements, an email form is provided on the presentation system that will be picked up by staff in the Digital Preservation department and special arrangements would be made as necessary. TNA Web Archive allows material to be requested via the web. Aside from these methods, material can be provided on request if users fill out a submission form or they simply ask the archives. User access requests are generally handled by TNA's online catalogue PROCAT, the Digital Archive Presentation system and the Web Archive online server.

The Catalogue unit manages the PROCAT system. Any additional technical and archival metadata which are not released through the catalogue and are particular to the Web Archive or the Digital Archive, are first reviewed by the Digital Preservation department. The department decides what additional metadata it wishes to release in order to support the presumption of authenticity of the record for users.

In TNA, DIPs are pre-generated by the Digital Preservation department at the ingest phase and are held separately from archival storage, in the presentation systems. The choice of file formats for DIPs is managed through the Technology Watch function, using migration plans and information from the online access part of the organisation. New DIPs are created as presentation technology evolves, or potentially, better migration tools become available which would substantially improve the quality of the material delivered.

In special cases, a DIP may be created by special request. Generally, no request is made for an AIP, but the only exception to this rule is a request made for closed information (Freedom of Information) material which is technically impossible to deliver over the Internet or via reading rooms (e.g., file is too large, not technically deliverable on that medium, etc.). These requests are made by email using a web form and are considered on a case by case basis, as they may involve work by archive staff which needs to be costed.



Access to public records is governed by legislation and users do not need to sign agreements or receive licences. Users can receive records from TNA via the Internet, reading rooms and via any other agreed medium or format (subject to charging as laid out in the relevant Statutory Instruments). Users can gain online access to both digital collections and their metadata but the user would have to be authenticated for accessing the collections.

## Access: Conclusions

In providing access to their collections, UKDA and TNA diverge perhaps most from the OAIS reference model. The long-standing tradition in these institutions is that all SIPs are processed at ingest and both an AIP and DIP version are created from it during this processing. DIPs are then stored and retrieved from storage once a user request is received, or placed directly on the online access system where users can download and process DIPs online. The OAIS model concept of creating DIPs from AIPs “on the fly” in response to a user query represents an exception, not a rule in the tested archives. This, however, is seen as an advantage, rather than weakness by these institutions, since creating the DIP at the time of ingest means that there is access to a richer information source, the depositor, for the staff carrying out the process. It is almost certain that no archive will be in a position to collect one hundred per cent of the information relating to a dataset and if anomalies are discovered, the depositor of the SIP is the most likely aid to resolution. If the DIPs are created years or even decades after deposit, the original depositors may no longer be able to answer detailed queries and, in the case of research data, the research teams are likely to have dispersed and thus the key source of information is lost to the archive.

Both institutions have a long tradition of participating in development of metadata standards and using international standards for describing their collections. Changes in the technology used for creating the digital resources are also likely to require the development of standards for description. Both archives are aiming to create DIPs with dissemination-specific metadata attached, making it easier for users to understand and make use of the digital resources they have requested.

## 5.7 GENERAL FUNCTIONALITIES

In addition to the entities described above, the OAIS reference model assumes various Common Services to be available. These services are considered to constitute another functional entity in the model.

Both institutions participating in the compliance testing maintain their own operating system and network services that comply with the requirements of the OAIS standard. Operating system services provide the core services needed to operate and administer the application platform, and provide an interface between application software and the platform. Network services provide the capabilities and mechanisms to support distributed applications requiring data access and applications interoperability in heterogeneous, networked environments.

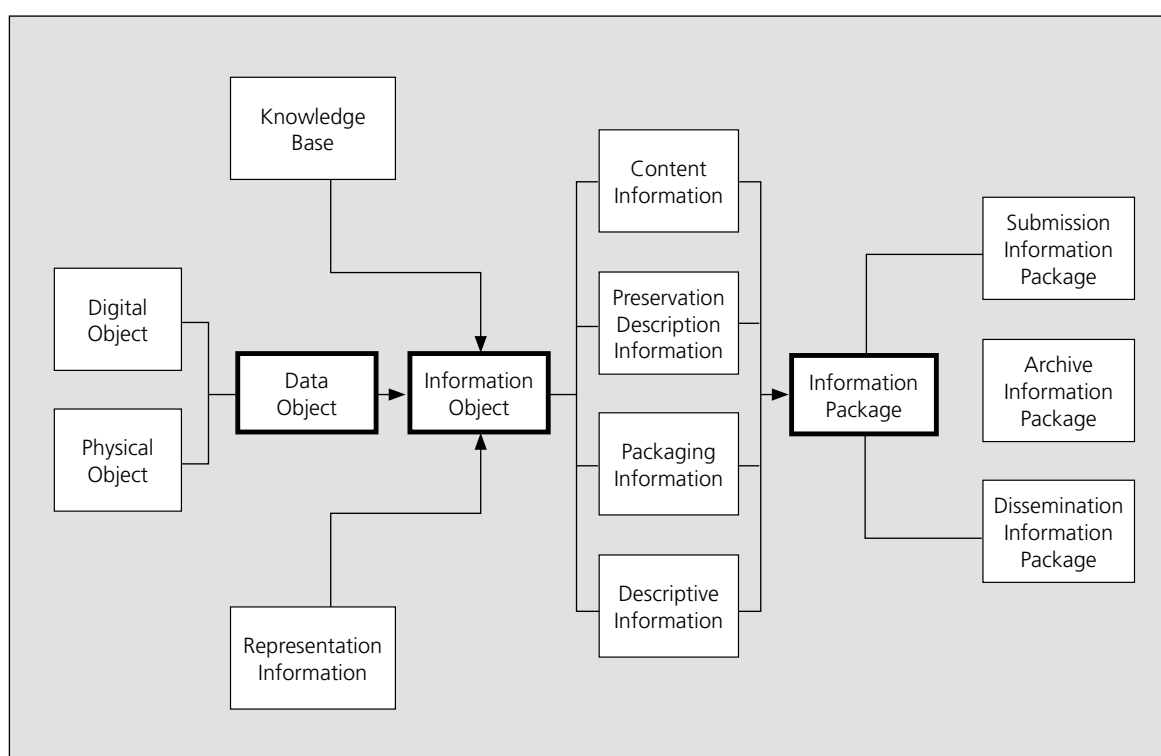
Security services capabilities and mechanisms protect sensitive information and treatments in the information system. UKDA are responsible for defining security on their systems whilst the University of Essex Computing Services act as the authentication to the set security. TNA conform to government information security standards for the level of material that they preserve and ensure that only publicly accessible information is placed in public-facing systems. Closed material is governed by legislation, policy and internal logical and physical access controls, and is managed in air-gapped systems. TNA does not require user authorisation to access public data.

The UKDA manages user authorisation via the Athens system which was described earlier. As a place of deposit for TNA, the UKDA also has to meet TNA requirements for its machine and storage rooms and for physical access to its buildings. These have all been described above in section 6 and fall under BS 5454.

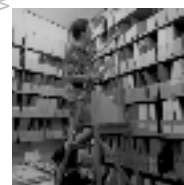


## 6 OAIS INFORMATION MODEL

The OAIS information model defines the broad types of information that would be required in order to preserve and access an information object stored in a repository. Information itself is defined as “any type of knowledge that can be exchanged, that is independent of the forms used to represent it, and this information is always expressed by some type of data”. The OAIS information model makes a basic assumption that all information objects are composed of a data object and the representation information that would permit the full interpretation of the data object into meaningful information. In order to successfully preserve such a generic information object, it is critical for an OAIS archive to clearly identify and understand the data object and its associated representation information, which together form what is called an ‘information package’.



**Figure 11. The OAIS information model.** (from Lavoie. 2000. p.27)



There are three types of OAIS Information package:

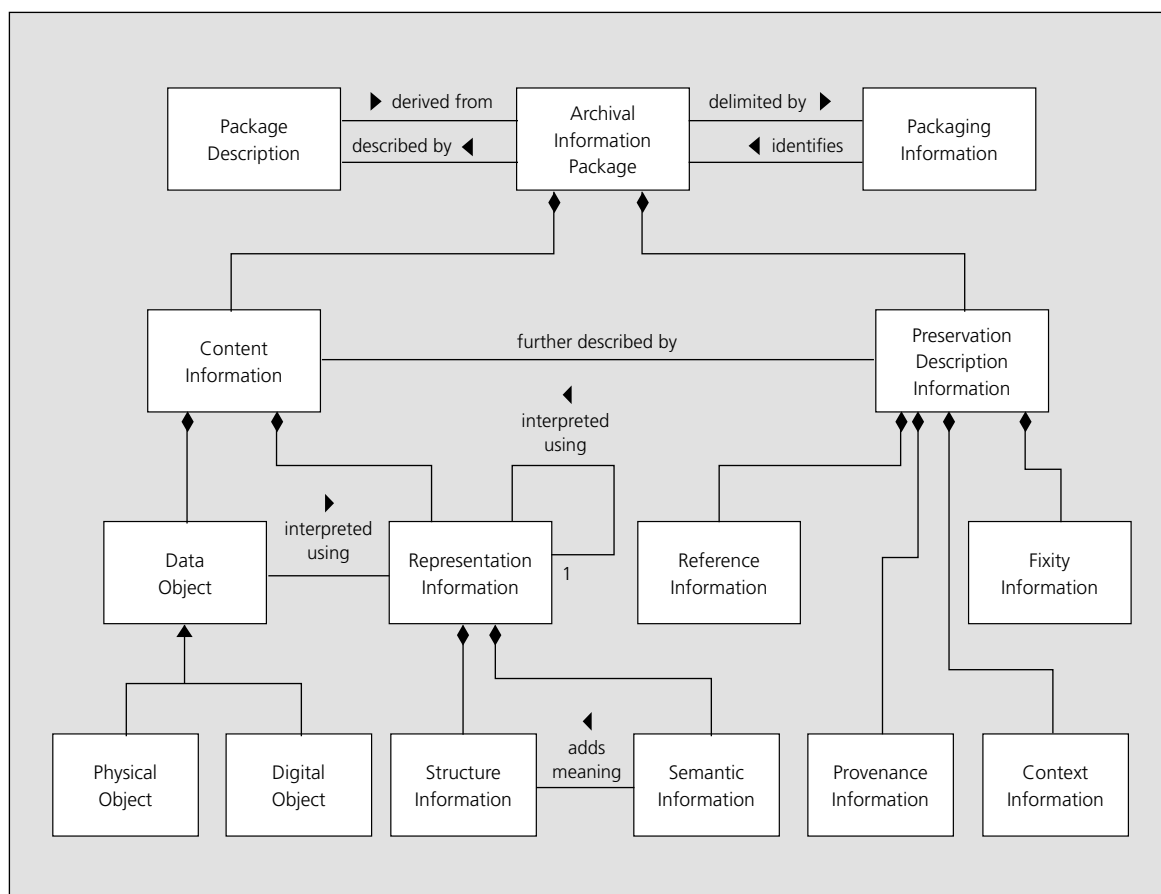
- Submission Information Package (SIP) – delivered to the Archive by the Producer;
- Archive Information Package (AIP) – generated from SIPs and stored by the Archive;
- Dissemination Information Package (DIP) – is transferred from the archive in response to a request by a Consumer.

The notion of a generic information package that must be managed through time is fundamental to the OAIS model. The information packages are the means of interaction for actors within the OAIS environment and define many of the specifics of the interfaces between the functions of an OAIS.

The OAIS reference model does not include a metadata standard with specific metadata elements. Instead, it defines four categories or blocks of metadata that are required for the OAIS to be able to successfully preserve and manage its collections: content information; preservation description information; packaging information; and descriptive information. These are defined below.

- Content Information (cf. ch. 4.2.1.4.1 in OAIS reference model).  
The content information is the original target of preservation. It consists of the content data object and its associated representation information needed to make the data object understandable to the designated user community.
- Preservation Description Information (cf. ch. 4.2.1.4.2 in OAIS reference model).  
The preservation description information (PDI) contains information that is necessary to adequately understand the content information over an indefinite period of time. It would include provenance information, unique identifiers for the content information and information validating the authenticity of the content information. The PDI is typically divided into four sections:
  - provenance information;
  - context information;
  - reference information;
  - fixity information.
- Packaging Information (cf. ch. 4.2.1.4.3 in OAIS reference model).  
The information that binds the components of the information package into an identifiable entity on specific media.
- Descriptive Information (cf. ch. 4.2.1.4.4 in OAIS reference model).  
Information that helps users of the archive to locate and access information of potential interest.

Each of the three information package types (SIP, AIP, DIP) has varying levels of information in the four information objects (e.g., it may require several SIPs to provide a complete set of Content Information and associated Preservation Description Information).



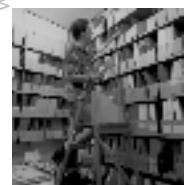
**Figure 12. An example of AIP components.** (from: CCSDS 650.0-B-1, 2002, p. 4-37)

Perhaps the area where library and archives communities have made the most fruitful contribution to the development of the OAIS reference model is in the various types of metadata that must be included with a digital object when it is archived in an OAIS. As this project reached completion, the international working group, led by the OCLC and RLG,<sup>54</sup> released a guide to core metadata for supporting the long-term preservation of digital materials as a report from the PREMIS working group. This timing meant that any attempt to map the preservation metadata used at the UKDA and TNA to the OCLC/RLG working group recommended preservation metadata set would have been premature. Instead, a broader-level mapping of the OAIS information packages and metadata categories to the digital preservation metadata and preserved objects was undertaken.

### Information model: UKDA

A typical SIP submitted to the UKDA is either quantitative data, e.g. SPSS files, or interview transcripts, e.g. MS Word files. The SIP is accompanied by a data submission form, data collection form and licence agreement. The accompanying documentation describes the data being submitted, such as its format, provenance, sources and file size. Larger surveys tend to have greater re-use value than smaller ones and therefore studies with large file sizes tend to make up the larger proportion of the UKDA data collections.

<sup>54</sup> RLG/OCLC, 2002



In January 2005, the Freedom of Information Act (Fol) was introduced and the UKDA was established as an agent of TNA. This means that since January 2005, staff at the UKDA are responsible, for Fol requests which relate to studies held on behalf of TNA. As a consequence, it has become a legal requirement for a TNA tagged SIP to be held in a manner which enables information held to be provided readily as understandable information. This is not, however, the case for SIPs deposited under the ESRC or AHDS History contracts, for which requests under the Fol would be referred back to the depositor.

The transformation of a SIP into an AIP involves the creation of a catalogue record for the submission and the addition of new metadata to the record. If necessary, the file format of the SIP is changed when creating the AIP, while the original SIP format is retained.

Information from the data submission form, which the depositor initially submits for evaluation is used to create a record in the MIRAGE acquisitions database. If data are accepted by the ARC then the information in MIRAGE is fed into the CALM processing database to initiate the creation of the catalogue record. Along with the data and licence agreement, the depositor also completes and returns a data and documentation transfer (data collection) form which contains further details about the data and is also used to create part of the catalogue record. This information includes details about any checks that have been made to the data, how the data were collected and sampling details.

