An introduction, for librarians, to long term digital preservation of the digital academic research record by academic libraries

Hilton Gibson - hgibson@sun.ac.za

ABSTRACT

The advent of the internet and the invention of the hypertext system has ushered in an information revolution, whereby anything that can be digitised can be easily reproduced and distributed at almost zero cost. However in some cases, the information revolution has been subverted and information has become an expensive and artificially scarce commodity. The most notable case of this subversion is occurring with the publication of academic research articles. The purpose of this paper is to explain how the subversion of the publication of academic research articles occurred and then to suggest long term sustainable remedies.

CHRONOLOGY OF THE DIGITAL ACADEMIC PUBLISHING PROBLEM

THE INTERNET IS BORN TO SHARE AND STORE ACADEMIC INFORMATION

In March 1989 Tim Berners-Lee published a paper at CERN proposing the world wide web using the inter-networking infrastructure developed by Vincent Cerf and Robert Khan in the 1970's and published as RFC 675. In January 1997 the HTTP 1.1 protocol is published as RFC 2068.

The original purpose of the world wide web (WWW) was to share and store information related to academic research.

THE SUBVERSIVE PROPOSAL AND THE SERIALS CRISIS

In 1994 Stevan Harnard presented a paper entitled "PUBLICLY RETRIEVABLE FTP ARCHIVES FOR ESOTERIC SCIENCE AND SCHOLARSHIP: A SUBVERSIVE PROPOSAL" whereby he proposed that academic authors archive digital articles on anonymous public FTP servers.

However, this did not happen because academic societies chose to publish via commercial publishing houses. Initially the commercial publishing houses provided cost effective publishing services but then costs started to escalate over the years, more than the annual consumer price index. This rapid cost increase has put enormous pressure on academic library budgets and has led to the "serials crisis".

OPEN ACCESS REMEDY

In February 2002 the Budapest open access declaration is published. In April 2003 the Bethesda open access statement is published. In October 2003 the Berlin open access declaration is published.

Open access is seen as the remedy to the "serials crisis" and a concept to provide public access to publicly financed research. In addition funders, public and private are beginning to mandate open access to research funded by themselves.

DIGITAL PRESERVATION BY ACADEMIC LIBRARIES AS THE LONG TERM REMEDY FOR THE ACADEMIC PUBLISHING PROBLEM

It is proposed that academic libraries provide platforms for storing and preserving the digital outputs of the academic research process. The platforms developed and maintained should be based on open digital technologies to ensure the best possible chance of the technology surviving for the benefit of future researchers. Digital preservation is a long term strategy that has two essential components:

1. Preserve and maintain a digital container for the digital research objects.
2. Preserve and maintain the digital research objects themselves.

If it is accepted that academic libraries should preserve the digital academic record then the next question is how to implement such a service and ensure that is available now and in the future?

