DESIDOC Journal of Library & Information Technology, Vol. 32, No. 4, July 2012, pp. 327-334

Implementing Open Archival Information System Model for Digital Preservation at Indian Institute of Geomagnetism

Smita Chandra* and Pratibha A.Gokhale**

**Indian Institute of Geomagnetism, Kalamboli Panvel Highway, New Panvel, Navi Mumbai-410 218*
E-mail: smitac@iigs.iigm.res.in

***DLIS, Jawaharlal Nehru Library, University of Mumbai, Vidyanagari, Santacruz(E), Mumbai-400 098*
E-mail: pratibha_gokhale@yahoo.com

ABSTRACT

The Open Archival Information System (OAIS) reference model ensures that an organisation charged with the task of preserving information (here, digital information) makes it accessible to the designated community in the long-term. It elucidates the functions and processes so as to promote interoperability amongst digital libraries, archives and other institutions and provides a framework for describing and comparing different long-term digital preservation strategies and techniques. This paper discusses what it means to be OAIS compliant for the pilot institutional repository at the Indian Institute of Geomagnetism (IIG). It maps the six entities of an OAIS-compliant repository (Ingest, archival storage, administration, data management, preservation planning and access) onto a pilot institutional repository (IR) structure of the IIG. It discusses the policies and procedures required for effective OAIS implementation within the IIG institutional repository and ways to effectively develop these based on experiences from abroad. The paper further assesses how the institute's operational structure can be informed by the OAIS and vice versa. It establishes the role of the preservation service provider for the IIG IR, and explores the potential for interaction between IIG and such service providers.It identifies the rights and responsibilities, services and actions, and apportions these between the IR and the third-party preservation repository service. It identifies the tools and develops the processes to implement the preservation services and actions. It is expected that this study can be of use to other organisations that might want to undertake implementation of the OAIS reference model.

Keywords: Open archival information system, digital preservation, institutional repository

1. INTRODUCTION

The Open Archival Information System (OAIS) reference model ensures that an organisation charged with the task of preserving information (here, digital information) makes it accessible to the designated community in the long term. It elucidates the functions and processes so as to promote interoperability amongst digital libraries, archives and other institutions and provides a framework for describing and comparing different long-term digital preservation strategies and techniquesby including terminology and concepts, for describing and comparing architectures and operations of existing and future archives[1].

The reference model addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, preservation planning, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It identifies both internal and external interfaces to the archive functions, and identifies a number of high-level services at these interfaces. It provides various illustrative examples and some 'best practice' recommendations. It defines a minimal set of responsibilities for an archive to be called an OAIS, and it also defines a maximal archive to provide a broad set of useful terms and concepts.

With this brief view of the OAIS reference model, this article aims to theoretically verify the argument that the pilot institutional repository (IR) at IIG complies with the OAIS reference model, by mapping the mandatory responsibilities and the OAIS functional entities onto the

Received 1 June 2012, online published 18 July 2012

services and architecture of the pilot repository. It assesses how the institute's operational structure can be informed by the OAIS and vice versa, and establishes the role of the preservation service provider for the IIG IR, and explores the potential for interaction between IIG and such service providers.

2. OAIS COMPLIANCE

The term 'OAIS-compliant' is a common term used in reference to a digital archiving system. The OAIS standard states that an OAIS-compliant digital archive implementation supports the OAIS information model and/or the OAIS functional model (OAIS Chapter 2.2)[1]. It is also committed to fulfilling the responsibilities listed in chapter 3.1 of the reference model, as also described in chapter 2, section on 'OAIS mandatory responsibilities'. The reference model finally notes that the 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. As noted by Lavoie[2], 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. The CCSDS reference model for an OAIS system provides a theoretical framework for developing and maintaining a strategy to ensure long-term preservation of objects within a digital repository. As such the OAIS reference model gives archives a conceptual flexibility. Such a framework therefore often results in ambiguity because the reference model is a conceptual framework rather than a concrete implementation and thus proves to be difficult over how the digital preservation community can endorse the certification of a digital repository as being 'OAIS compliant'. Conformance to the reference model can imply an explicit application of the OAIS concepts, terminology, and the functional and information models in the course of developing a digital repository 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 the various components in the archival system to the corresponding features of the reference model. To test the OAIS compliance, the first step was to have an IR developed to satisfy the same, therefore, the pilot repository for IIG was developed.