The additional metadata are incorporated into Read and Note files which are created in the CALM database as the dataset is being processed. The Read file is used to record any problems encountered with the data, new edition information and any legal conditions of use. The Note file details any file format changes, as well as the validation checks performed and reports on any data or documentation problems.<sup>55</sup>

The DIP is usually created from the preservation copy (AIP). Different format DIPs can be created at the request of users. Documentation, such as user guides and Read files, are also supplied to the user and can be downloaded from the catalogue record located on the UKDA web site. The Read File may contain notes regarding accompanying documentation as well as any changes or information that is necessary to fully understand the data, such as details about weighting factors or their application. Standard study related information is also made available for download and includes acknowledgements, disclaimers and citations. A user guide or data dictionary will also be included for download.

The UKDA is using a number of metadata standards for managing its collections (see Table 1). The main metadata standard used at the UKDA is the Data Documentation Initiative (DDI).

The UKDA's interest in standards for data exchange date back to its origin in the late 1960's. If archives were to realise their philosophy of international data exchange for research purposes, they needed a standard to enable them to read data received from partner archives. Until the 1990's archives used the OSIRIS codebook, developed at Michigan for the Inter-university Consortium for Political and Social Research, as their standard.

OSIRIS gradually became outdated with the rapid changes to the technical environment in which archives functioned and the DDI, to which OSIRIS files can be converted, has emerged as the new standard.

The DDI is an XML-based international standard for the content, presentation, transport and preservation of documentation for datasets in the social and behavioural sciences. The Nesstar software was developed by a consortium of European archives and is used as a dissemination method by the UKDA which has also developed in-house tools to facilitate the creation of the DDI compliant files from commercially available statistical software packages.

<sup>55</sup> See Appendix 4 for extracts from typical Read and Note files



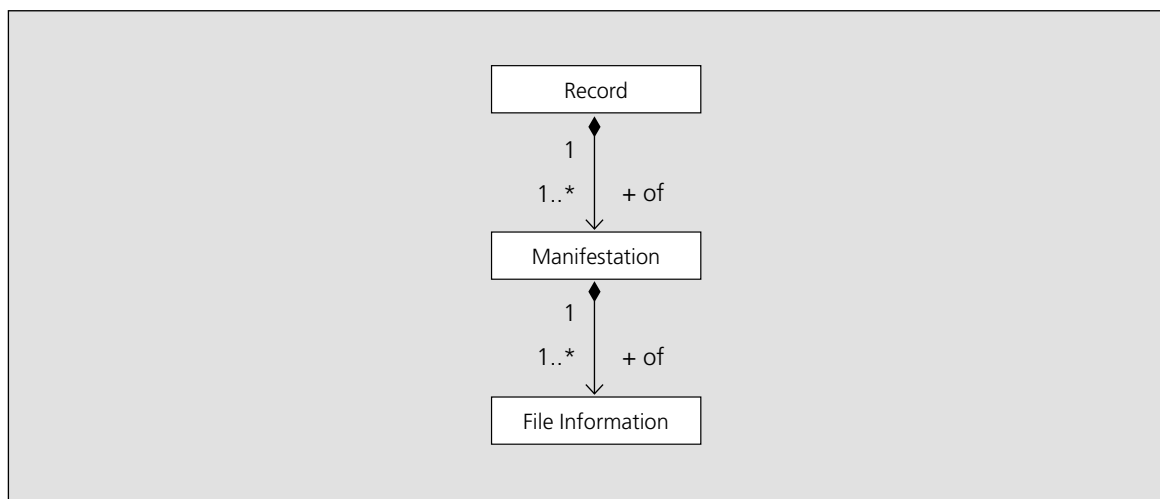
The DDI is now managed by an alliance of approximately 25 member institutions.

| STANDARD                             | STATUS            |
|--------------------------------------|-------------------|
| DDI                                  | in use            |
| Dublin Core                          | in use            |
| METS                                 | project - test    |
| Z39.50                               | in use            |
| e-gif                                | watching brief    |
| e-gms                                | watching brief    |
| NGDF                                 | watching brief    |
| HE/FE geo profile                    | watching brief    |
| Data Protection Act                  | legal requirement |
| Freedom of Information Act           | legal requirement |
| Web design standards <sup>56</sup>   | legal requirement |
| ISO & TNA standards for preservation | in use            |
| Internal quality control standards   | in use            |
| IMS                                  | in use            |
| UKCMF                                | in use            |
| ISO 19115                            | watching brief    |
| FGDC                                 | watching brief    |

**Table 1. Status of metadata standards used in the UK Data Archive.**

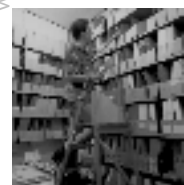
### Information model: TNA

It is broadly possible to map the OAIS information model to the information packages in use at TNA. The OAIS reference model informed the design of the Digital Archive system at TNA, in addition to other standards like the Government Metadata Standard (e-GMS).



**Figure 13. A simple metadata model used at TNA.**

<sup>56</sup> These are requirements for the presentation of material.



Information in the collections of the National Archives is managed per record accessioned, which may be a hierarchy of records. Information is first grouped at the record level. Each record may contain a series of manifestations (e.g., migrated versions of the record). Finally, each manifestation consists of a number of files (the OAIS digital objects) constituting the record. Extensive metadata are kept at each level.

The Digital Archive accepts submissions as a collection of files, arranged in a hierarchical file system consistent with the limitations imposed by CD-ROM and DVD file systems. The files consist of the digital objects making up the record, and the single XML metadata file that contains all the metadata surrounding the record known so far.

At this stage it is expected the SIP will contain information for most of the information categories as defined in the OAIS reference model, except migration histories and other entries that relate to TNA work processes.

When a SIP has been received, the objects are validated against the information contained in the SIP. Catalogue enhancements may be made and additional context and reference information may be added, along with any additional provenance information (e.g., audit trail of processes applied to the record). The digital objects making up the record are not touched in any way during this process: TNA treats these objects as the original manifestation of the record that can be produced in support of the presumption of authenticity.

They are read in order to be scanned for virus infection and copied to serve as the basis for immediate DIP creation. This forms the AIP that is loaded into the digital archive. In many cases the final DIP creation will take place before the AIP has been loaded. This serves the dual purpose of having a browsable, navigable record immediately available, to open to the public or to serve in answering FOI enquiries, and also to validate the structural integrity of these original digital objects through migration to presentation formats.

The Digital Archive system also adds some technical metadata during the loading process.

A DIP is produced from the AIP by technical staff, migrating the digital objects into a form that can be delivered and viewed using open standards or through widely available software on the Internet.<sup>57</sup> For example, a WordPerfect file may be converted into a PDF file for use over the Internet. Information about the migration process is gathered, along with provenance, context, reference and descriptive information. Together these form the DIP.

A full TNA metadata model is presented in Figure 14. While this model may initially look quite complex, the metadata elements it includes broadly map to the OAIS information objects. It can be seen that the OAIS information packages also broadly correspond to the categories of TNA metadata, with a few exceptions, notably the Record Content, which takes information from several OAIS packages.

All of the information packages at TNA are managed inside a single XML file per manifestation which contains all the record metadata for that manifestation. Some of the entities in the XML file are extracted into relational tables for faster querying or update but the more invariant data are recorded directly as XML using Oracle 9i's XML relational technology.

The model does not naturally decompose into the separate information packages of SIP, AIP and DIP as OAIS suggests, as the model is explicitly supposed to encompass the entire workflow of managing the digital objects. It is, nevertheless, possible to identify these logical packages of information within the metadata model.

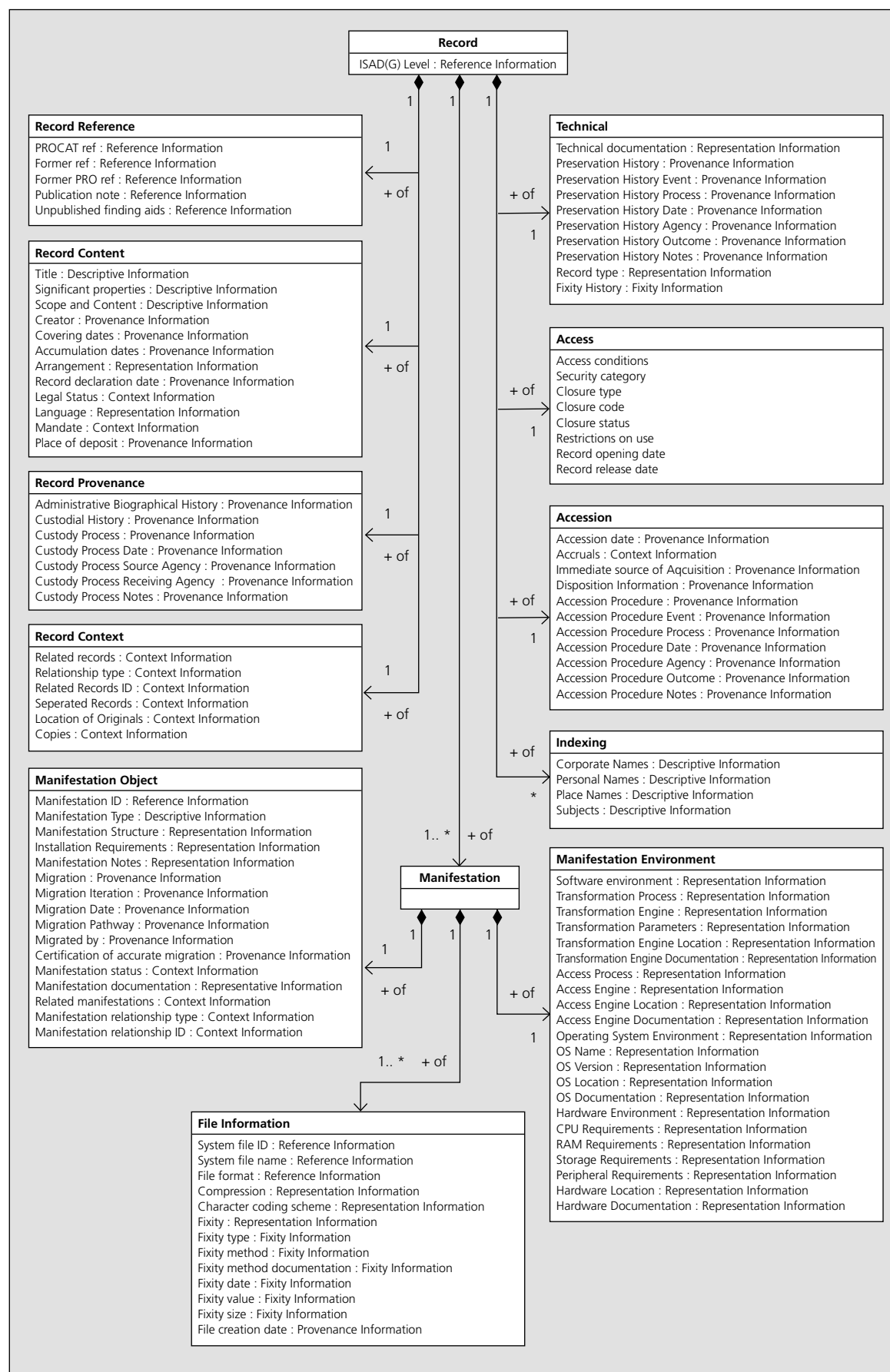
<sup>57</sup> Note that at present the objects are only accessible via TNA reading rooms – the application is still being readied for release on the general Internet.



When developing its digital preservation metadata standard, TNA's working group found that the OAIS reference model has very little to say about the structure of DIPs and does not attempt to model much of the ongoing management metadata that are required for a functioning archive. For example, the migration of records and metadata required to document activities related to it, is not modelled in the OAIS. Therefore, TNA metadata do not map directly to the OAIS concepts but, instead, follows the workflow of the Digital Archive system.

For a sample extract of the mapping of TNA digital preservation metadata and OAIS information model see Appendix 3 of the report, although note that this data model is in active development to take account of advances in the modelling and management of multiple manifestations.

Figure 14. The National Archives digital preservation metadata model





## OAIS Information Model: Conclusions

The OAIS information model is instructive with its categorisation of information required for preservation of digital objects. The description of these information categories, however, sometimes remains too general to allow the placing of metadata elements with confidence into one or other category. Therefore, the mapping exercise of metadata elements to the OAIS information model, includes some subjectivity.

Both institutions participating in the compliance testing have traditions of using their own specific metadata standards which they also helped to develop. Some of these standards (e.g., DDI, ISAD(G)) are used internationally and are widely accepted within the archival domain. These have, however, not yet been mapped to the OAIS reference model in the way the OCLC/RLG working group has done with the digital preservation metadata standards developed in the library domain. Undertaking such an extensive metadata mapping was outside the scope of the current project.

A general level metadata mapping does, nevertheless, demonstrate that the digital preservation and collection management metadata used at both institutions does correspond to OAIS's understanding of information objects. The information packages (SIP, AIP and DIP) are clearly defined at both institutions and are handled, processed, stored and accessed separately. Both institutions enhance metadata and thereby add value to the SIPs they receive. The majority of these additional metadata are directed towards helping users to retrieve and interpret the DIPs but some are added to document the preservation processes applied to AIPs.

It was felt that a more thorough mapping between the DDI and ISAD(G) metadata standards and the OAIS reference model might be undertaken in the near future (if resources allow) and an international discussion started from that basis.

TNA have found the OAIS reference model very useful, but in regard to the categorisations of metadata, they are extremely broad, functionally organised (as one would expect), and do not reflect the way metadata are packaged and used across particular archival practices.

TNA have instituted a working group to look at the provision of other levels of metadata categorisation alongside the OAIS definitions, in accordance with international standards where they exist, and to analyse where particular metadata categories are created, used and managed. Defining these responsibilities and information flows is vitally important in a functioning archive, so for TNA at least, the process of mapping to OAIS metadata was interesting but not directly helpful, as our own models and metadata categorisation system were already more complex than OAIS as a reference model can ever be.

This possibly reflects a basic tension in the OAIS – its need to be general outweighs any attempt to be specific – and for some naturally detailed and specific areas (like metadata), it must be very difficult for OAIS to say anything of much interest to a functioning archive with established systems.

As a starting point for a new archive, it has very much more value, and indeed, informed the original design of TNA's metadata, along with other standards like the Government Metadata Standard and ISO 15489 but was nowhere near complete or detailed enough to be directly usable as a metadata categorisation system.