CAPACITY PLANNING FOR DIGITAL PRESERVATION BY ACADEMIC RESEARCH LIBRARIES

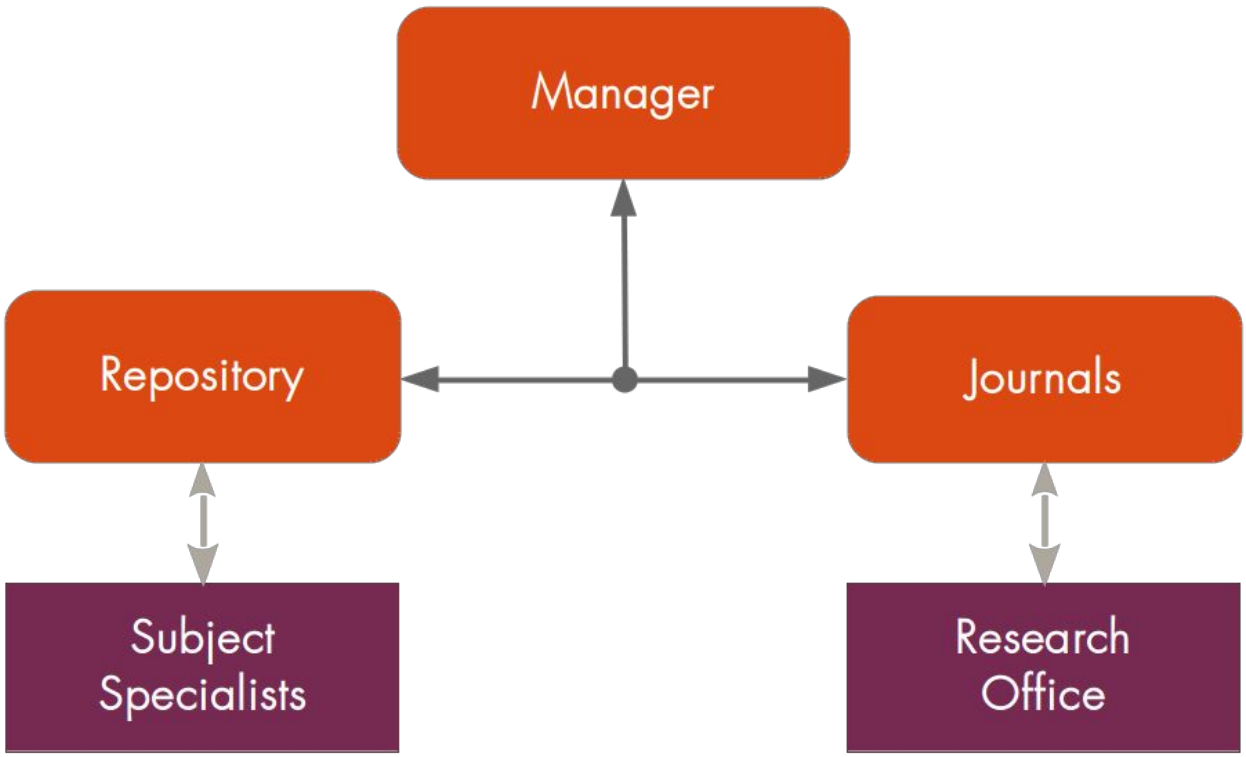
The platform for storing the research digital objects is normally called an institutional repository. Therefore a digital preservation plan for the repository and the digital objects in the repository should exist. This digital preservation plan should ensure that there is ample capacity for the long term operation and maintenance of the repository.

The capacity for the long term preservation of the repository and the digital objects contained therein can be facilitated by two distinct groups, namely:

1. An operations team.
2. A systems team.

The bare minimum for a digital preservation group with the bare minimum for hardware is detailed below.