3. PILOT INSTITUTIONAL REPOSITORY FOR IIG

It was decided by the members of the library committee to have a pilot IR for IIG, as running a pilot study would help to:

- Identify the hardware and software requirements
- Identify the skills and expertise available or that needs to be developed amongst the staff
- know ways of organising the content
- establish the service model for IR that addresses digital preservation

DSpace was chosen as the IR software for IIG for a number of reasons:

(a) DSpace open source platform is available for free to anyone and can be downloaded.

(b) It allows any organisation or individual to use, modify, and even integrate the code into their commercial application without paying any licensing fees.

(c) DSpace is deeply informed by the OAIS reference model[3].

(d) Where possible, DSpace has adopted the OAIS model and vocabulary to articulate DSpace design objectives and terminology.

(e) DSpace platform provides the tools for the IIG institutional repository to administer digital preservation within the archive, as well as to accept submissions from producers and allow access appropriately to the IIG communities.

A survey of literature led to the decision that as IIG takes responsibility for managing its digital collections through its IR (here, a pilot study), collaborative long-term digital preservation and access strategies would be advantageous. Collaboration between institutions is essential, and this collaboration requires both organisational and technical investments. Not only a pre-coordinated technological solution, but also strong, long-term inter-institutional agreements must be put in place, or there will be insufficient commitment to act in concert over time.

4. DISTRIBUTED DIGITAL PRESERVATION SERVICE

The OAIS reference model provides a framework in which preservation services in terms of IR can be constructed. At a very general level, it can be seen that IR provide a similar range of functionality as found in OAIS-input and output, data management, and storage. The OAIS imposes more formality and discipline on these processes for the purpose of long-term preservation. Thus, deposit becomes, ingest, and are concerned with archival storage, all enveloped by preservation planning, administrative and management roles. Hitchcock[4] proposes the following three models for digital preservation that provides IRs the flexibility in managing preservation risks.

- Preservation Service provider model (service provider is OAIS)
- Institutional model (institution is OAIS), here an institution may incorporate more than repository
- Software model (repository is OAIS), preservation features are built into the IR software.

This categorisation established the need for a disaggregated or distributed digital preservation service. Preservation service is the generic term used to describe third-party institutions responsible for active preservation. The preservation service should possess relevant knowledge and capacity to convert the disaggregated model into a practical implementation. It is primarily responsible for the construction and maintenance of the archival version of the e-print and the creation of supplemental metadata. The words 'disaggregated' and 'distributed' are synonymous when referred to in terms of third-party preservation service provider. Guide to Distributed Digital Preservation[5] defines distributed digital preservation service as 'the distribution, management, and maintenance of digital information over a wide geographical area and over a long period of time—maintaining its viability, authenticity, and accessibility across changing technologies, formats, and user expectations'.

The role of the preservation service provider model was identified for the IR for IIG due to the following reasons:

(a) The range of services offered by the third-party distributed preservation service provider is greater and more flexible than the institutional or the software model of OAIS for digital preservation.

(b) Distribution ensures more security, as content is distributed to locations over time.

(c) In the service provider model, the IR could be OAIS compliant, but it need not necessarily be if the service provider delivers that compliance.

(d) Compliance to OAIS standards needs technical expertise for various procedures, like for mappings various metadata schema like METS object model onto the DSpace object model or for monitoring technology which is a part of the preservation planning entity in the OAIS model. This requires a certain level of technical expertise and familiarity with programming languages, that staff handling the IIG lack. Such a lack of expertise could be compensated for such third-party distributed preservation services.

(e) Costs associated if any with this model of digital preservation would be affordable, as it is perceived that such a model allows a cooperative approach amongst a range of existing institutions to build up their capacity to ensure the viability of data over long period. This would allow costs to be sunk or shared amongst the cooperatives.

(f) Such an effort would result in the standardisation of preservation practices across multiple repositories.