## 7 USING THE METS METADATA STANDARD IN A DIGITAL ARCHIVE

### 7.1 INTRODUCTION

The Metadata Encoding and Transmission Standard (METS) is a structure used for encoding and wrapping administrative, descriptive and structural metadata for a digital object. METS is a flexible and software independent platform, using XML, which provides the potential for interoperability between metadata schemes by providing a framework for integrating various types of metadata. METS promotes interoperability of descriptive, administrative and technical metadata and provides the means for digital repository management. It uses XML, which ensures that it is interchangeable. However, it does require XML-compliant software to be used with it.<sup>58</sup>

The METS format is an attempt to provide a standard but flexible format to hold the diverse metadata associated with a digital object in a form in which it can easily be shared, exchanged, searched and rendered for browsing and display purposes. METS is intended primarily as a flexible, yet tightly structured, container for all metadata necessary to describe, navigate and maintain a digital object (descriptive, administrative and structural metadata). All metadata relating to a single digital object (which could be either a single image or all the components making up an item as complex as a digitised volume) are integrated into a single file: within this file, each type of metadata is described in a separate section, which is linked to its counterparts by a comprehensive system of internal identifiers. The metadata themselves may be held physically within the METS file, or may be held in external files and referenced from within the METS document: they may follow any preferred scheme, although a number of these are recommended specifically for use within METS. METS is written in XML, a generic language designed for marking up electronic texts.

METS aims to address the lack of standardisation of library metadata by encoding images, video and sound files and texts and can be seen to be the equivalent of 'collection level' metadata.<sup>59</sup> However, it does not ensure that metadata content is standardised as it is not a metadata standard, but rather a framework within which to organise existing metadata. An attempt has been made to extend the METS through recommendation of extension schemas to allow for the standardisation of specific types of metadata.<sup>60</sup>

### 7.2 THE METS DOCUMENT

A METS document consists of seven major sections.

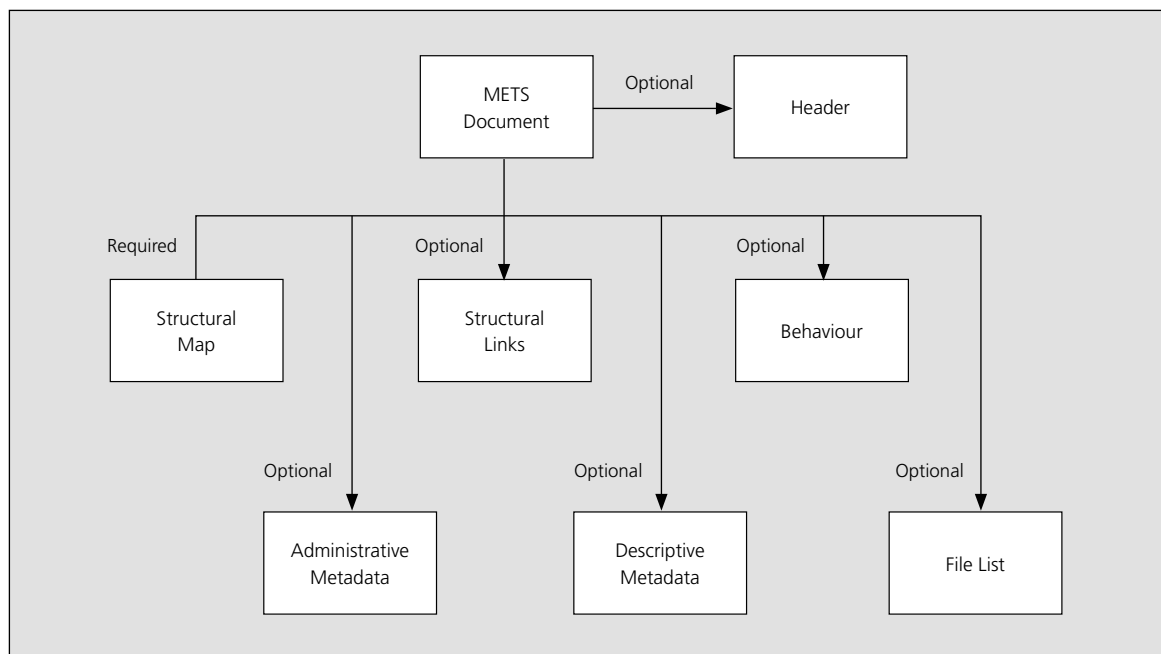
- **METS Header** – the METS Header allows the recording of minimal descriptive metadata about the METS object itself within the METS document. These metadata include the date of creation for the METS document, the date of its last modification, current status for the METS document, the names of one or more agents who have played some role with respect to the METS document, the role they have played, and a variety of alternative identifiers for the METS document to supplement the primary identifier.<sup>61</sup>
- **Structural Map** – the structural map is the heart of a METS document. It outlines a hierarchical structure for the digital object that is archived and links the elements of that structure to content files and metadata that pertain to each element. The METS <div> element could be used to describe how a book is divided into separate chapters and if these chapters themselves contain sections or subsections, it will show how these are nested together as each of these divs points to its component files in the file group.

58 Waibel, 2003

59 L'Hours, 2004

60 Gartner, 2002

61 cf. LOC, 2004

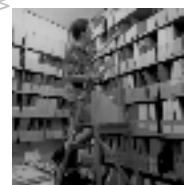


**Figure 15. Structure of a METS document**

- **Structural Links** – the structural links section of a METS document is the simplest of the METS sections, containing only a single element, <smLink>, although that element may be repeated as necessary. The structural links section of METS is intended to allow the existence of hyperlinks between items within the structural map, usually <div> elements. This is a useful facility for archiving, for example web sites, and for maintaining a record of the hypertext structure of the sites separately from the HTML files of the site itself, in METS.
- **Behaviour** – the behaviour section in METS allows information to be recorded on how components of the digital object should be rendered for the user. This may include information on specific software packages to be used or on particular parameters to be used when rendering a file. A behaviour section contains one or more <behaviour> elements, each of which has an interface definition element that represents an abstract definition of the set of behaviours represented by a particular behaviour section.
- **Administrative Metadata** – the administrative metadata section can include information about how the digital objects have been created and stored, IPR, metadata about the original source object (e.g., an analogue publication) from which the digital object derives, and information regarding the provenance of the files comprising the digital object.

There are four main forms of administrative metadata provided for in a METS document:

- Technical Metadata (information regarding file creation, format and use characteristics);
- IPR Metadata (copyright and licence information);
- Source Metadata (descriptive and administrative metadata regarding the analogue source from which a digital library object derives);



- Digital Provenance Metadata (information about source/destination relationships between files, including master/derivative relationships between files and information about migrations/transformations employed on files between original digitisation of an artefact and its current incarnation as a digital object).

Each <amdSec> is labelled with a unique ID to allow it to be referenced from within the structural map or the file group. Each of these four elements may occur more than once in any METS document.

- **Descriptive Metadata** – this section may point to descriptive metadata external to the METS document, or contain internally embedded descriptive metadata, or both. Multiple instances of both external and internal descriptive metadata may be included in the descriptive metadata section. The descriptive metadata consist of one or more <dmdSec> elements. Each <dmdSec> is labelled with a unique ID to allow it to be referenced from within the structural map or the file group. An <mdRef> element provides a URI which may be used in retrieving the external metadata.
- **File Section** – the file section lists all files containing content which comprise the digital object. <file> elements may be grouped within <fileGrp> elements, to provide for subdividing the files by object version. Each file element provides an ID and the physical location of the file (usually a URL). The ID of the file is referenced from the structural map. Different manifestations of the digital file may be grouped together in surrounding <fileGrp> elements, and so referenced together if necessary.

According to the METS web site, users may select extension schemas to provide additional metadata and make them more exact. The Library of Congress Audio-Visual Prototyping Project began in February 2003 and has made use of such extensions.<sup>62</sup>

## 7.3 USES OF METS

METS was created with a digital library environment in mind and although there are no strong reasons why it cannot be applied to an archival system, examples are difficult to find. A few examples of projects which use METS are described later in this chapter. METS was initially intended to describe metadata for images, text and audio. Many of the projects currently in progress are using METS to describe collections of still images, mostly digitised images of books and artefacts.

Following its initial purpose, a METS document is essentially a 'hub' that draws together potentially dispersed but related files and data. METS uses XML to provide a vocabulary and syntax for identifying the digital components that together comprise a digital object, for specifying the location of these components, and for expressing their structural relationships. The digital components comprising a digital object could include the content files, the descriptive metadata, and the administrative metadata. This feature makes a METS document compatible with the OAIS information package concept and indeed, using METS for defining, creating and managing SIP, AIP and DIP in a digital archive has been suggested. The role of METS as a container format for information packages could be manifold, for example:<sup>63</sup>

- **Submission Information Package (SIP)**  
METS as transfer syntax, providing a standard for transmitting or exchanging digital objects; METS can be used to transfer data from local systems to union systems by pulling together data and encoding them into highly portable markup;
- **Archival Information Package (AIP)**  
METS documents can be stored internally in an archive, providing a standard for archiving digital objects; METS includes an extensible section on administrative metadata to allow a digital object to carry information such as IPR or technical metadata;

62 See <http://www.loc.gov/rr/mopac/avprot/metsmenu2.html> for more details

63 Cf. Waibel, 2003; Proffitt, 2003



#### ■ Dissemination Information Package (DIP)

METS as transfer syntax, offering the basis for providing end users with the ability to view and navigate digital content and its associated metadata; the XML mark-up can be used as a media player through the use of a METS viewer utility. METS as input to display applications with a structural map that divides the object into distinct sections.

However, the adoption of METS is not yet widespread, particularly in the UK. A number of current projects are exploring the application of METS to their work but are not yet sufficiently advanced to offer comments on its use. For those that are, a key concern is that associated with any new standard: there is a reticence to commit entirely to a standard that has not yet achieved the critical mass required to ensure its longevity.

Of these, the Oxford University's Digital Library is using METS as the central component of its metadata system and has established detailed procedures for mapping metadata from legacy digital projects to the scheme. The British Library and the National Library of Wales both plan to implement METS for their digital collections.<sup>64</sup> The AHDS History are currently using METS as a wrapper to export data for their Online Historical Population Reports project. The Digital Curation Centre is considering using METS to encode archival objects into SIP's, AIP's and DIP's.<sup>65</sup>

Little analytical information could be found about the implementation of METS, probably because it is relatively new and its use is not currently widespread particularly amongst archives. However, some of the UK-based projects which do use METS have been described in the section below.

### 7.3.1 AHDS HISTORY: ONLINE HISTORICAL POPULATION REPORTS

The JISC funded Online Historical Population Reports (OHPR) project is being undertaken by AHDS History, based at the UK Data Archive, University of Essex. The core aim of the project is to digitise Census Reports (England, Wales, Scotland, Ireland 1801-1933), Registrar General Reports (England, Wales, Scotland, 1801-1920), associated material from TNA and a number of other parliamentary papers (including relevant legislation). This material will then be made publicly available via the web with special emphasis on HE and FE users.

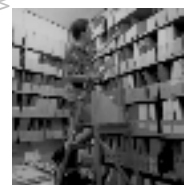
In addition to the public provision of the data the project provides for long-term archiving of data and metadata, including harvesting, interoperability and future format migration.

JISC provided the project with a number of recommended Metadata standards including DDI, TEI, MODS, MIX and METS. METS is still a relatively new format gaining currency within the archiving community and as such the OHPR implementation will be an experimental opportunity to evaluate the standard. It serves as a 'wrapper' for other metadata standards and provides a method of describing the structure of the digitised data. As METS had been recommended by the project funder, the main aim was to confirm that it met all of the project requirements rather than to evaluate METS in comparison to other standard 'wrappers'.

Early in the course of the project the project team needed to make a decision on whether its development cycle should be XML or database led. Opinions were canvassed at a number of digitisation and XML meetings and conferences and the consensus was that a database-led approach was more appropriate to the long-term aims and the available expertise. It would be technically possible to develop all of the metadata for the project in XML form but the sheer quantity of hand-crafting required and the relative poverty of mainstream XML-enabled databases made this impractical. The OHPR approach is to store all of the data required, whether for the web site, archiving or interoperability, in a database. Thus it is clear that multiple METS documents will not be manipulated as part of the core metadata creation process.

<sup>64</sup> Gartner, 2002

<sup>65</sup> Patel, 2004



At this stage of the project (June 2005) the major task is to create the structural metadata to describe 200,000 images of printed pages from 700 volumes of population reports. These data are being made available incrementally via a demonstration web site<sup>66</sup> but the creation and output of XML versions of the metadata (descriptive and administrative) will not take place until a much wider selection of the raw data has been examined. In the latter stage of the project the database contents will be exported into specialised XML templates covering all of the required metadata standards.

The core of any METS document is a structural map which represents the data in a series of nested <div> elements. Each of these elements may possess a unique identifier and a pointer to files associated with it. In the case of the OHPR project these <divs> could be at the level of a single Census volume, describing the structure of each section and sub-sections (introductions, summary tables, tabular data by county) or page. A single page reference may contain pointers to data relating to the page in multiple formats including tiffs, OCR text and tabular versions. METS may also be valuable at collection level as the structural map could equally describe the sequence of Census or Registrar General Reports over the period described.

Each <div> may contain links to the Descriptive and Administrative metadata sections of the METS document through unique metadata IDs. These metadata sections may be repeated as often as is required and either point to external instances of metadata records or 'wrap' an instance of the record within the METS file itself. In the case of OHPR these metadata could be a DDI instance of a table associated with a <div> or a MIX document describing everything from the basic image parameters of a tiff to the details of the scanner used to create it.

It is clear that much of the structural content described is required for the provision of a database to support browsing and searching of the content via the web site. Equally there is no barrier to the storage of the more technical metadata elements in the database. Once the bulk of the metadata are created and most of the potential anomalies that may affect the output of clear structural maps have been detected, METS versions of the content will be made available. In addition to providing a suitable format for archiving as a flat file, these data may also be made available for interoperability or harvesting purposes.

The OHPR, database approach provides a key benefit in that changes may be made during and even after the lifetime of the project to match the changes in metadata standards. For example, the MIX standard was only a proposal when this project was initially funded but the team is confident that, barring any fundamental changes in the content requirement, they will be able to output MIX that matches the latest requirements of the standard.

METS provides an effective means of describing the structure of the content and 'pointing' to related metadata. However, there have been problems sourcing detailed documentation from projects implementing the standard and although this may change if the standard becomes more popular, it does appear that a critical mass of projects will be required to drive effective interoperability between METS users. This also raises the question of just how similar underlying datasets would need to be for effective interoperation using METS.

### 7.3.2 CULTURNET CYMRU: BOOKS FROM THE PAST

Books from the Past is an online collection of books from Wales, of national cultural interest which have long been out of print, and are unlikely to be reprinted by traditional means. The web site is in Welsh and English and uses METS for structural metadata and for linking image files to page divisions within the texts.<sup>67</sup> It also uses TEI for the full text content of the books. The software is based on Greenstone Digital Library Software, but with some customisation and additions, including a METS/TEI to GAF converter. The texts are available in two forms – images of the original book pages, together with a fully searchable electronic text which is also suitable for printing.

66 [www.histpop.org](http://www.histpop.org)

67 <http://www.booksfromthepast.org>



Culturenet digital library add-ons to Greenstone are being developed to include a METS/TEI to Greenstone Archive Format converter and Welsh/English functionality. The software being developed should be released as open source.

