Digital Preservation Center of NSLC

Zhenxin Wu
National Science Library, Chinese
Academy of Sciences
33 Beisihuan Xilu, Zhongguancun
Beijing P.R.China ,10019
+86-(10)-82628382
wuzx@mail.las.ac.cn

ABSTRACT
This paper briefly introduces the context and major functions of Digital Preservation Center of National Science Library, Chinese Academy of Sciences (NSLC), and describes its digital preservation system, as well as its preservation services and future work.

Categories and Subject Descriptors
H.3.6 [**INFORMATION STORAGE AND RETRIEVAL**]: Library Automation – Large text archives

General Terms
Management.

Keywords
Digital Preservation Center, Digital Preservation System, Preservation Services.

1. CONTEXT
It is generally acknowledged that digital literature has become the main mode for creating, publishing and disseminating academic information in science and technology fields. This is true in China and globally. The most important function for the staff at the National Science Library, Chinese Academy of Science (NSLC) is to guarantee access to this literature, not only for researchers of the Chinese Academy of Sciences (CAS) who have come to rely on this literature but also with mandate for researchers in the natural sciences and high technology fields across China.

NSLC has been working to build digital infrastructure for digital resources management, in which long-term preservation is regarded as one of key importance, commanding increased attention and investment. A key part of this infrastructure is a digital preservation system (DPS) [1] for preserving digital resources purchased from commercial publishers. Recently NSLC has signed preservation agreements with six publishers and has signed preservation service agreement with three domestic information institutions in the national archive manner. At the time of writing, there are 23,633 electronic journals archived in DPS, from Springer Verlag, Institute of Physics Publishing (IOP), Nature Publishing Group (NPG), BioMed Central, Royal Society of Chemistry (RSC), Chinese VIP STM journals (VIP): more than 28 million articles and 100 million files. The development of processes for preserving eBooks is in progress.

The Institutional Repository Grid (IRGrid) [2] is another component in the digital infrastructure at NSLC, for collecting and storing publications written by researchers of the Chinese Academy of Science. Additional components are being added. Later in 2013, information will be launched about web archiving for the networked resources regarded as important for science and technology and preservation of scientific data.

NSLC has contributing to efforts for long-term preservation over a ten year period, promoting national developments and participating in international meetings, including hosting the iPRES conference in 2004 and 2007. It also now reports its activity on archiving e-journal content to the international Keepers Registry [3] facility.

2. DIGITAL PRESERVATION CENTER
Staff at the National Science Library, Chinese Academy of Science (NSLC) identified four functions [4] for a Digital Preservation Centre (DPC) [5].

(1) Strategy & Planning. It is important at the outset to clarify the objectives for the intended preservation services, defining the scope and selection criteria for archival content, and determining the preservation procedures and processes.

(2) Rights Protection & Management. The interests of stakeholders must be taken into account in order to keep access to resources sustainable. Stakeholders include all parties in the supply chain: libraries as purchasers and their users, publishers, service providers and agents, and the authors/producers of the literature. NSLC implements a comprehensive rights protection and management for its digital preservation.

(3) A Trusted National Archiving System. NSLC wished to provide nationwide preservation services that complied with international standards, referencing international best practices. This is in order to provide entire preservation life-cycle management using scalable technology, one that could be sustainable, reliable and efficient. The infrastructure for preservation is gradually forming for the provision of preservation services nationwide.

(4) Promoting a Cooperative Preservation Network. NSLC has been dedicated to promoting the development of digital preservation nationally through cooperation with major domestic libraries and information institutions. This includes developing and sharing policy and practice, knowledge of standards, and sharing digital preservation services countrywide. An initial step has been a collaborative network for coordinated distribution of multiple secure copies and replacing each other to provide preservation services when necessary. Furthermore, public certification and audits which ensure standardized and transparent cooperation management will be provided within the network.

3. DIGITAL PRESERVATION SERVICES
NSLC has succeeded in its planning and implementation and has established a digital preservation system. This is initially being operated for NSLC and three national organizations with

agreement to preserve the e-journals which they subscribe from Springer, with provision for public access given a trigger event.

The digital preservation system is a dark archive system with have two service platforms:

(1) The archival data management platform. Only accessed by representatives of the national organizations that have agreed to archive their content, this web platform enables auditing of the archived resources, with facility for regular automatic report sent by the DPS. It also allows management of the associated subscription information.