(g) It would also help in the reduction of possible duplication of effort by different repositories.

To implement the distributed model, the IR and the third-party preservation service provider must clarify the terms of their agreement. Just like the SHERPA-DP[6], it is envisaged that the following three parties will perform specific functions within the disaggregated model:

- Institutional Repository of IIG--This is the institutional archive responsible for managing the ingest process and provide access to digital content.
- Preservation Service--An external organisation allocated to implement the technical infrastructure and preserve digital content in the long-term.
- Repository Advisory Board--A group, composed of advisors from different organisations that will provide advice and recommendations to the preservation service.

5. MAPPING OAIS CONFORMANCE

Having established the reasons for choosing distributed digital preservation service for the IIG pilot IR, specifications for compliance by the IIG IR and the third-party distributed preservation service provider was studied in detail, such that the IR of IIG can be entirety regarded as an OAIS. The SHERPA-OAIS model was modified to suit the requirements at hand. The preservation service, to be provided by the third-parties is best viewed as a combination of certain aspects of the OAIS ingest and archival store functions (Fig. 1).

At a high-level, the progress of an information package within the distributed OAIS model will consist of six broad stages:

(1) The depositor (OAIS producer) submits a submission information package (SIP), consisting of an e-print and associated metadata to the institutional archive.

(2) Repository staff refines the resource discovery metadata that accompanies the e-print, as defined by internal archive specifications.

(3) On a pre-determined schedule, the updated SIP is transmitted to the preservation service, the preservation service updates the SIP according to the specification provided by the IR of IIG.

(4) Preservation service also generates an archival information package (AIP) intended for preservation.

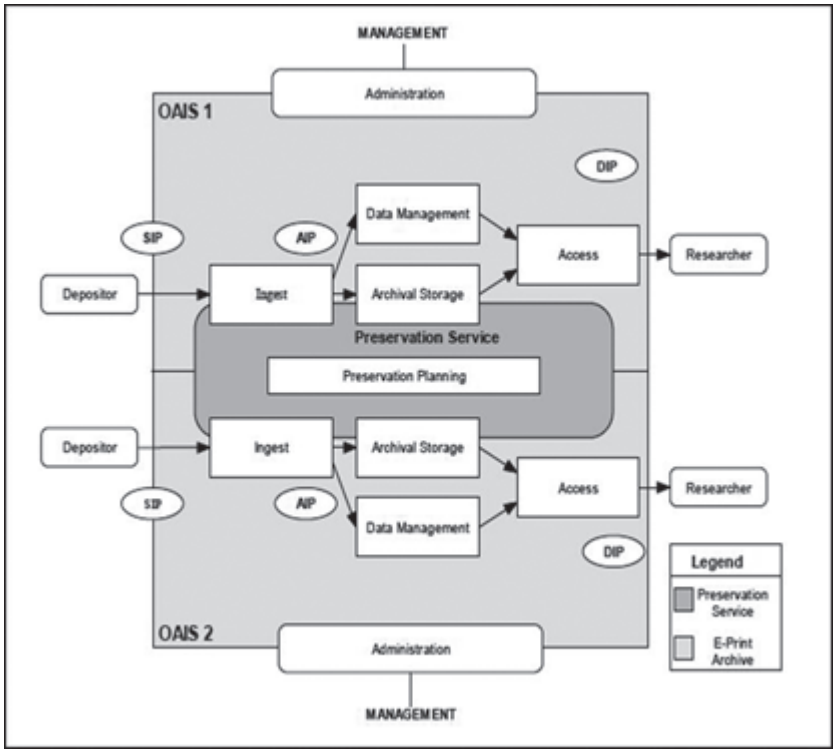

Figure 1. Adapted from the SHERPA-DP OAIS compliant model.[7]

(5) AIP is stored within the Archival Store at the third-party distributed preservation service and an appropriate backup strategy is implemented.

(6) IR of IIG generates a dissemination version intended for use by the OAIS designated community and make it available via their search catalogue. A user is able to request and download a copy of the dissemination information package.

The conformance of IR of IIG to the OAIS reference model will result in the implementation of each of the six OAIS functional entities in one form or another, in the course of developing is archival system.