While transcribed electronic text in TEI would satisfy the searching and accuracy criteria for the project, it would not be sufficient (on its own) for the presentation of the original orthography and layout. The nature of the web means that layout and presentation of web pages is controlled by the user. Ideally the project team wanted to present to the user an image of each page and electronic text of each page. For this purpose, a metadata system was required which could link an image of each page to the encoded text for each page in a given TEI document. A metadata system was also needed which could wrap up all the different parts which make up each electronic book (structural metadata). METS was chosen for this purpose.<sup>68</sup>

### **7.3.3 LLYFRGELL GENEDLAETHOL CYMRU / NATIONAL LIBRARY OF WALES: Y DRYCH DIGIDOL / THE DIGITAL MIRROR**

Y Drych Digidol/The Digital Mirror contains the outcomes of the National Library of Wales digitisation programme.<sup>69</sup> METS is being used for management of files and in the presentation of complex digital objects, which have been reformatted for access as part of the Treasures programme and which require a page-turning facility. The MARC21 standard is used for the main descriptive and preservation metadata. Delivery is based on an adaptation of one of the METS viewers developed by Leslie Myrick at New York University. Since 2004, an automated system has been developed for generating METS documents and there have been further adaptations to the METS viewer to deal with the requirements of different object types. This project resolved the problem of a lack of tools by developing their own which it is their intention to make available as open source.

### **7.3.4 OXFORD UNIVERSITY: OXFORD DIGITAL LIBRARY**

The Oxford Digital Library is using METS to complement their image files with searchable electronic texts and to provide integrated access to both types of digital resource. METS forms the central metadata for the Oxford Digital Library which is currently being established using materials from libraries within Oxford University.<sup>70</sup> An in-house MySQL/PHP system of web forms is used for cataloguing, using qualified Dublin Core fields. METS files are automatically generated by this system when the metadata input is complete with the DC fields being converted to embedded MODS records for descriptive metadata. This system has been in operation since 2002.

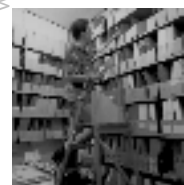
### **7.3.5 THE UK DATA ARCHIVE**

The UK Data Archive has been considering METS in the context of the DDI metadata standard. The DDI Structural Reform Group (SRG) has been asked to recommend a conceptual model that can be used to produce a future version of the DDI in certain formats. Initially this will be an XML Schema. This conceptual model is based on the life-cycle of a dataset from concept to recycling, with different types of information being added at each stage. It also wanted to break down these stages and the existing document type definition into a more workable and extensible modular model and has therefore been looking at METS as a method of linking or wrapping these individual modules.

<sup>68</sup> See <http://sunsite.berkeley.edu/mets/registry/>

<sup>69</sup> [http://www.llgc.org.uk/drych/index\\_s.htm](http://www.llgc.org.uk/drych/index_s.htm)

<sup>70</sup> <http://www.odl.ox.ac.uk>



The DDI team is particularly interested in two sections of a METS document: the structural map; and the structural links. The DDI SRG may also be considering using the administrative metadata. However, an alternative approach may be to learn from and adapt METS rather than adopt it fully because it does not adequately describe the richness between nodes and because xlink may be a barrier. To date, no decision has been made on whether or not to use METS.

### 7.3.6 THE NATIONAL ARCHIVES

From TNA's perspective, the need to conform with other metadata standards (e.g. eGMS), government interoperability standards (eGIF) and, pragmatically, to produce a working model rapidly has meant that METS has not played a significant part in metadata management at TNA. TNA remains interested in METS and may implement it at some point in future if a clear driver to do so arises.

## 7.4 STRENGTHS AND WEAKNESSES OF METS

The METS schema provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital object, and for expressing the complex links between these various forms of metadata.<sup>71</sup> METS shows great potential as a file-exchange format and can provide a useful standard for the exchange of digital objects between repositories.<sup>72</sup> It is written in XML, which is a widely accepted interoperable standard for the exchange of metadata. One of its major advantages is that it is software independent, another is that it is a robust archival medium and human readable.<sup>73</sup>

In the majority of cases, as described above in various projects, METS is implemented to wrap several files that belong to one conceptual object into a single METS XML file. The hierarchy of the METS structural map enables the navigation of files embedded in, or referenced by, the METS object and by browsing, for example, through individual pages of a book or by jumping to specific segments. If XML metadata documents are output from a database (for archiving or interoperability purposes), one huge METS file could be created, either with various metadata wrapped inside it or with hundreds of files which are all managed (pointed to) from within the METS file.<sup>74</sup>

METS has a clearly delineated internal structure which is easy to generate automatically from other applications (such as relational databases), although this results in it being less verbose than other more traditional XML applications. It is designed to be flexible and easy to use as it offers the choice of retaining the metadata internally or externally. As METS is also designed to be flexible and extensible, it provides the opportunity for future use within the digital library community. However, its degree of flexibility could prevent widespread applicability as a standardised container if the content of the metadata is not standardised, and documented use cases are not provided.

XML is being used by a number of digital library projects but it is clear that no agreement has been reached on a standard, as the choice of DTD's used has been diverse. To a certain extent, many of these DTD's can be mapped to each other allowing for the exchange of metadata but these are almost always partial mappings and require negotiations between organisations to encourage metadata conversions. METS is attempting to satisfy the need for a universally-applicable standard that will incorporate all the main types of digital object metadata.<sup>75</sup> The recommendation of extension schemas for use with METS attempts to address some of these issues. External XML schemas, such as Dublin Core can also be incorporated into the METS schema.

71 Waibel, 2003

72 LOC, 2004

73 Gartner, 2002

74 L'Hours, 2004

75 Gartner, 2002



One view of METS is that it offers no more than a framework into which metadata can be added and this flexibility provides its greatest weakness. Its full potential as a standard will only be realised if it becomes widely accepted along with the use of standardised metadata content. The projects which have been reviewed in this document have discovered their own solutions to the problems of metadata content, by developing their own automated tools for use with METS. This shows that some effort and input is required in using METS and this also suggests that there is a lack of usable standardised tools available.

The need for better articulation of the uses of METS is clear. The standard may be flexible enough to provide solutions to a number of general problems but this benefit to an organisation has to be weighed against the investment in time required to understand such a large and flexible standard. This is especially likely to be a problem in institutions that are already applying established metadata standards. Additionally, the lack of standardised tools is clearly a barrier to adoption.

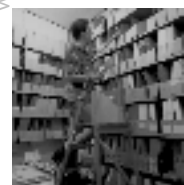
The potential for aggregating the metadata required for different purposes, such as resource discovery, rendering, processing and preservation, into one METS document to act as an OAIS information package, has not been realised sufficiently in practice. METS can contribute to but cannot guarantee interoperability, usability and longevity of digital resources because it only provides a container for the metadata that supports these features. These metadata have to be standardised first.

However, all new standards are likely to be hampered by lack of widespread adoption and a dearth of tools during the development period. The success of new standards often lies in the recognition of the need for such a standard by a committed group of organisations that is prepared to test and adopt it through a gradual process, as happened with the DDI and the community of data archives. It is clear that METS is able to solve many problems that digital libraries face, even though, at this stage, it does not necessarily resolve problems of interoperability. It is, however, offering a starting point towards interoperability (as only one problem an archive or library may wish to solve) even though much more is required. If the value of METS is framed entirely in terms of interoperability, then a critical mass of users will be required before any such benefit can be seen. If a critical mass develops, then so will tools and the standard will survive.

In terms of digital preservation for born digital material, there is the potential to model the metadata contained in the information packages of the OAIS model but there is no clear driver to do so. The metadata categories in OAIS are so broad, and are not organised in terms of creation, usage or management, that it is hard to see what the specific value of doing so would be. Practically, this would constitute an application of METS, in the same way that METS is an application of XML. This would result in two standards having to be absorbed and the individual metadata structures would still have to be defined.

Where there is a need to capture the internal structure of a record for internal management purposes, especially at the point of creation, it will provide a useful standard. Again, this is clearly evidenced in digitisation programs, where, for example, a single document may be turned into a variety of digital assets which need to be linked to one another.

It may be that METS is useful in modelling the relationships of internally generated digital assets, and in processing such assets from other institutions, but may have less value in modeling complex semi-structured records which are created externally and already have appropriate, non-METS records metadata defining them as records.



## 7.5 METS: CONCLUSIONS

METS has its origin in the digital library and archives world, where it has definite potential as a framework, particularly for data. It offers significant flexibility for linking together existing blocks of metadata and managing them as one collection or outputting them as a single XML file for exchange or retrieval. However, given that it has not, as yet, been widely adopted, the interoperability value is limited at present.

This presents a particular problem between organisations that do not use the same tools and metadata content standards. The flexibility of METS has resulted in its being considered, most usefully, as a wrapper for various bits of existing metadata and for metadata following other standards. Nevertheless, the use of METS is only a starting point for interoperation, given that it wraps arbitrary metadata. Interoperation alone cannot be the primary driver for its adoption and it must be acknowledged that METS does solve some of the problems of managing and transmitting collections of structured information. Hopefully its increasing adoption will lead to standard tools which can be shared, and lead towards the possibility of greater interoperation in the future.

Both of the institutions participating in this project are using metadata standards defined by their own legal and best practice requirements. The UKDA implements a policy of 'standards watch' and is keeping a watching brief on the uptake of METS whilst undertaking limited development to permit the output of METS documents for harvesting through the OHPR project. It is also considering the potential of METS, as a container, in the wider context of its application of the DDI. As with any new technology, the UKDA does not adopt new standards until they are considered robust and workable and there is some confidence in their longevity.

The obvious reason for this is that the UKDA has key responsibilities for both preservation and dissemination and a feature of its systems and processes must be the assurance of continuity for both. The early adoption of unproven standards has the potential to jeopardise the organisation's ability to carry out these responsibilities effectively in the long term. Nevertheless the UKDA does recognise that the uptake of such new standards by established archives may act as a spur to their wider adoption and is therefore mindful of the importance of giving due consideration to emerging standards.

Like the UKDA, TNA has other metadata standards which take precedence (e.g. eGMS). An investigation of METS by TNA concluded that it was an interesting wrapper format, but that the organisation had to define its own metadata formats and had no interoperability requirements with other institutions that required METS.

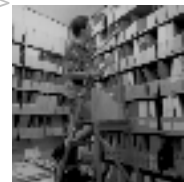
One reason for this is that many digitisation programmes pull apart the structure of existing material to capture it, then it needs to be capable of reassembly for presentation and management. So far, TNA has not needed to pull apart the existing born digital records in quite the same way, as born digital records are already typically structured in the digital formats immediately needed, allowing for relatively straightforward format translations or possessing representation information that allows for their reconstruction and repurposing.

It is still possible that TNA may adopt METS at a later stage, particularly as a packaging format for SIPs, AIPs and DIPS, but in the initial stages of building the archive, there was no great value in adopting a format which had not attained the status of an accepted international standard, with no other direct driver to invest in understanding and implementing it. This highlights the point made earlier, that METS looks to be a very capable and flexible metadata wrapper format but its very flexibility can get in the way of being able to articulate exactly what it is useful for. Possibly this is because it solves a problem which is evident in the digital library world, but less so in the born digital archive world - the pulling apart (for digitisation) and re-assembly of logical record structures.



The lack of good documentation, use cases and worked examples is a particular problem that has been identified and has been problematic both in terms of preparing this report and, from the experience of the OHPR project, when actively considering a METS implementation. The very flexibility and abstraction of METS demands material that would help a potential customer of METS to evaluate its use and determine whether it solves real problems or provides additional benefits.

METS is less useful when dealing with born digital material, although it does have potential for being used for AIPs, SIPs and DIPs. Once a real-life interoperability situation emerges, for example for AIPs, it would be useful to test METS in this situation but within the framework of this project, there was no room for mapping the existing metadata and creating METS-based OAIS IPs. Thus detailed mapping fell outside the scope of current project.



## 8 FINAL CONCLUSIONS AND RECOMMENDATIONS

Whilst specific conclusions have been drawn within sections of this report, a number of these are worth repeating and other, more general conclusions are added in this section. Conclusions relating to the METS standard are not included in this section as they have been treated separately in the previous section.

The outcome of the compliance testing of mandatory responsibilities was good for both organisations in that it did, as hoped, confirm that the practices of each organisation conform to, and exceed the requirements of, the reference model. This latter finding, that both organisations exceeded the requirements, is considered to be important because at a very early stage it became apparent that both the UKDA and TNA would conform to the OAIS reference model based on a naïve approach that to conform means achieving all the mandatory requirements. Insofar as the OAIS model is only for reference and is designed for all types of archives, it becomes clear that any institution with a responsibility for preservation could meet these high level requirements. It was evident, therefore, that a deeper investigation would be required to determine the extent to which the organisations comply with the functional model. Indeed, only by taking this approach could TNA and UKDA offer useful or critical comment on the application of the model to other working archives. Nevertheless, a decision was also made that neither organisation would undertake a very detailed mapping of its systems to the OAIS reference model or of its metadata to the METS standard. The time required for this would be extensive and is not covered by the scope and funding of this award. It is, however, something that might be considered for the future if further funding were to become available.

The biggest discrepancy that the testing of mandatory responsibilities revealed was with the OAIS concept of the Producer and the Designated Consumer Community. The OAIS model frequently points to the strong link between the user community and the way the material in the archive should be described and preserved. In reality, however, it is often difficult to limit the user groups or communities to groups as narrow as the OAIS standard examples. As already stated, both partners to this project are of the view that the OAIS has an inbuilt limitation in that it overly assumes an identifiable and relatively homogeneous consumer (user) community. This is not the case for the user communities of either the UKDA or TNA.

Different archives are also likely to have varying levels of control over the material provided by information producers, as described in chapter 7.1 for the UKDA and TNA. Consequently, there are also likely to be differences between organisations at the pre-ingest and ingest stages, in terms of what processes an organisation must undertake. For example: it may only be necessary to check for the existence of metadata; it may be necessary to verify the metadata; or there may be a need to create new metadata and to convert file delivery formats.

Other discrepancies were noted when compliance of functions was tested. In terms of the archival storage function, the strict separation of the storage of archival preservation copies (AIPs) from the storage of dissemination versions of the archived material (DIPs) is the biggest discrepancy from the Reference Model. Access to the archive's collections represents an interaction with the DIP storage, rather than the creation of the DIP 'on the fly' from the AIP retrieved from storage, as recommended by the OAIS standard. As explained, the reason for this is that both TNA and UKDA are following archival tradition in which original records are often copied and only the copies are made available to users. This important mechanism to protect the archival version of the original record is not explicitly mentioned in the OAIS reference model although at the same time it is not precluded.



Consideration of the data management function proved to be considerably more time-consuming than originally planned. The reason for this was that, for each of the archives, comparison of the data management functions with the OAIS model resulted in the potential for numerous mappings to all the different systems and processes that an archive undertakes. The OAIS presents this as a simple, contained function but in practice it is highly distributed.