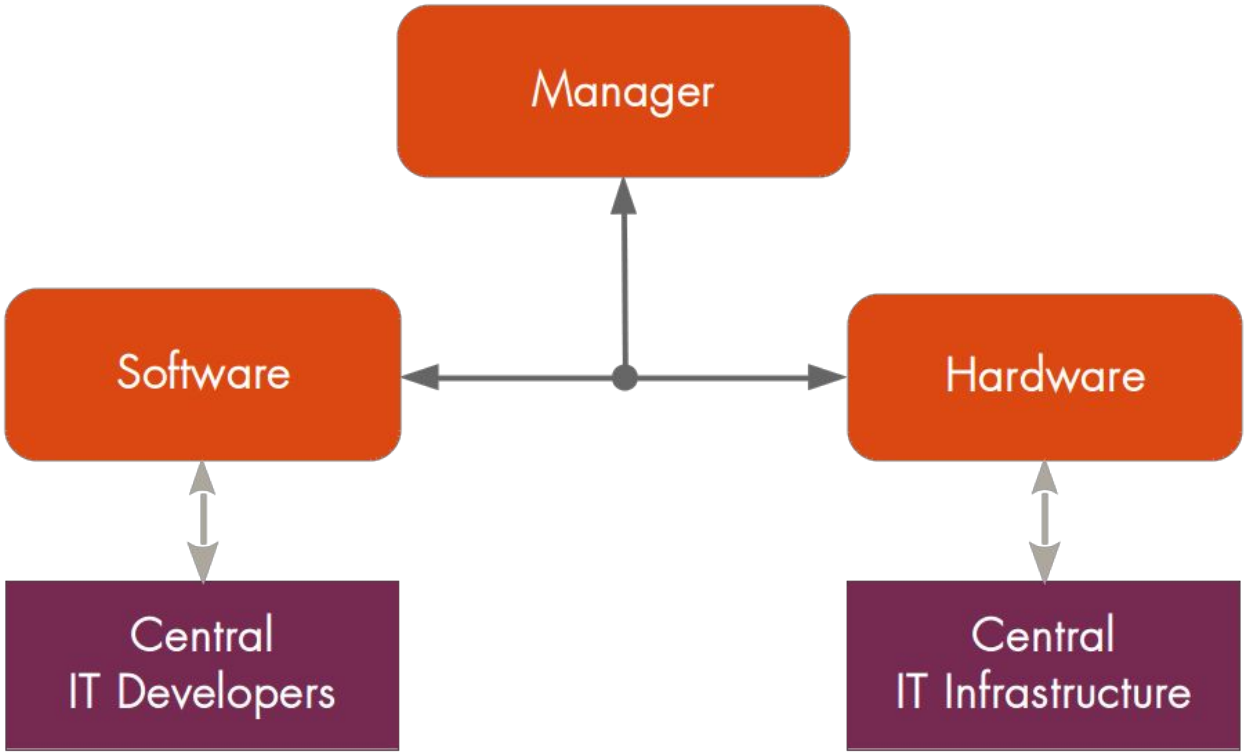
Operations Team



The operations team consists of the following persons:

- Director
- Manager
- Repository Librarian and Journal Librarian

Systems Team



The systems team consists of the following persons:

- Manager
- Java Web App Developer
- Ubuntu Linux System Administrator

For further details of capacity building and planning, please go to:
http://wiki.lib.sun.ac.za/index.php/SUNScholar/Capacity_Building

Proposed Business Model

From the team definitions above it is possible to determine a provisional business model. The only missing component is the cost of hardware which will be dealt with first.

Hardware Costs

These hardware costs can be amortised over 4 years, which is the normal warranty period for hardware.

Production Server	R250,000.00 each	R250,000.00
Backup Server X2	R150,000.00 each	R300,000.00

Internet Connection	Usually for the central IT department to provide.	n/a

Personnel Costs

Operations and Systems Managers	R450,000 each	R900,000.00
Operations Librarians X2	R350,000.00 each	R700,000.00
Systems Technicians X2	R350,000.00 each	R700,000.00

Therefore annual costs are as follows:

Hardware => (R250,000.00 + R300,000.00) / 4 years = R137,500.00 annually

Personnel => R900,000.00 + R700,000.00 + R700,000.00 = R2,300,000.00 annually

Total Cost => R137,500.00 + R1,850,000.00 = **R 2, 437,500.00 annually**

APPENDIX - OPEN DIGITAL TECHNOLOGIES

For long term digital preservation, the use of open digital technologies is critical, because we have no reliable way of predicting what technologies will be used by future researchers. The best bet to ensure availability in the future is to use open technologies. Now that we have established the need for open digital technologies, the next step is to identify the open digital technologies available today.

File formats (bitstreams)

The first consideration should be the file formats used **which should always be uncompressed**. Only file formats that have open, royalty free and patent free, published digital format standards and metadata schemas should used. Examples are provided below:

Documents

The following are the recommended open document formats:

- Open document formats at: http://opendocumentformat.org
- PDF open formats at: http://www.pdfa.org

Images

The following are the recommended open image formats:

- https://en.wikipedia.org/wiki/BMP_file_format
- https://en.wikipedia.org/wiki/Portable_Network_Graphics
- https://en.wikipedia.org/wiki/Tagged_Image_File_Format

Audio

The following are the recommended open audio formats:

- https://en.wikipedia.org/wiki/FLAC
- https://en.wikipedia.org/wiki/Vorbis

Video

The following are the recommended open video formats and containers:

- https://en.wikipedia.org/wiki/WebM
- https://en.wikipedia.org/wiki/Theora
- https://en.wikipedia.org/wiki/Matroska

Database

The following are the recommended open database formats:

- https://en.wikipedia.org/wiki/Comma-separated_values
- https://en.wikipedia.org/wiki/SQL

Systems Software

The next consideration regarding open digital technologies is the software used to build and maintain the repository because the repository itself must also be preserved for future use. A digital repository is built using server hardware and a server operating system on top of which is installed the actual repository software. For the server operating system and repository software it is recommended that open source software be used.

Open Source Server Operating System

The recommended open source server operating system software is Ubuntu LTS. For more details about the selection of Ubuntu as the server operating system, please go to:
http://wiki.lib.sun.ac.za/index.php/SUNScholar/DSpace/Why_Ubuntu_Server

Open Source Repository Software

The recommended open source repository software is DSpace. For a detailed listing of available repository software products, please go to: http://wiki.lib.sun.ac.za/index.php/List_of_Repository_Software