5.1 Mapping OAIS Mandatory Responsibilities

The following mandatory responsibilities are listed by the OAIS. Details of the compliance to these requirements by IIG are.

(a) *Negotiates for and accepts information from producers*: The IIG library collects e-resources based on the digital collection development policy. The depositors are required to read the submission guidelines, which contains information about the type of materials to be deposited, who can deposit it, the supported file formats, the metadata policy, the take down policy, etc. Further, data depositors complete and sign a Deposit Form, which elicits information about the intellectual characteristics of the collection being deposited. The Copyright information of major publishers for the collections at IIG, gives further assurance that confidential information has been removed and that the proper approvals have been sought and granted.

(b) *Obtains sufficient control for preservation*: The Submission or Deposit Form asks the depositor to attest to the fact that he or she has copyright to the collection being deposited and thus has the authority to grant approval to the institute to redistribute the data. This form also obtains permission for the institute to migrate or transfer content for preservation purposes, as also to catalogue, enhance, validate and document the data collection. The IIG will make sure that the agreement with other organisations or external parties ensures that it does not have to preserve all the common representation information related to its content information objects.

(c) *Determines designated consumer community*: The library continuously assesses its current and potential user community. Feedback of staff, scientists and students is continuously taken face-to-face, on their experiences with the library, the collection housed within it and their expectations.

(d) *Ensures information is independently understandable*: The staff and scientists of IIG will evaluate each data collection to determine what kinds of explanatory or descriptive information need to be provided to use the data most effectively. The institute has the technical expertise to provide support to

users of the data and it should avail this support. Acknowledging the fact that no file format or physical storage media lasts forever, the IR at IIG to collect data in open and standardised file formats and would follow data migration and media refreshment strategies. All these are reflected in the Digital Preservation Policy developed for this purpose.

(e) *Follows established policies and procedures*: The digital preservation policy document has been scrutinised by the scientists of the Institute. The Institute will monitor its digital preservation policy as necessary to account for technological shifts, changes in perceived best practice and the nature of IIG library holdings.

(f) *Make information available*: The IIG will provide the designated community with access to its information and data holdings via the IR website. Restricted data will be made available through separate means or a policy decision can be taken to have tighter access constraints and made available for analysis only on-site. The goal is to provide access, even when there are stringent constraints on information use. IIG has no formalised policy on dissemination of its holdings. The access policy illustrated in its collection development' policy and illustrates basic access mechanisms employed by the IIG library, these can be further elaborated for the IR for IIG.

5.2 Mapping OAIS Functional Entities

The OAIS functional model comprises of functional entities which form a collection of six high-level services that fulfill the OAIS's dual role of preserving and providing access to the information in its custody. The conformance of IR of IIG to the OAIS reference model will result in the implementation of each of the six services or functional entities in one form or another, in the course of developing is archival system.

5.2.1 Ingest

The IIG Ingest procedure are drawn on the lines of the AHDS Archive Ingest Procedures Framework[8] and the Tufts and Yale Ingest Guide for University Electronic Records[9]. It is based on the presumption that a producer will create, acquire, organise and manage records in a recordkeeping system. Here the IR of IIG will ingest some of those records into a separate preservation system that a neutral third party administers. Following are the steps followed for ingest by the IIG institutional repository.

Section A: Negotiate Submission Agreement

This stage describes the actions needed by the producer and the repository to generate a Submission Agreement. This agreement defines the nature and scope of the records to transfer to the preservation system and how the archive will execute transfer, validation, and transformation of these records.

To begin with the repository or the information producer get into contact with the other. Information about the information producer is collected and documented. The repository and the producer come to an agreement regarding which records the repository will consider for accession. This step also determines if the producer has proper custody of the records under consideration. The administrative module of DSpace is responsible for identity management here, along with this IIG would maintain an activity log of all contacts made and each step followed.

Once the repository has confirmed it is working with the proper custodian, it will conduct a survey of the records to assess their continued value and authenticity. Such an authenticity check confirms whether the records intellectually belong in the repository, asking if it should accession the records. The repository then appraises the formats of the records that it should accession into the third party preservation system and determines if any of these records are in formats that are not one of the formats that the preservation system supports. The repository then determines if it will:

(a) Transform the records into one of the existing preservation formats, and

(b) Transform the records into a new preservation format.