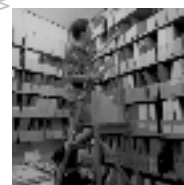
The reference standard places much emphasis on the administration function, in particular on the monitoring and management of decision-making processes associated with preservation planning. Whilst it is acknowledged that effective administration is essential for good archiving practice, the UKDA in particular considered that this is an area where it may be difficult for small archives to match the reference model and one where it is difficult to see where the model offers scalability for professional but less well resourced organisations. The model makes implicit assumptions about technical infrastructures (by assuming the full or partial automation of processes) when in reality the technical infrastructure of a small organisation may be extremely limited.

Consideration of the access function demonstrated that the UKDA and TNA have various limitations on access to their collections and each has its own mechanisms for controlling it (e.g. user agreements, FoI requests, unrestricted access). These controls have developed over time and fit into a particular legal and regulatory framework. The OAIS model has to be general and cannot accommodate all possible methods of provision, control and limitation of access to an archive's holdings. Hence, for large organisations, the mapping becomes unproductive because in the real world such organisations are very likely to require much more sophisticated and specific processes than the reference model requires. It is acknowledged that the interaction with the file storage and the Administration's oversight of the access regulations needs to be there but the actual mechanisms for providing access are often much more complex in the working environment.

In some cases, it has been possible to see where the OAIS model does not provide enough detail to make good statements about the archive's operation. This is particularly so in the areas of ingest, access and migration. This is not a weakness of OAIS as it is attempting to be very neutral about the actual preservation and access implementation methods but it does mean that OAIS does not provide a conceptual map or a common terminology for some very important digital preservation areas common to both TNA and UKDA, and to many other archives.

One of the unexpected outcomes of the work has been the utility of the OAIS language as a means of communication between the UKDA and TNA. The working terminologies of each organisation is different but as dialogue progressed each found that it was useful and less confusing to use OAIS terminology as a common means of communication. This result is likely to be opportune for both organisations as the UKDA was designated an official place of deposit for TNA in 2005. TNA also found the terminology to be useful for cross-divisional discussion and is encouraging its wider use internally.

One concern is that the model is not obviously scalable. A possible solution is to consider the development of an 'OAIS Lite' to help smaller archives gain value from it as a reference model. As it stands, the model is daunting and requires a substantial investment in understanding and mapping to make use of it. It is not always clear what the really important concepts are, as they are mixed in with a very heavy level of functional detail once the rather basic mandatory responsibilities have been left behind.



The task of mapping the UKDA and TNA functions to the OAIS reference model would have been easier if there had been a guideline or manual to aid and inform the process. Such a document could now be produced, drawing on our experience that the model is very general, designed to be non-implementation-specific, and includes entities that have very little knowledge built into them and little knowledge of other external entities or systems. The main body of knowledge resides in the administration and preservation planning entities of the OAIS reference model which represent what an archive normally does. In reality, these responsibilities are spread across various units of an archive, some of which also perform responsibilities that are included in other OAIS entities rather than as discrete functions (e.g., ingest). Essentially, the OAIS model has a clever design with functional entities exchanging the minimal amount of information needed to perform their duties, however, an archivist looking at the model needs to be aware that a real-life archive can do more within a function than the OAIS prescribes.

For anyone considering a comparison of their systems to the OAIS reference model the watchwords must be 'resource intensive'. Both organisations found the exercise to be more labour intensive than expected and invested greater resources into the project than had been originally anticipated.

Nevertheless, the consensus is that the exercise has been valuable. The fact that both organisations can now say with confidence that they conform to the model gives formal and demonstrable support to claims that were previously based on reputation and the assumption of 'trusted archive' status.

A particular value of undertaking such exercises, especially when the results are made publicly available, is that it demonstrates a willingness on the part of organisations to open their processes to scrutiny by their peers and interested individuals. Organisations such as TNA and the UKDA have nationally recognised custodial responsibilities and should be expected to be transparent in their approach to their work.

Whilst it is acknowledged that this process has been both time-consuming and resource intensive, it is suggested that others who wish to undertake such work might wish to consider the adoption of a similar methodology. From the perspective of both TNA and the UKDA, the methodology has been very successful and it has been possible to map most digital archive activities to the OAIS reference model, although it is complicated by the fact that many activities cross over established organisational boundaries, as both organisations were formed before OAIS. The approach provided a clearer understanding of why both organisations are doing what they do in the particular ways in which they are doing it.

For TNA, the value in mapping processes to the functional model has been to gain a clearer understanding of how the mixture of digital and established archival processes can be viewed more coherently within a single functional model. Digital preservation activities at TNA are not so old that they span the same degree of technological changes as the UKDA but the archive itself had many pre-existing aspects which had to be accommodated. Mapping to OAIS provided a means to join it all up again.

For the UKDA, the examination of the total body of the organisational procedures and similar documents during the early phase of the work provided some interesting insights into the effects of technological change on organisational roles and modus operandi. In particular the advent of the desktop computer and the Internet have resulted in significant changes to work patterns, data distribution methods and user expectations and many of the recorded changes to procedures and work flow patterns coincide with technical developments.



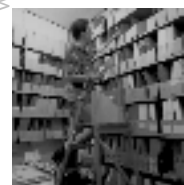
The observed technological change over time has, therefore, had an impact on what we do with the OAIS now. For decades archivists have striven to maintain high standards using information technology that was developed for purposes other than archiving or archives. IT has now developed so far that we are in need of standards and models for archiving the digital materials that we all create.

Hopefully software will soon be available that is built on the basis of the OAIS reference model and which will make the task of archiving and preservation easier for many communities that have no archiving tradition.

The experience of both the UKDA and TNA is that there is value to be gained in mapping to the OAIS reference model and we would encourage others to go ahead and do it themselves, if resources permit. As yet, very few similar exercises have been undertaken, hence we would encourage others to publicise their results because any new example is likely be of use to the wider community

### Recommendations

- Compliant organisations should be encouraged to use the OAIS reference model terminology and newly existing organisations should be encouraged to adopt the terminology from their inception.
- Consideration should be given to the development of an “OAIS Lite” to help smaller archives gain value from the reference model.
- Organisations wishing to undertake a similar exercise to this need to be aware that it is resource intensive and requires the full support of senior management to enable it to be completed successfully.
- Consideration should be given to the extent that pre-ingest and ingest activities overlap and how this is reflected in the model.
- Consideration should be given to the relationship between AIPs and the created DIPs within the model.
- Consideration should be given to whether there may be benefits to the community of providing a guide or manual, or a self-testing toolkit for mapping an archive’s functions to the OAIS reference model.
- It is recommended that the OAIS model should include some reference to the importance of, and need for, archives to operate a system of ‘standards watch’ as well as the technology watch which is already mentioned.



## BIBLIOGRAPHY

---

- CCSDS 651.0-B-1. (2003) *Producer-Archive Interface Methodology Abstract Standard*. Blue Book. <http://www.ccsds.org/CCSDS/documents/651x0b1.pdf>
- CCSDS 650.0-B-1. (2002) *Reference Model for an Open Archival Information System (OAIS)*. Blue Book (Standard). Issue 1 <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>
- Dale, Robin. *Metadata: The Big Picture* [http://www.leeds.ac.uk/cedars/pubconf/umist/Dale\\_RSlides.ppt](http://www.leeds.ac.uk/cedars/pubconf/umist/Dale_RSlides.ppt)
- Day, Michael. (2004) *The OAIS Reference Model: Current Implementations*. Chinese-European Workshop on Digital Preservation, Beijing, China <http://www.ukoln.ac.uk/preservation/presentations/2004/beijing/oais/day-slides.ppt>
- Gartner, Richard. (2002) *METS: Metadata Encoding and Transmission Standard* [http://www.jisc.ac.uk/uploaded\\_documents/tsw\\_02-05.rtf](http://www.jisc.ac.uk/uploaded_documents/tsw_02-05.rtf)
- L'Hours, Herve. (2004) *OHPR Metadata Strategy (unpublished)*
- Lavoie, Brian. (2000) *Meeting the challenges of digital preservation: the OAIS reference model*. OCLC Newsletter, January/February 2000, pp. 26-30 <http://www2.oclc.org/oclc/pdf/news243.pdf>
- Lavoie, Brian. (2004) *The Open Archival Information System Reference Model: Introductory Guide*. DPC Technology Watch Report 04-01 [http://www.dpconline.org/docs/lavoie\\_OAIS.pdf](http://www.dpconline.org/docs/lavoie_OAIS.pdf)
- Lazorchak, William. (2004) *The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation* [http://www.ibiblio.org/squealer/butchhome/papers/ghost\\_in\\_the\\_machine.pdf](http://www.ibiblio.org/squealer/butchhome/papers/ghost_in_the_machine.pdf)
- Library of Congress. (2004) *Metadata and Encoding Transmission Standard: METS and Overview and Tutorial* <http://www.loc.gov/standards/mets/METSOverview.v2.html>
- McDonough, Jerome. (2004) *Introduction to METS (Metadata Encoding and Transmission Standard)*. New York University [http://www.ukoln.ac.uk/events/jisc-cni-2004/presentations/ppt-2000-html/j-mcdonough\\_files/frame.htm](http://www.ukoln.ac.uk/events/jisc-cni-2004/presentations/ppt-2000-html/j-mcdonough_files/frame.htm)
- METS Implementation Registry <http://sunsite.berkeley.edu/mets/registry/>
- Patel, Manjula. (2004) *Scoping Study for a DCC Metadata Standards Registry* [http://www.dcc.ac.uk/docs/Scoping\\_Study\\_for\\_a\\_DCC\\_Metadata\\_Standards\\_Registry.pdf](http://www.dcc.ac.uk/docs/Scoping_Study_for_a_DCC_Metadata_Standards_Registry.pdf)
- Proffitt, Merrilee. (2003) *News from the Digital Library: The Metadata Encoding and Transmission Standard, the Open Archival Information System* [http://dc-mrg.english.ucsb.edu/conference%202003/documents/merrilee\\_proffitt.ppt](http://dc-mrg.english.ucsb.edu/conference%202003/documents/merrilee_proffitt.ppt)
- RLG/OCLC. (2002) *Trusted Digital Repositories: Attributes and Responsibilities* [http://www.rlg.org/en/page.php?Page\\_ID=20477](http://www.rlg.org/en/page.php?Page_ID=20477)
- RLG/OCLC. (2002) *Preservation Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects*. A Report by The OCLC/RLG Working Group on Preservation Metadata [http://www.oclc.org/research/projects/pmwg/pm\\_framework.pdf](http://www.oclc.org/research/projects/pmwg/pm_framework.pdf)
- Waibel, Gunter. (2003) *Like Russian Dolls: Nesting Standards for Digital Preservation*. RLG DigiNews, Volume 7, Number 3 <http://www.rlg.org/legacy/preserv/diginews/diginews7-3.html#feature2>



## APPENDIX 1. GLOSSARY

---

A brief explanation of specific UKDA and TNA terms, concepts as used in the OAIS reference model and terms used in relation to the METS standard.

AACR2: Anglo-American Cataloging Rules, the standard application guidelines for machine-readable cataloguing practice.

ARC: Acquisitions Review Committee – internal UKDA committee.

Archive: an organization that intends to preserve information for access and use by a designated community.

ASCII: American Standard Code for Information Interchange – the most common format for text files on computers and the internet.

Archival storage: the OAIS entity that contains the services and functions used for the storage and retrieval of Archival Information Packages.

Attribute: within an XML document, a component of an element which modifies its meaning: for example: type="PHYSICAL"

CALM: UKDA information package tracking database.

Common services: the supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security and directory services necessary to support the OAIS.

Consumer: the role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail. This can include other OAISs, as well as internal OAIS persons or systems.

Content Data Object: the data object, that together with associated representation information, is the original target of preservation.

DAG: Director's Advisory Group, UKDA internal management committee.

Data Management: the OAIS entity that contains the services and functions for populating, maintaining, and accessing a wide variety of information. Some examples of this information are catalogues and inventories on what AIPs may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data, consumer access statistics, consumer billing, security controls, and OAIS schedules, policies, and procedures.

Data Object: either a physical object or a digital object.

DDI: Data Documentation Initiative – a metadata standard used for documenting datasets.

DTD: Document Type Definition.

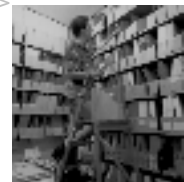
Designated community: an identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.

Digital Archive: the repository for born digital AIPs at TNA.

Digital Object: an object composed of a set of bit sequences.

DPD: Digital Preservation Department at TNA.

DROID: a file format identification tool created by TNA using information stored in PRONOM.



Dublin Core (DC): a basic set of 15 metadata elements designed to represent core fields for the description of any electronic resource.

EAD: Encoded Archival Description, an XML standard for the encoding of archival finding aids.

Element: the tags within an XML document, corresponding to the fields of a database. They have the format <tag>....</tag>, for instance <structMap>....</structMap>

EROL: Electronic Records Online – the presentation system for born digital records at TNA.

e-print: electronic versions of scholarly papers intended to increase access to scholarly research.

Federated Archives: a group of archives that has agreed to provide access to their holdings via one or more common finding aid.

Independently understandable: a characteristic of information that has sufficient documentation to allow the information to be understood and used by the designated community without having to resort to special resources not widely available, including named individuals.

Ingest: the OAIS entity that contains the services and functions that accept Submission Information Packages from Producers, prepare Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting descriptive information become established within the OAIS.

Long term preservation: the act of maintaining information, in a correct and independently understandable form, over the long term.

Management: the role played by those who set overall OAIS policy as one component in a broader policy domain. Management provides the OAIS with its charter and scope.

MARC: MACHine Readable Cataloguing: the standard within traditional libraries for cataloguing information.

METS: Metadata Encoding and Transmission Standard.

MIRAGE: UKDA acquisition tracking database.

MOA2: Making of America II – a DTD which attempted to encode digital library metadata, now superseded by METS.

NARA: National Archives and Records Administration.

NDAD: National Digital Archive of Datasets – a digital archive for statistical datasets run by University of London Computer Centre under contract with TNA.

NISO: National Information Standards Organisation.

OCLC: Online Computer Library Center.

Open-source: software released (usually free-of-charge) with its source code under a licence which allows it to be modified by users.

Producer: the role played by those persons, or client systems, who provide the information to be preserved. This can include other OAISs or internal OAIS persons or systems.

PRONOM: the technical file format registry provided on the Internet by TNA.

Representation information: the information that maps a data object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a data object) is mapped into a symbol.



RLG: Research Libraries Group.

RMD: Records Management Department at TNA.

SMT: Senior Management Team, UKDA.

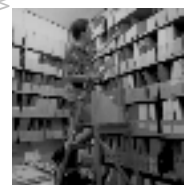
TEI: Text Encoding Initiative – a widely used XML application for electronic text.