(2) The public access platform. This is intended for the users of the national organizations and for access to subscribed content which may have been triggered according to certain procedures. Similar to the common publishers' service platform, this platform provides browse and search functions, as well as full-text download restrictions based on the subscription information with monitoring of malicious download behavior, providing monthly usage statistics and reports.

Regular audits already have been planned on the schedule and will be carried out by a third party expert group drawn from the National Science and Technology Libraries Group [6] and the Chinese Academic Library and Information System. There is also reporting into The Keepers Registry: for example, search http://thekeepers.org for 'Chinese Journal of Chemical Physics' (1674-0068). In the future, NSLC will provide more services, such as public certification and audit services.

4. DIGITAL PRESERVATION SYSTEM

The IT system department of NSLC is responsible for designing and developing the DPS, as a digital preservation system that was in compliance with the Open Archival Information Systems (OAIS) standard [7]. This includes systematic procedures and policies for the entire lifecycle management. Meanwhile audit management and access control based on preservation agreement have been provided, and system security management and multi-level disaster recovery mechanism have been established, noting the Trustworthy Repositories Audit and Certification (TRAC) standard [8].

The system architecture of the DPS is shown in Figure 1.

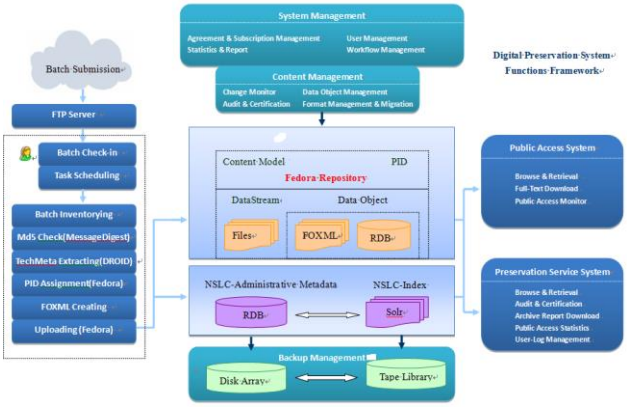

Figure I. Digital preservation system function framework

The core of DPS is the open-source Fedora repository suite [9]. This meant much custom extending and developing had to be done. The DPS can ingest different formats included in the submitted information packages (SIP) and transform these into unified format information packages according to Fedora digital object model, and then store these in the Fedora repository. Utilizing the API provided by Fedora, DPS provides many functions to manage the digital objects, such as fixity checks, version management, etc. With an external relational database, the DPS can store and manage all the metadata which be required for the preservation life-cycle management, from the ingest of the SIP to access of the archived content. With indexing documents, the DPS is easier for browsing, retrieving and statistics. The DPS established at the NSLC is able to provide a complete ingest workflow management and basic preservation management.

4.1 Customizable Ingestion Process

As noted, the DPS can receive and process digital content in a variety of formats, and then generate a unified format for each SIP ingested. Therefore there is a great demand to provide a more flexible workflow mechanism, which can be customized for processing different format SIP. Based on the concept of modular programming, the ingest stage is divided into several modules each with minimum function, such as batch inventory, virus detection, SIPs backup and SIPs transmission, package unzipping, document format validation and technical metadata extraction, metadata verification, standard SIP creation, standard SIP validation, SIPs uploading. The archivist can deploy an appropriate mix by combining different modules to meet the special requirement for a particular SIP.

Function Module Description:

(1) Batch Inventory. Creating and submitting an inventory list of SIPs in order to carry out MD5 checks by using the Java Message Digest [10]. This ensures that the SIPs are unchanged during the transmission process, and later, this list will be use to review against list submitted by the publishers.

(2) Virus Detection. DPS runs Kaspersky [11] on all SIPs for virus scanning.

(3) SIPs Backup and SIPs Transmission. By using the API of JAVA FTPClient [12], DPS replicates the SIPs onto a pre-specified FTP site as a backup, copying to the specified working directory, which is prepared for the next step in the ingest process.

(4) Unzipping Package. DPS runs tar or unzip command to decompress different format data package.

(5) Document Format Validation and Technical Metadata Extraction. Taking into account the efficiency and the quality of the metadata extraction, DPS use Apache PDFBox [13] for PDF format validation and technical metadata extracting, and uses Droid [14] for other format documents.

(6) Metadata Verification. According to the metadata specification agreed with the publishers, DPS verifies the metadata content by using SAX [15].