The copyright status of the records is determined to see if it already has the copyright of the records, or needs to obtain the copyrights agreement for the records in an ingest. After this the repository determines if the active recordkeeping system, DSpace, managing the records in the ingest allows, or is capable of enabling, a trustworthy transfer of records to the third-party preservation system or archive in a feasible, scaleable manner. Measures to enable a trustworthy transfer may range from building special tools to undertaking a software re-engineering project. Even if the records are authentic and are stored in acceptable record types and formats, and even if there are no problems with copyrights, or access rights, it may still simply be too difficult or expensive to facilitate a trustworthy transfer from DSpace to the third-party recordkeeping system.

The repository excludes from ingest the records that it and the producer cannot or are not willing to make the effort to transfer to the third-party preservation repository in a trustworthy manner. All these are noted and/or accordingly modifiedin the Survey Report of records.

The repository then assesses itself if it can accession, manage, and preserve the records it should accession in the third-party preservation system, either on it own, or with help from the producer, allow the third-

party preservation services to do so on its own. This assessment is based on the information collected in the survey report. The repository must determine if its existing resources for preservation formats, record types, creator records, security procedures, and system capabilities meets the needs of preservation of the various records deemed for preservation. It then determines if it should modify or add to its resources to meet those assessments or if it should reject or modify the scope of the records involved in the ingest process. The repository then with the help of the third-party preservation service provider adds description of Metadata Encoding Rules, Transfer Procedures and Schedules, and Validation Procedures to the Submission Agreement. Then the repository and producer work on finalising the Submission Agreement, until they both agree to endorse it. Figure 2 below encapsulates the complete process of negotiating a submission agreement.

Section B: Transfer and Validation

This section describes the actions needed for the repository and the producer to deposit records into a third-party preservation system. The producer creates a SIP of the records submitted to the repository, with the help of the third-party preservation service provider according to the terms of the Submission Agreement. The producer will then transfer the SIP to the repository.

Validation is a quality control check to ensure minimum standards on the objects being entered into the digital archive. The minimum standards for the IIG digital archive are:

- All expected objects are present
- SIP encapsulation is correct
- All of the mandatory metadata is present
- Content is in an approved long-term preservation format
- Objects are not infected with viruses.

During this part, the repository validates the SIP and its record components received from the producer. The repository checks that the SIP and its components are well-formed and whether they contain viruses. It also validates that the producer was authorised to transfer the SIP. If the SIP fails any of these validations, the repository rejects the SIP and notifies the producer to generate another SIP. The repository then attaches to the records metadata it can automatically infer from the Submission Agreement. The third-party preservation service provider turns the records involved in an ingest into AIP according to the rules and procedures specified by the Submission Agreement of the IR of IIG. The third-party preservation service provider conducts a final appraisal of the records involved in ingest. It ensures that the records in the newly formed AIPs have the proper metadata associated with them. If the AIP does not contain the correct records, the repository rejects the records and notifies the producer to generate a new SIP for the affected records. If the records in the AIP do not have the proper metadata, the archive makes sure the metadata is enriched by the third party. The third-party preservation service provider deposits the AIPs it has formulated during ingest into its preservation system. Then it formally notifies the producer that it has accepted and accessioned the records the producer transferred to the repository in a SIP or SIPs. This is the moment of formal transfer from the repository to the third-party preservation service provider.

5.2.2 Archival Storage

The IIG third-party preservation service provider will undertake the complete responsibility of archival storage of the AIPs including providing the specified hardware as the main preservation file system, appropriate storage capacity and drivers. It is IIG third-party preservation service provider's responsibility to ascertain error checking, disaster prevention and a range of backup and recovery measures to such a preservation system. It is advisable for the IIG third-party preservation service provider to have a Disaster Recovery Planning Guide. The IIG will rely on advanced technology from its preservation service provider for storage management and media checking regime, but will also use a policy of multiple copies of every AIP to reduce the risk of data loss.