URL: Universal Resource Locator – the “address” of an Internet resource (for example, <http://www.loc.gov/>).

VRML: Virtual Reality Markup Language.

XML: eXtensible Markup Language.

XML parser: a software package used to validate an XML-encoded document (i.e., checks that it conforms to the DTD or schema in which it is encoded).



## APPENDIX 2. UKDA PREFERRED FILE FORMATS

| TYPE OF DATA                                      | PRINCIPAL INGEST FORMATS                                                                       |
|---------------------------------------------------|------------------------------------------------------------------------------------------------|
| Alphanumeric data (held in a statistical package) | Delimited text, SPSS portable (.por), SPSS system (.sav), STATA (.dta), SAS (transport file)   |
| Alphanumeric data (held in a database)            | Delimited text, MS Access, dBase, FoxPro, SIR (export file), XML, Filemaker Pro, Paradox       |
| Alphanumeric data (held in a spreadsheet)         | Delimited text, MS Excel, Lotus, Quattro Pro                                                   |
| Textual data (e.g. transcripts)                   | MS Word, Rich Text Format (RTF), XML, SGML, HTML, WordPerfect, Plain (unformatted) Text, paper |
| Video data                                        | MPEG-1, M-PEG-2, MPEG-4, Apple Quicktime                                                       |
| Audio data                                        | MS Wave, MPEG-1 Audio Layer 3 (MP3)                                                            |
| Raster (bitmap) images                            | TIFF, PNG, GIF, BMP                                                                            |
| Vector images                                     | DXF, SVG, Adobe Illustrator                                                                    |
| Documentation/metadata                            | MS Word, RTF, XML, SGML, HTML, WordPerfect, Adobe PDF, paper                                   |

| TYPE OF DATA                                              | PRESERVATION FORMATS                                                                       |
|-----------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Alphanumeric data                                         |                                                                                            |
| (held in a statistical package, database, or spreadsheet) | Delimited ASCII with data definition statements, Tagged ASCII or Unicode                   |
| Textual data                                              |                                                                                            |
| (e.g. transcripts)                                        | Rich Text Format (RTF), SGML, XML, Plain (unformatted) ASCII or Unicode                    |
| Video data                                                | MPEG-2                                                                                     |
| Audio data                                                | Currently holding MS Wave, MPEG-1 Audio Layer 3 (MP3) (preservation standard under review) |
| Raster (bitmap) images                                    | TIFF                                                                                       |
| Vector images                                             | DXF, SVG                                                                                   |
| Documentation/metadata                                    | RTF, TIFF, XML, SGML, Plain (unformatted) ASCII                                            |

### APPENDIX 3. TNA PRESERVATION METADATA MAPPING TO OAIS INFORMATION MODEL

| NO.  | ELEMENT                                                         | DESCRIPTION                                                                                                                                                                                 | LEVEL (DA)               | LEVEL (PROCAT) | SEARCHABLE/<br>EDITABLE/<br>REPEATABLE | MANDATORY                       | TYPE             | OAIS MAPPING |
|------|-----------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|----------------|----------------------------------------|---------------------------------|------------------|--------------|
| A    | Record                                                          | This metadata applies to the abstract record within the archive                                                                                                                             |                          |                |                                        |                                 |                  |              |
| A0.1 | Record Level<br><Aggregation>                                   | The level of record that the metadata is describing. This is derived from PRO cataloguing standards or standards such as ISAD(G), 2 <sup>nd</sup> edition.                                  | Accession,<br>Record (P) | Series - Item  | E                                      | Mandatory                       | Value list       | Reference    |
| A1   | Record Reference                                                | This metadata describes how the record is referenced                                                                                                                                        |                          |                |                                        |                                 |                  |              |
| A1.1 | System Identifier<br><SystemID>                                 | The system generated unique identifier for each accession or record                                                                                                                         | Accession,<br>Record     | N/A            | S                                      | Mandatory                       | System generated | Reference    |
| A1.2 | PROCAT Reference<br><Identifier>                                | A code that is specific to an individual records or group of records to distinguish it from other records and display the relationship between certain records.                             | Accession,<br>Record (P) | Series - Item  | S/E                                    | Mandatory                       | Text             | Reference    |
| A1.3 | Former Reference<br>– External<br><FormerExternalRef<br>erence> | A reference code previously used for an individual record or group of records by an agency outside the PRO.                                                                                 | Accession,<br>Record (P) | Series – Item  | S/E                                    | Optional                        | Text             | Reference    |
| A1.4 | Former Reference<br>– PRO<br><FormerPRO<br>reference>           | A PRO reference previously used for an individual record or group of records.                                                                                                               | Accession,<br>Record (P) | Series - Item  | S/E                                    | Optional                        | Text             | Reference    |
| A1.5 | Publication Note<br><PublicationNote>                           | A bibliographical reference to published finding aids, editions, transcripts and reproductions of records, and significant publications based on the use, study or analysis of the records. | Accession                | Series – Item  | E                                      | Optional                        | Text             | Reference    |
| A1.6 | Unpublished Finding<br>Aids<br><Unpublished<br>FindingAid>      | A reference to current unpublished finding aids                                                                                                                                             | Accession                | Series – Item  | E                                      | Optional                        | Text             | Reference    |
| A2   | Record Content                                                  | This metadata relates to the content of the record                                                                                                                                          |                          |                |                                        |                                 |                  |              |
| A2.1 | Title<br><Title>                                                | The name given to the unit of description.                                                                                                                                                  | Accession,<br>Record (P) | Series - Item  | S/E                                    | Mandatory at<br>Accession level | Text             | Descriptive  |

|       |                                                                  |                                                                                                                                                                                             |                   |                         |     |  |                              |            |                |
|-------|------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|-------------------------|-----|--|------------------------------|------------|----------------|
| A2.2  | Significant Properties<br><Significant Properties>               | A description of the significant properties of the record which must be preserved.                                                                                                          | Accession, Record | N/A                     | E   |  | Optional                     | Text       | Descriptive    |
| A2.3  | Scope and Content<br><Description>                               | Synopsis of form and subject matter within the records.                                                                                                                                     | Accession, Record | Series – Item           | S/E |  | Mandatory                    | Text       | Descriptive    |
| A2.4  | Creator<br><Creator>                                             | Organisation or individual responsible for content of record up to its declaration as a record                                                                                              | Accession, Record | Series – Item           | S/E |  | Mandatory at Accession level |            | Provenance     |
| A2.5  | Covering dates<br><CoveringDates>                                | Dates that span the creation of the records.                                                                                                                                                | Accession         | Series – Item           | E   |  | Mandatory                    | Text       | Provenance     |
| A2.6  | Accumulation dates<br><Accumulation Dates>                       | Dates that span the accumulation of the records. These dates refer to the record keeping activities of the creator(s), e.g. series began in May 1925 or series ran from 1920 to 1960.       | Accession         | Series – Sub sub-series | E   |  | Optional                     | Text       | Provenance     |
| A2.7  | Arrangement<br><Arrangement>                                     | The logical structure of the record, e.g. file structure or data model                                                                                                                      | Accession, Record | Series - Item           | E   |  | Optional                     | Text       | Representation |
| A2.8  | Record declaration date<br><DateDeclared>                        | Date (and time) a document of business activity was declared to be a formal record                                                                                                          | Record            | N/A                     | S/E |  | Optional                     | Date       | Provenance     |
| A2.9  | Legal Status<br><LegalStatus>                                    | A note as to whether or not the records are Public Records.                                                                                                                                 | Accession         | Series – Item           | E   |  | Mandatory                    | Value List | Context        |
| A2.10 | Language<br><Language>                                           | The native language that the information on the records is captured in. Comply with ISO 639                                                                                                 | Record            | Piece, Item             | E   |  | Mandatory                    | Value list | Representation |
| A2.11 | Mandate<br><Mandate>                                             | Purpose for which information contained within the record was processed                                                                                                                     | Accession         | N/A                     | E   |  | Optional                     | Text       | Context        |
| A2.12 | Place of Deposit<br><Location>                                   | The authority controlled name of the repository holding the record.                                                                                                                         | Accession         | Series – Item           | S/E |  | Mandatory                    | Value list | Provenance     |
| A3    | Record Provenance                                                | This metadata relates to the provenance of the record                                                                                                                                       |                   |                         |     |  |                              |            |                |
| A3.1  | Administrative/ Biographical History<br><Administrative History> | Administrative, biographical details, or other historical statement about the corporate body/bodies or individual(s) responsible for creating and accumulating the records being described. | Accession         | Series – Item           | E   |  | Optional                     | Text       | Provenance     |



|        |                                                                     |                                                                                                                                                                                                   |                   |   |                            |                             |
|--------|---------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|---|----------------------------|-----------------------------|
| A3.2   | Custodial history<br><CustodialHistory>                             | Information on where and how records have been held from creation to transfer to the PRO, giving details of changes of ownership and/or custody that may be significant to the records integrity. |                   | R |                            | Provenance                  |
| A3.2.1 | Custody Process<br><CustodyProcess>                                 | The process which occurred, e.g. change of ownership                                                                                                                                              | Accession         | E | Series - Piece<br>Optional | Text<br>Provenance          |
| A3.2.2 | Custody Process Date<br><CustodyProcessDate>                        | The date on which the process occurred                                                                                                                                                            | Accession         | E | Series – Piece<br>Optional | Date<br>Provenance          |
| A3.2.3 | Custody Process Source Agency<br><CustodyProcessSourceAgency>       | The name of the source agency for the process                                                                                                                                                     | Accession         | E | Series - Piece<br>Optional | Text<br>Provenance          |
| A3.2.4 | Custody Process Receiving Agency<br><CustodyProcessReceivingAgency> | The name of the receiving agency for the custody process                                                                                                                                          | Accession         | E | Series – Piece<br>Optional | Text<br>Provenance          |
| A3.2.5 | Custody Process Notes<br><CustodyProcessNotes>                      | Any additional notes on the custody process                                                                                                                                                       | Accession         | E | Series - Piece<br>Optional | Text<br>Provenance          |
| A4     | Record Context                                                      | This metadata describes the relationship between this record and others in the archive or held externally                                                                                         |                   |   |                            |                             |
| A4.1   | Related records<br><Relation>                                       | Reference to any material directly related to the records being described                                                                                                                         |                   | R |                            | Context                     |
| A4.1.1 | Relationship Type<br><RelationshipQualifier>                        | Type of relationship between records                                                                                                                                                              | Accession, Record | E | Series - Item<br>Optional  | Text<br>Context             |
| A4.1.2 | Related Records ID<br><RelatedRecordID>                             | ID of related records                                                                                                                                                                             | Accession, Record | E | Series – Item<br>Optional  | As for System ID<br>Context |
| A4.2   | Separated records<br><SeparatedRecords>                             | Reference to any material which was originally part of the archival record but which has been removed, separated or destroyed.                                                                    | Accession, Record | E | Series - Item<br>Optional  | Text<br>Context             |
| A4.3   | Location of Originals<br><LocationOfOriginals>                      | Location of original records if PRO holds copies.                                                                                                                                                 | Accession         | E | Series – Item<br>Optional  | Text<br>Context             |
| A4.4   | Copies<br><LocationOfCopies>                                        | Location of copies or renditions of the records                                                                                                                                                   | Accession         | E | Series - Item<br>Optional  | Text<br>Context             |



|      |                                                                   |                                                                                                                                                                                                                                    |           |                     |     |           |            |                            |
|------|-------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|---------------------|-----|-----------|------------|----------------------------|
| A5.6 | Password<br><Password>                                            | Password to access record if system does not permit password removal                                                                                                                                                               | Record    | N/A                 | E   | Optional  | Text       | Representation / Context ? |
| A5.7 | Encryption<br><EncryptionMethod>                                  | Method used if any (would be removed prior to loading).                                                                                                                                                                            | Record    | N/A                 | E   | Optional  | Text       | Representation             |
| A6   | Access                                                            |                                                                                                                                                                                                                                    |           |                     |     |           |            |                            |
| A6.1 | Access Conditions<br><AccessConditions>                           | Conditions governing access to the records, e.g. subject to 30 year closure.                                                                                                                                                       | Accession | Series – Sub-series | E   | Mandatory | Value List |                            |
| A6.2 | Security category<br><SecurityCategory>                           | The protective marking on the record, e.g. "Restricted".                                                                                                                                                                           | Record    | N/A                 | E   | Optional  | Value List |                            |
| A6.3 | Closure type<br><ClosureQualifier>                                | Conditions governing access to the records at piece or item level.                                                                                                                                                                 | Record    | Piece, Item         | E   | Mandatory | Value List |                            |
| A6.4 | Closure code<br><ClosureCode>                                     | Code associated with the closure type.                                                                                                                                                                                             | Record    | Piece, Item         | E   | Mandatory | Value List |                            |
| A6.5 | Closure status<br><ClosureStatus>                                 | Current closure status of the record and description.                                                                                                                                                                              | Record    | Piece, Item         | S/E | Mandatory | Value List |                            |
| A6.6 | Restrictions on Use<br><RestrictionsOnUse>                        | Restrictions on the use or reproduction of the material after access has been granted, e.g. copyright restrictions. This can also include information on physical restrictions, such as "only available on Computer Lab terminal". | Record    | Series - Item       | E   | Optional  | Text       |                            |
| A6.7 | Record Opening Date<br><DateIssued>                               | Date when the record was/will be opened.                                                                                                                                                                                           | Record    | Piece, Item         | S/E | Mandatory | Date       |                            |
| A6.8 | Record Release Date<br><DateAvailable>                            | Date when the record was/will be made accessible to the public                                                                                                                                                                     | Record    | N/A                 | S/E | Mandatory | Date       |                            |
| A7   | Accession                                                         |                                                                                                                                                                                                                                    |           |                     |     |           |            |                            |
| A7.1 | Accession Date<br><DateAcquired>                                  | The date the records became part of the Digital Archive.                                                                                                                                                                           | Accession | N/A                 | S   | Mandatory | Date       | Provenance                 |
| A7.2 | Accruals<br><Accruals>                                            | Indicates if additions of further records to an accession are expected.                                                                                                                                                            | Accession | Series              | E   | Optional  | Text       | Context                    |
| A7.3 | Immediate Source of Acquisition<br><ImmediateSourceOfAcquisition> | The authority controlled name of the corporate body or person from which the records were acquired.                                                                                                                                | Accession | Series - Piece      | S/E | Mandatory | Value List | Provenance                 |