(7) Standard SIP (FOXML) Creation. DPS adopts the Fedora FOXML [16] as the standard SIP format. It uses SAX to parse the original XML file and extract associated metadata to create a FOXML file, then generates a unique identifier (PID) using the Fedora API-M, establishing the relationships between objects and their data streams. The original XML files, PDF files and other multimedia files are

copied onto the designated directory, as parts of a data object, to be uploaded with the FOXML file.

(8) Standard SIP Validation. Before uploading, DPS once again checks all of the components of the SIPs (including internal and external content).

(9) Uploading SIPs. DPS provides local and remote uploading modes. In the remote mode, DPS uses the Fedora SOAP APIs directly in order to ingest a SIP into Fedora, This keeps flexibility. In the local mode, DPS uses Fedora's underlying function directly without using APIs, which greatly improves the efficiency of uploading.

4.2 Basic Preservation Management

Taking advantage of the features and functions in Fedora, DPS has developed many basic preservation management functions. This use of the Fedora API-A and API-M includes:

(1) Browsing and retrieval of archival data
(2) Multi-level (collection, journal, paper) audit (integrity & fixity)
(3) Archival data maintenance
(4) Tracking changes on data objects
(5) Statistic & Report
(6) Document format management and data migration.

4.3 Agent Execution Mode

It is important to monitor and look for ways to reduce human resource costs and improve efficiency of the system. For this purpose, DPS develop an agent module to execute ingest and other processes deployed by the archivist. The agent module runs automatically in the background which greatly improves the automation level of system. For example, after receiving the data package submitted by publishers, the archivist logs into the DPS to register the receipt of the data and to customize the ingest task. This includes checking the predefined profile (including the designated backup directory and the designated work directory, the selected workflow, etc.) and task scheduling. The Agent monitors task instructions and starts the background data process automatically, the results are sent to the archive administrator by email after the task is completed.

5. FUTURE WORK

There is still much to be done, and to be accorded priority of attention and effort. The following list is put forward for comment and consideration:

(1) Signing preservation agreements with more other publisher, in order therefore to ingest more e-journal content
(2) Serving as the core (node) for domestic long-term preservation community, promoting progress of preservation nationwide
(3) Increasing the types of resource archived: web archiving for important network resources and preservation of scientific data
(4) Increasing preservation service agreements with more resources and more customers
(5) Playing the leading role for the national digital preservation network of China.

6. ACKNOWLEDGMENTS

Thanks are due to all my colleagues who contribute to digital preservation activity at NSLC, especially to Honghu Fu, Yuju Wang and Li Qian from IT department.

7. REFERENCES

NOTE: all URLs successfully accessed June 16, 2013

[1] Digital preservation system (limited access), http://dps.las.ac.cn.

[2] Institutional Repository Grid of CAS, http://www.irgrid.ac.cn/.

[3] The International Keepers Registry , http://thekeepers.org

[4] Xiaolin Zhang, Jiancheng Zheng , Zhenxin Wu, etc. 2012.The Long-term preservation strategy of NSLC (Internal document).

[5] Digital Preservation Centre (DPC) of NSLC, http://dpc.las.ac.cn.

[6] National Science and Technology Library (NSTL), http://www.nstl.gov.cn/.

[7] Open Archival Information Systems (OAIS), http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284.

[8] Trustworthy Repositories Audit and Certification (TRAC), http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56510.

[9] Open-source Fedora repository, http://fedora-commons.org/

[10] Java Message Digest (A class of JAVA which provides applications the functionality of a message digest algorithm, such as MD5 or SHA), http://doc.java.sun.com/DocWeb/api/java.security.MessageDigest

[11] Kaspersky (An antivirus and internet security software), www.kaspersky.com/

[12] FTPClient (A tool which encapsulates all the functionality necessary to store and retrieve files from an FTP server), http://commons.apache.org/proper/commons-net/apidocs/org/apache/commons/net/ftp/FTPClient.html.

[13] Apache PDFBox (An open source Java tool for working with PDF documents), http://pdfbox.apache.org/

[14] Droid (Digital Record Object Identification, is an automatic file format identification tool developed by the National Archives, U.K.), http://sourceforge.net/projects/droid/

[15] SAX (An open source Java-only API for XML), http://www.saxproject.org/

[16] FOXML (A simple XML format that directly expresses the Fedora digital object model), http://fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html.