5.2.3 Data Management

Metadata management is carried out within the DSpace software itself. Basic information about each SIP is captured at Ingest within DSpace, that also has the facility to enhance extremely detailed metadata, outputting the results as standard XML files. Any paper documents relating to a SIP will be digitised and attached to the AIP. Reports from the DSpace can be generated to show what records have been accessioned, their status, the XML metadata for each record, etc. More specific reports can be hand-crafted through direct SQL queries and scripts to read the XML.

5.2.4 Administration

The Institute will appoint and delegate responsibility for the service administration that would take care of the strategic, financial and personal elements of the organisation. While the administration function can be clearly vested in the Director, who will maintain direct financial control, the specific functions can be spread across the organisation. Submissions to the IR of IIG can be reviewed and approved by the committee that will be scheduled to meet fortnightly and review all new acquisitions or submissions that have been received in the period since the last meeting. When need arises expertise can be sought from the relevant subject

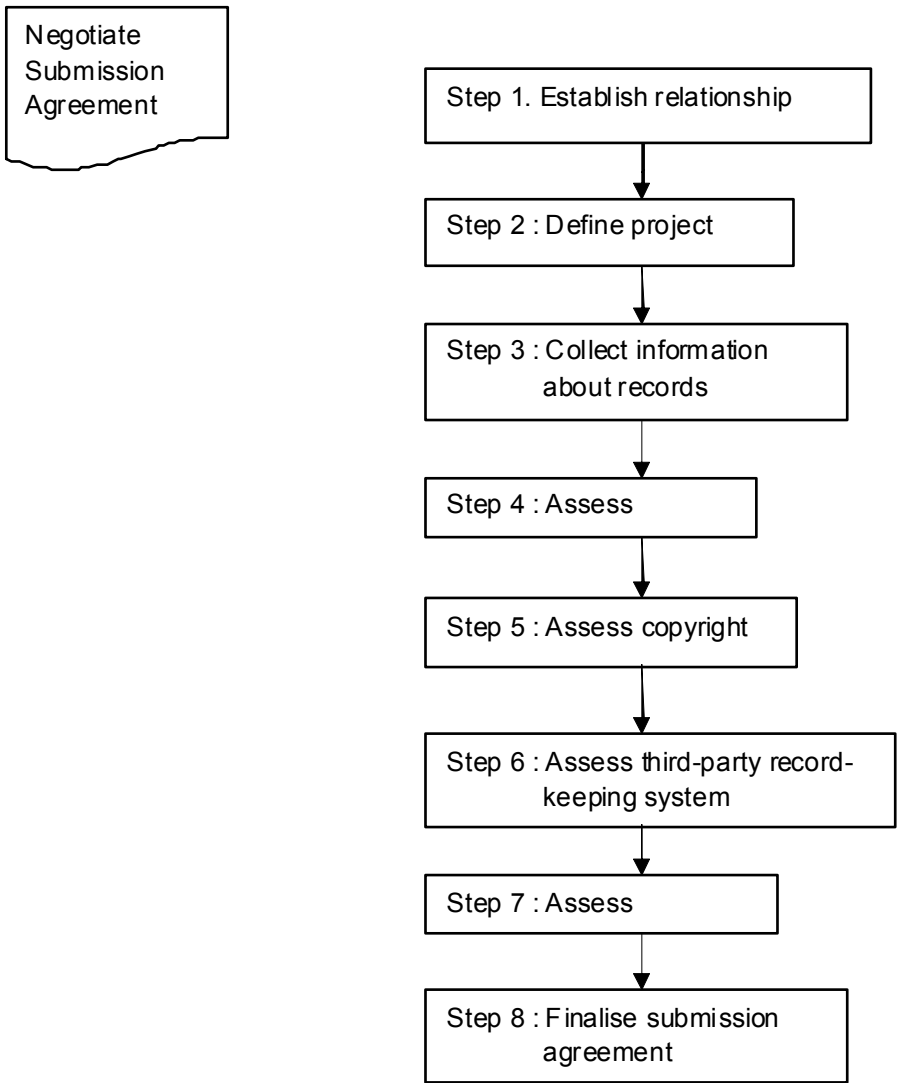

Figure 2. Negotiating a submission agreement.