|        |                                                                   |                                                                                                                                                                                                                        |                          |                         |   |           |            |             |
|--------|-------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|-------------------------|---|-----------|------------|-------------|
| A7.4   | Disposition information                                           | Information on any appraisal, scheduling and destruction actions, e.g. is it a sample, was it transferred under FOI? This should include the section of the acquisition policy under which the record was accessioned. | Accession                | Series – Sub sub-series | E | Mandatory | Text       | Provenance  |
| A7.5   | Accession Procedure<br><Accession<br>Procedure>                   | Describes the events undertaken as part of accession into the archive                                                                                                                                                  |                          |                         | R |           |            | Provenance  |
| A7.5.1 | Accession Procedure<br>Event<br><AccessionProcedur<br>eEvent>     | Describes the event which occurred, e.g. files transferred to DP server                                                                                                                                                | Accession                | N/A                     | E | Mandatory | Text       | Provenance  |
| A7.5.2 | Accession Procedure<br>Process<br><AccessionProcedure<br>Process> | Describes the process involved, e.g. files copied from DIT to server and extracted from Backup Exec archive format                                                                                                     | Accession                | N/A                     | E | Mandatory | Text       | Provenance  |
| A7.5.3 | Accession Procedure<br>Date <accession<br>ProcedureDate>          | Describes the date on which the process took place                                                                                                                                                                     | Accession                | N/A                     | E | Mandatory | Date       | Provenance  |
| A7.5.4 | Accession Procedure<br>Agency<br><AccessionProcedure<br>Agency>   | The agency which undertook the process, e.g. DP Department                                                                                                                                                             | Accession                | N/A                     | E | Mandatory | Text       | Provenance  |
| A7.5.5 | Accession Procedure<br>Outcome<br><AccessionProcedure<br>Outcome> | The outcome of the process. e.g. all files successfully transferred                                                                                                                                                    | Accession                | N/A                     | E | Mandatory | Text       | Provenance  |
| A7.5.6 | Accession Procedure<br>Notes<br><AccessionProcedur<br>eNotes>     | Any additional notes on the process                                                                                                                                                                                    | Accession                | N/A                     | E | Mandatory | Text       | Provenance  |
| A8     | Indexing                                                          |                                                                                                                                                                                                                        |                          |                         |   |           |            |             |
| A8.1   | Corporate Names<br><CorporateName>                                | Authority control data (PROCAT index)                                                                                                                                                                                  | Accession,<br>Record (P) | Series - Item           | S | Optional  | Value List | Descriptive |
| A8.2   | Personal Names<br><PersonalName>                                  | Authority control data (PROCAT index)                                                                                                                                                                                  | Accession,<br>Record (P) | Series - Item           | S | Optional  | Value List | Descriptive |
| A8.3   | Place Names<br><CoverageSpatial>                                  | Authority control data (PROCAT index)                                                                                                                                                                                  | Accession,<br>Record (P) | Series - Item           | S | Optional  | Value List | Descriptive |
| A8.4   | Subjects<br><Subject>                                             | Authority control data (PROCAT index)                                                                                                                                                                                  | Accession,<br>Record (P) | Series - Item           | S | Optional  | Value List | Descriptive |



| B      | Manifestation                                                       | This applies to specific manifestations of an individual record. There will always be at least one manifestation for each record which has files associated, i.e. the original accessioned form. |     |     |  |           |                  |                | Repeatable for each manifestation of a record. |
|--------|---------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|-----|--|-----------|------------------|----------------|------------------------------------------------|
| B1     | Manifestation Object                                                | This describes the actual object being manifested                                                                                                                                                |     |     |  |           |                  |                |                                                |
| B1.1   | Manifestation ID<br><ManifestationID>                               | The unique ID for each manifestation                                                                                                                                                             | N/A | S   |  | Mandatory | System generated | Reference      |                                                |
| B1.2   | Manifestation Type<br><Manifestation<br>Qualifier>                  | The type of manifestation, e.g. HTML version                                                                                                                                                     | N/A | S/E |  | Mandatory | Text             | Descriptive    |                                                |
| B1.3   | Manifestation Structure<br><ManifestationComp<br>onents>            | Description of any complex components of the manifested object and their interrelationships                                                                                                      | N/A | E   |  | Optional  | Text             | Representation |                                                |
| B1.4   | Installation Requirements<br><InstallationRequire<br>ments>         | Any procedures required to install the manifestation for use, e.g. unpack zip file to specified directory                                                                                        | N/A | E   |  | Optional  | Text             | Representation |                                                |
| B1.5   | Manifestation Notes<br><Manifestation<br>Notes>                     | E.g. note of preservation strategy adopted.                                                                                                                                                      | N/A | E   |  | Optional  | Text             | Representation |                                                |
| B1.6   | Migration<br><Migration>                                            | Describes the migration process which may have been used to create the manifestation                                                                                                             |     | R   |  |           |                  | Provenance     |                                                |
| B1.6.1 | Migration iteration number<br><MigrationIteration>                  | The iteration number of the migration, e.g. 3 <sup>rd</sup> migration.                                                                                                                           | N/A | E   |  | Optional  | Number           | Provenance     |                                                |
| B1.6.2 | Migration Date<br><MigrationDate>                                   | The date on which the records were migrated.                                                                                                                                                     | N/A | E   |  | Optional  | Date             | Provenance     |                                                |
| B1.6.3 | Migration Pathway Used<br><MigrationPathway>                        | Reference to the relevant migration pathway documentation.                                                                                                                                       | N/A | S/E |  | Optional  | Text             | Provenance     |                                                |
| B1.6.4 | Migrated By<br><MigratedBy>                                         | Details of the individual or agency responsible for undertaking the migration                                                                                                                    | N/A | E   |  | Optional  | Text             | Provenance     |                                                |
| B1.6.5 | Certification of accurate migration<br><Migration<br>Certification> | Details of any layout, content or functionality lost in migration, discrepancies and corrections                                                                                                 | N/A | E   |  | Optional  | Text             | Provenance     |                                                |

| B1.7     | Manifestation status<br><Manifestation<br>Status>                                    | Eg. Original, previous, current                                                                                          | Manifestation | N/A | S/E | Mandatory | Value List                    | Context                                 |
|----------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|---------------|-----|-----|-----------|-------------------------------|-----------------------------------------|
| B1.8     | Manifestation<br>Documentation<br><ManifestationDocu<br>mentation>                   | Link to any documentation associated with<br>the manifestation                                                           | Manifestation | N/A | E   | Optional  | Text                          | Representation                          |
| B1.9     | Related<br>Manifestations<br><Related<br>Manifestation>                              | Describes related manifestations.                                                                                        |               |     | R   |           |                               | Context                                 |
| B1.9.1   | Manifestation<br>Relationship type<br>ManifestationRelatio<br>nshipQualifier>        | Eg. Migrated from, migrated to, equivalent<br>to.                                                                        | Manifestation | N/A | E   | Optional  | Value List                    | Context                                 |
| B1.9.2   | Manifestation<br>Relationship ID<br><Related<br>ManifestationID>                     | ID of related manifestations                                                                                             | Manifestation | N/A | E   | Optional  | As for<br>Manifestation<br>ID | Context                                 |
| B2       | Manifestation<br>Environment                                                         | This describes the environment required to<br>render the manifested object.                                              |               |     |     |           |                               | Each sub-element<br>below is repeatable |
| B2.1     | Software<br>Environment<br><Software<br>Environment>                                 | Information about the software<br>environment required to render the stored<br>bitstream into a displayable record.      |               |     | R   |           |                               | Representation                          |
| B2.1.1   | Transformation<br>Process<br><Transformation<br>Process>                             | Describes the process required to<br>transform the stored bitstream into a<br>displayable object                         |               |     | R   |           |                               | Representation                          |
| B2.1.1.1 | Transformation<br>Engine<br><Transformation<br>Engine>                               | The software required to transform the<br>object, e.g. Winzip Version 8.1                                                | Manifestation | N/A | E   | Optional  | Text                          | Representation                          |
| B2.1.1.2 | Transformation<br>Parameters<br><TransformationPara<br>meters>                       | Runtime parameters required for the<br>transformation engine, e.g. output<br>directory                                   | Manifestation | N/A | E   | Optional  | Text                          | Representation                          |
| B2.1.1.3 | Transformation<br>Engine Location<br><TransformationEngi<br>neLocation>              | A link to the location of an operational<br>version of the transformation engine, e.g.<br>to ISO in the software library | Manifestation | N/A | E   | Optional  | Text                          | Representation                          |
| B2.1.1.4 | Transformation<br>Engine<br>Documentation<br><TransformationEngi<br>neDocumentation> | A link to documentation for the<br>transformation engine                                                                 | Manifestation | N/A | E   | Optional  | Text                          | Representation                          |

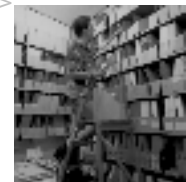


| B2.1.2   | Access Process<br><AccessProcess>                          | Describes the process required to access and display the transformed bitstream                                                     | Manifestation | N/A | R | R | Optional | Text | Representation |
|----------|------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|---------------|-----|---|---|----------|------|----------------|
| B2.1.2.1 | Access Engine<br><AccessEngine>                            | Describes the software required to display the object and access its intellectual content, e.g. Adobe Acrobat Reader Version 5.0.5 | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.1.2.2 | Access Engine Location<br><AccessEngineLocation>           | A link to the location of an operational version of the access engine, e.g. to ISO in the software library                         | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.1.2.3 | Access Engine Documentation<br><AccessEngineDocumentation> | A link to documentation for the access engine                                                                                      | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.2     | Operating System Environment<br><OSEnvironment>            | Information about the OS required to access the records.                                                                           |               |     | R | R |          |      | Representation |
| B2.2.1   | OS Name<br><OSName>                                        | The name of the OS, e.g. Windows XP                                                                                                | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.2.2   | OS Version<br><OSVersion>                                  | The OS edition, version, and any service packs installed, e.g. Home Edition, Version 2002, Service Pack 1 installed                | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.2.3   | OS Location<br><OSLocation>                                | A link to the location of an operational version of the OS, e.g. to ISO in the software library                                    | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.2.4   | OS Documentation<br><OSDocumentation>                      | A link to OS documentation                                                                                                         | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.3     | Hardware Hardware<br><HardwareEnvironment>                 | Information about the hardware environment required to access the records.                                                         |               |     | R | R |          |      | Representation |
| B2.3.1   | CPU Requirements<br><CPUrequirements>                      | The minimum CPU type and speed required, e.g. Athlon XP 1700+ (version 6), 1.44 GHz                                                | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.3.2   | RAM Requirements<br><RAMrequirements>                      | The minimum RAM required, e.g. 256MB                                                                                               | Manifestation | N/A | E | E | Optional | Text | Representation |
| B2.3.3   | Storage Requirements<br><StorageRequirements>              | The minimum hard disk space required, e.g. 10GB free disk space                                                                    | Manifestation | N/A | E | E | Optional | Text | Representation |

|        |                                                            |                                                                                                                                  |               |     |     |           |                  |                                                  |
|--------|------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|---------------|-----|-----|-----------|------------------|--------------------------------------------------|
| B2.3.4 | Peripheral Requirements<br><Peripheral Requirements>       | Additional peripheral requirements, including sound and graphics cards, dongles, drives, e.g. soundblaster-compatible sound card | Manifestation | N/A | E   | Optional  | Text             | Representation                                   |
| B2.3.5 | Hardware Location<br><HardwareLocation>                    | A description of the location of the necessary hardware environment, e.g. Computer Laboratory                                    | Manifestation | N/A | E   | Optional  | Text             | Representation                                   |
| B2.3.6 | Hardware Documentation<br><Hardware Documentation>         | A link to any hardware documentation                                                                                             | Manifestation | N/A | E   | Optional  | Text             | Representation                                   |
| C      | File                                                       | This describes the individual files which comprise a specific manifestation.                                                     |               |     |     |           |                  | Repeatable for each file within a manifestation. |
| C1     | File Information                                           |                                                                                                                                  |               |     |     |           |                  |                                                  |
| C1.1   | System File ID<br><SystemFileID>                           | Unique system ID assigned to each file.                                                                                          | File          | N/A | S   | Mandatory | System generated | Reference                                        |
| C1.2   | System File Name<br><SystemFileName>                       | The system file name, assigned by the digital archive.                                                                           | File          | N/A | S   | Mandatory | System generated | Reference                                        |
| C1.3   | Original File Name<br><OriginalFileName>                   | The original file name.                                                                                                          | File          | N/A | S   | Mandatory | Text             | Reference                                        |
| C1.4   | File Format<br><FileFormat>                                | The name and version of the file format.                                                                                         | File          | N/A | S/E | Mandatory | Text             | Representation                                   |
| C1.4.1 | Compression<br><Compression>                               | Details of any form of compression used.                                                                                         | File          | N/A | E   | Optional  | Text             | Representation                                   |
| C1.4.2 | Character Coding Scheme<br><EncodingScheme>                | Details of any character set and encoding scheme used (e.g. Unicode, UTF-8).                                                     | File          | N/A | E   | Optional  | Text             | Representation                                   |
| C1.5   | Fixity<br><Fixity>                                         | Describes the current fixity method used by the system to ensure the authenticity of the file                                    |               |     |     |           |                  | Fixity                                           |
| C1.5.1 | Fixity type<br><FixityTechnique>                           | The technique used to generate the fixity value, e.g. MD5 digest algorithm                                                       | File          | N/A |     | Mandatory | Text             | Fixity                                           |
| C1.5.2 | Fixity Method<br><FixityMethod>                            | The method used to generate the fixity value, e.g. MD5Summer version 1.1.0.22                                                    | File          | N/A |     | Mandatory | Text             | Fixity                                           |
| C1.5.3 | Fixity Method Documentation<br><FixityMethodDocumentation> | A link to the documentation for the fixity method.                                                                               | File          | N/A |     | Mandatory | Text             | Fixity                                           |



|        |                                          |                                                                                         |      |  |     |  |           |         |                                                                                 |
|--------|------------------------------------------|-----------------------------------------------------------------------------------------|------|--|-----|--|-----------|---------|---------------------------------------------------------------------------------|
| C1.5.4 | Fixity Date<br><FixityDate>              | The date on which the fixity value was generated                                        | File |  | N/A |  | Mandatory | Date    | Fixity                                                                          |
| C1.5.5 | Fixity Value<br><FixityValue>            | The value generated by the fixity method.                                               | File |  | N/A |  | Mandatory | Text    | Fixity                                                                          |
| C1.6   | File size<br><FileSize>                  | The file size in bytes.                                                                 | File |  | N/A |  | Mandatory | Numeric | Fixity                                                                          |
| C1.7   | File creation date<br><FileCreationDate> | The system file creation date.                                                          | File |  | N/A |  | Mandatory | Date    | Provenance                                                                      |
| D      | Type Specific Metadata                   | This comprises additional metadata which may be required to support specific file types |      |  |     |  |           |         | This section is still under development, and will be required to be extensible. |
| D1     | Web                                      |                                                                                         |      |  |     |  |           |         |                                                                                 |
|        | Web site/page name                       |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Content owner                            |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Databases used                           |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Plug ins used                            |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Browser version required                 |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Other proprietary software               |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Third party web sites details            | Original URL, date URL link was broken                                                  |      |  |     |  | Optional  |         | (Not used)                                                                      |
| D2     | Email                                    |                                                                                         |      |  |     |  |           |         |                                                                                 |
|        | Recipient name                           |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Sender name                              |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Copy identification                      | Author or recipient version                                                             |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Date/time record sent                    |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Date/time record opened                  |                                                                                         |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Action Requested                         | As a result of record                                                                   |      |  |     |  | Optional  |         | (Not used)                                                                      |
|        | Recipient specific configuration         | Could the recipient see all the information                                             |      |  |     |  | Optional  |         | (Not used)                                                                      |



## APPENDIX 4. EXTRACTS FROM TYPICAL UKDA READ AND NOTE FILES

***The following is an extract from a Note File which is held internally and records information about the history of the files/dataset. It relates to a specific quarter of the Quarterly Labour Force Survey.***

### UK DATA ARCHIVE DOCUMENTATION

SN:4965 - Quarterly Labour Force Survey, December 2003 - February 2004

#### Details of study received

-----

Received from ....., Operations and Research Support, D2/24,  
Office for National Statistics, 1 Drummond Gate, London SW1V 2QQ,  
during May 2004.

email: .....@ons.gov.uk

tel: 020-7533-5376

2nd edition received November 2004 from above depositor.