experts. The appointed committee will develop all the management policies and issues that relate to the acquisition function of the archive. As a matter of principle, the IIG Institutional Repository can aim to apply open standards. A work plan will be created for each 'accepted' submission specifying that:

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

At the submission stage, there will be formally signed Submission Agreement, which will be handled by the library acquisitions staff. This section will work closely with data depositors to provide guidance and advice with regards to data creation and deposit

5.2.5 Preservation Planning

The third-party distributed preservation service provider must monitor and take help of the international community on the preservation of digital objects. The preservation service will monitor technical developments in the digital preservation community in file format characterisation, automated risk assessment, and migration as a mode of format preservation. The preservation service provider will monitor the status of formats for which instances exist in the archive. The IIG third-party preservation service provider will develop, review and implement various preservation strategies that best suits the institute.

5.2.6 Access

The management of IIG's DSpace catalogue will be done by the library staff engaged in archiving, who will be responsible for structure and quality control. Catalogue records are created during ingest processing, when the main metadata record is generated. Fundamental authorisation takes place via the DSpace software that grants permission to access the resource or not.

6. CONCLUSIONS

Thus, the compliance testing of the pilot IIG institutional repository against the OAIS mandatory responsibilities results in a good match of expectations of the institute. It resulted in the development of a regulatory framework to ensure long-term preservation of digital objects within the IR of IIG. It successfully identified the need and role for a third-party third-party distributed preservation service. The various policies, guidelines and agreements developed during the course of the study, like the collection development policy, the submission agreement and the submission guidelines, etc., when exercised would result in a strong control over the producers of material that is handed over to the archive.

The OAIS model frequently points to the strong link between the user community and the way the material in an archive should be described and preserved. To this end with limited user groups, data contained in the institutional repository will not be subject to many uses. The IR of IIG has an identifiable and relatively homogeneous user community. The collection within the IR of IIG will not be large enough to solicit different preservation strategies. The designated community concept assumes an identifiable and relatively homogeneous user community, which is supported by the limited user group of the IR of IIG.

Amongst operating archives, mapping OAIS conformance reveals the biggest discrepancy often to be the understanding of the producer and the designated consumer community. While in an ideal world the archive would exercise strong control over the producers of materials that is 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. An in situ digital repository like the pilot institutional repository of IIG endowed with the responsibility of conformance to the OAIS would address this issue right at the stage of creation of the digital object.

REFERENCES

1. Reference model for an open archival information system (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf

2. Lavoie, Brian F. The open archival information system reference model: Introductory guide. 2004. http://www.dpconline.org/docs/lavoie_OAIS.pdf

3. Tansley, Rovert; Bass, Mick & Smith, MacKenzie. DSpace as an open archival information system: Current status and future directions. *In* ECDL 2003, LNCS 2769, edited by T. Koch & I.T. Solvberg. Springer-Verlag, Berlin, 2003, pp.446-60.

4. Hitchcock, Steve; Brody, Tim; Hey, Jessie M.N. & Carr, Leslie. Digital preservation service provider models for institutional repositories: Towards distributed services. *D-Lib Magazine,* 2007, **13**(5/6), http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

5. Skinner, Katherine & Schultz, Matt. A guide to distributed digital preservation. 2010. http://www.metaarchive.org/sites/default/files/GDDP_Educopia.pdf

6. Sherpa-DP. http://www.sherpadp.org.uk/

7. Knight, Gareth. Sherpa-DP OAIS Report: An OAIS compliant model for disaggregated services. 2005. http://www.sherpadp.org.uk/documents/sherpa-dp-oais-report.pdf

8. Ruusalepp, Taivo. AHDS archive ingest procedures framework, 2003. http://www.ahds.ac.uk/preservation/ingest-procedures-review.pdf

9. Tufts and Yale Ingest Guide for University Electronic Records. Fedora and Preservation of University Electronic Records: 2.1 Ingest Guide. 2006.

10. http://dca.lib.tufts.edu/features/nhprc/reports/ingest/index.html