#### Details of data files received

-----

d03fSPRP.EXP      SPSS (exported from SIR)      Data file

2nd edn:

d03fSPRP.EXP      SPSS (exported from SIR)      Data file

#### Details of doc files received

-----

Uses generic

#### Details of hard copy doc received

-----

None received

#### File name changes

-----

#### FILES FOR ISSUE:

d03fSPRP.EXP (saved as SPSS POR format): name changed to qlfsd03f.por  
as per UKDA policy.

2nd edn - replacement file with same name and format

#### Level of processing

-----

The study was processed to A\* standard.



#### Data file conversion: methods, formats created and validation

---

The data file was converted into SPSS portable file qlfsd03f.por.  
Stata and tab versions created using script as per normal procedure.

#### Doc file conversion: methods, formats created and validation

---

n/a - uses generic.

#### Dataset validation checks and solutions to problems raised

---

All files created checked added as per A\* standard.  
No problems encountered.

#### Data and documentation problems

---

None encountered. Missing variable and value labels have been added as necessary - syntax file qlfsd03f.sps is stored under dp/code/spss/.

#### Useful Notes

---

ONS have advised that until further notice, the variable GB should not be used for analysis, as it contains anomalies. Users are advised to use other geographies, e.g. COUNTRY & URESMC instead.  
NUTS2 geography development is still continuing. Northern Ireland cases have a default value of -9 for this variable.

Please note that the weighting variables in the QLFS are now as follows:

PWT03 - Person weight

PIWT03 - Person income weight

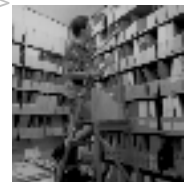
***The following is an extract from a Read File which is created during processing. It is generic in that it relates to all of the Quarterly Labour force surveys. It is available from the catalogue without the need for registration.***

GN:33246 Quarterly Labour Force Survey

#### QUARTERLY LABOUR FORCE SURVEY - reweighting project

---

ONS have recently published reweighted QLFS estimates for 1992-2003, based on survey microdata that have been weighted to post-2001 Census population estimates for the first time. This means that new editions of all quarters from 1992 to 2003 have recently been deposited (April/May 2004) at the UK Data Archive. The names of some variables have changed as a result of the update - details are listed in the individual READ file for each affected study. User should note that the QLFS documentation may still refer to variables by their previous names until the next documentation update.



Further information on the reweighting project may be found in the April 2004 edition of 'Labour Market Trends' periodical. The relevant article is also available online from the ONS website at:  
<http://www.statistics.gov.uk/cci/article.asp?id=887>

#### Identifying Households in the LFS Data

Households in the survey can be uniquely identified by using REMSERNO (which is a concatenation of WEEK, W1YR, QRTR, ADD WAVFND and HHLD) plus either QUOTA or CASEID, as these are identical. These numbers will remain the same for a household while it is in the survey. If REMSERNO does not appear in a dataset it can be calculated using the following equation:

$$\begin{aligned} &\text{week} \times 10000000 + \\ &\text{w1yr} \times 1000000 + \\ &\text{qrtr} \times 100000 + \\ &\text{add} \times 1000 + \\ &\text{wavfnd} \times 100 + \\ &\text{hhld} \end{aligned}$$

Persons within households can be identified by using RECNO.

#### Individual Identifier - CASENO

A new variable CASENO has been included in the dataset from the March - May 2002 quarter (SN:4547) onwards. This variable gives every case in the dataset a unique identifier.

CASENO is calculated by the following equation:

$$\begin{aligned} &\text{quota} \times 100000000000 + \\ &\text{week} \times 1000000000 + \\ &\text{w1yr} \times 100000000 + \\ &\text{qrtr} \times 10000000 + \\ &\text{add} \times 100000 + \\ &\text{wavfnd} \times 10000 + \\ &\text{hhld} \times 100 + \\ &\text{persno} \end{aligned}$$

#### Weights

When using income variables on LFS databases it is necessary to use the income weights. Please refer to Volume 3 of the documentation, 'Details of LFS Variables', for further information on these and the other weighting variables. Following the reweighting exercise of 2004 (see above), the names of the weighting variables have been changed as follows:  
PWT03 - Person weight (replaces INTWT02, NEWWT02, NEWWT and WEIGHT1)



PIWT03 - Person income weight (replaces INTIW02, NEWIW02, NEWIWT and WEIGHT2).

Please note that the documentation may still refer to variables by their previous names until the next documentation update.

#### Zero weights

Users should be aware then when running tables at person level, the weighting variable is sometimes missing or has a value of zero. These cases can be ignored, because they are non-responders about whom there is no information other than knowing that they exist. This situation arises if a member of a household was not present at the time of interview, and the respondent was unwilling, or unable, to give any information about that person.

#### Publication Thresholds

For LFS analysis not involving earnings data, ONS recommend that estimates of less than 10,000 people are not used as they are likely to be unreliable.

Estimates which do involve earnings data are subject to different thresholds because the sample size is smaller, and the grossing procedure is different. Much of the output from analysis of earnings data is in the form of 'means' or 'proportions', so users need to be aware of the size of the population on which such estimates are based.

The threshold when using earnings data is set at 60,000 people for estimates based on wave 5 only (for quarters prior to Spring 1997), and at 30,000 for estimates based on combined waves 1 to 5 data (for Spring 1997 onwards).

Users are advised to consult Volume 1 of the documentation, 'Background and Methodology', for further information on this subject.



## APPENDIX 5. A SET OF QUESTIONS FOR OAIS COMPLIANCE SELF-TESTING

This appendix provides a set of questions that the UKDA and TNA found useful when testing for OAIS compliance. This question set can be extended and tailored for use as a self-assessment tool for OAIS compliance testing by other institutions. Answers to these questions will help when comparing the functions and workflows in an archive with those recommended by the OAIS standard.

### ARCHIVE'S FUNCTIONS

- What is the main responsibility of the archive, what does it currently do? What is the domain of your archive?
- What kind of material does your archive collect and take responsibility for?
- How is the scope of your collection(s) specified or limited? (e.g., in official archive policies, mandates)?
- Who are the main donors of information to your archive, who are the information producers?
- Who are the main users of the information in your archive?
- How would you describe the interaction of the main OAIS 'actors' (Producer, Consumer, Management, Archive) in the context of your archive?
- What standards and requirements must your archive comply with, apart from the OAIS reference model?

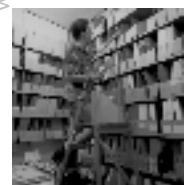
### THE OAIS MANDATORY RESPONSIBILITIES

- How does your archive define or limit the pool/group of its depositors (information producers)?
- Does your archive set quality criteria for the material it acquires (e.g., file formats, description standards, IPR clearance, etc.)?
- Does your archive have a depositor agreement(s) that is/are signed with every acquisition?
- How does the depositor agreement treat the IPR, DP and FoI issues?
- How extensive are the rights given to the archive over the deposited material under the agreement? Is the archive free to change the original submitted material as it sees fit during the preservation processing; or are there limitations to what the archive is allowed to do with the SIP, for example are file format changes to the SIP limited by authenticity requirements resulting in the need to prepare DIPS in different formats?
- What policy and/or strategy does your archive follow to ensure the understandability and usability of the preserved information over time? Is there a formal definition of the chosen preservation strategy of the archive?
- Has your archive defined its designated user community and how? Is the definition fixed by an external body? Is the definition of the designated community likely to change in the (near) future?
- Is access to the materials in your archive limited by some (legal or other) regulation? If so, how does the archive control the adherence to these limits?



## THE OAIS FUNCTIONAL ENTITIES

- What are the ingest actions/functions in your archive?
- Describe a typical submission/delivery session in your archive.
- Who or what department is responsible for:
  - negotiation for a submission agreement;
  - organising the transfer of files and validating the transfer results;
  - analysing and validating the transferred SIPs;
  - providing feedback to the Producer;
  - generating the AIP;
  - generating the necessary additional metadata for the AIP?
- What validation is performed on the submitted information (SIP) in your archive?
- What security measures are taken to ensure the safety of the submission and the submitted data before their storage in the archive?
- Describe a typical SIP and AIP in your archive.
- What transformation processes are performed prior to storage of AIPs in your archive?
- Is there a formal definition of the preservation strategy of the archive?
- Does your archive have a preservation policy?
- How does the policy define digital preservation measures?
- If migration is used for preservation, how is migration defined and used?
- Does the preservation policy set standards for information producers and consumers?
- How is the AIP being stored and archived in your institution?
- How many copies of one AIP is made? Does this policy include every AIP in the archive?
- Is the choice of storage media based on the anticipated frequency of use in your archive?
- Are the storage media checked for errors and how frequently?
- Does the preservation policy include a technology usage plan?
- Does the preservation policy include a disaster plan?
- Is there a disaster recovery policy in place at your institution? Does it cover recovery actions for faulty or lost storage media?
- How does the preservation system in your archive handle the user access requests? (e.g., an automatic online query system, human intervention, no access to AIP storage)?
- How are updates to holdings handled and managed in your archive?
- Is there a report-generation system for archive collections information in your archive? How does it work, what is being reported on?
- Is the administration in your archive involved in managing the preservation system configuration? Who initiates changes in system configuration?
- Who initiates and develops migration plans in your archive? Who is checking the quality of migration decisions and plans?

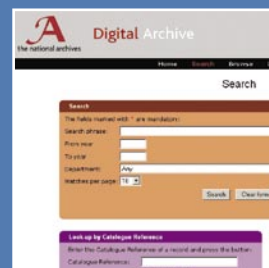
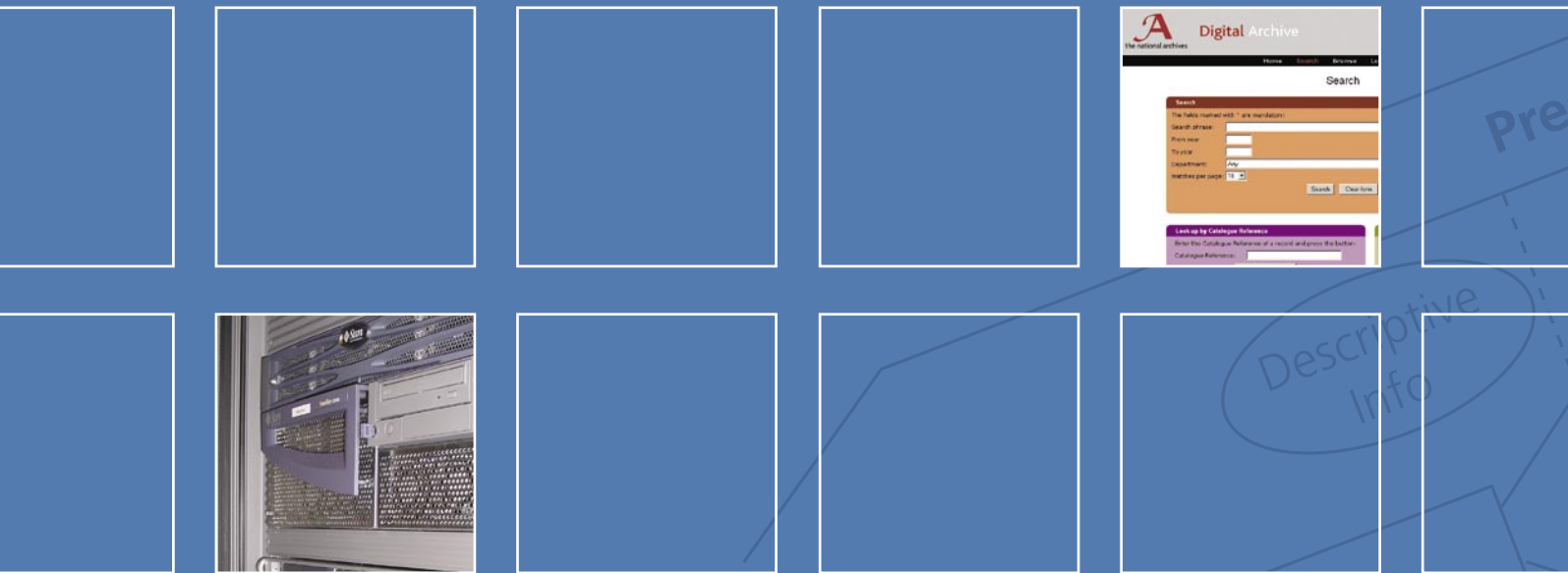


- Who is responsible for monitoring the technology development or providing the Technology Watch in your archive? How are its results reported?
- Does your archive have a policy for developing its own standards and strategies?
- Which unit or department is responsible for developing the strategies and standards for preservation in your archive? Are the current strategies and standards documented in a policy or a written preservation strategy?
- How is physical access to your archive collections controlled?
- How are the requests from users handled in your archive? Are any administrative decisions made before the data are released to the users?
- Does your archive provide information in exchange for money? If so, how is the billing for services to users managed?
- Who is responsible for interacting with the data Producers and Consumers in your archive?
- What methods can users of the archive use to request material from your archive? What finding aids are provided?
- Does your archive have an access/dissemination policy?
- Does the policy refer to limitations on access to information in the archive? If yes, what are the limitations?
- Does the archive have a user agreement or licensing system in place?
- What are the main methods of access to your collections? What media do you use? (e.g., FTP, offline media by post, online, reading room access only)?
- How is the DIP created in your archive? Are any value added services provided to users?
- On what conditions is a new DIP created? Is DIP creation related to resource planning by the administration of the archive?
- Who is responsible for managing the catalogues in your archive?
- What security measures are used during access?

## THE OAIS INFORMATION MODEL

- How does your archive understand the SIP, AIP and DIP packages in the context of their own work?
- What metadata are received from information producers as part of SIPs?
- What metadata, (if any), are being created by the archive for inclusion in the AIP and DIP?
- How does your archive comply with the general OAIS reference model logic (report on discrepancies, mismatches and problems of understanding the OAIS)?





PRODUCER

SIP

Ingest

Descriptive Info

Pre

