erpanet

Information Society
Technologies

erpastudies

project gutenberg

Electronic Resource Preservation and Access Network

erpanet

www.erpanet.org

ERPANET – Electronic Resource Preservation and Access Network – is an activity funded by the European Commission under its IST programme (IST-2001-3.1.2).

Further information on ERPANET and access to its other products is available at Hhttp://www.erpanet.orgH.

A great deal of additional information on the European Union is available on the Internet. It can be accessed through the Europa server (Hhttp://europa.eu.intH).

ISSN 1741-8682

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK

Table of Contents

Executive Summary

Project Gutenberg is the first and largest collection of eBooks made freely available to the public. The project philosophy is that the greatest value of computers is not their computing power, but rather their potential for the searchable storage and retrieval of library materials. The premise for the project is that any object – whether text, picture, sound or 3D image – that can be entered into a computer can be replicated indefinitely. The eBooks generated by Project Gutenberg are stored on two main servers and can then be downloaded to local servers around the world. This case study differs form many other ERPANET studies in that the project is volunteer-driven. As such, there are no financial or business incentives to preserve the eBooks. The real incentive lies in the belief that literary works in the public domain should be freely accessible to as many people as possible for as long as possible. By digitising everything in 'plain vanilla ASCII' as well as many other formats, the eBooks are readable by over 99% of computer operating systems. By avoiding proprietary formats whenever possible, Project Gutenberg also helps to guarantee the long-term survival of the bit streams of the eBooks. The combination of open formats and the proliferation of copies downloaded around the world should ensure that the Project Gutenberg eBooks currently in existence are still accessible far into the future.

Chapter 1: The ERPANET Project

The European Commission and Swiss Confederation funded ERPANET Project[1] (Electronic Resource Preservation and Access Network) works to enhance the preservation of cultural and scientific digital objects through raising awareness, providing access to experience, sharing policies and strategies, and improving practices. To achieve these goals ERPANET is building an active community of members and actors, bringing together memory organisations (museums, libraries and archives), ICT and software industry, research institutions, government organisations, entertainment and creative industries, and commercial sectors. ERPANET constructs authoritative information resources on state-of-the-art developments in digital preservation, promotes training, and provides advice and tools.

ERPANET consists of four partners and is directed by a management committee, namely Seamus Ross (HATII, University of Glasgow; principal director), Niklaus Bütikofer (Schweizerisches Bundesarchiv), Hans Hofman (Nationaal Archief/National Archives of the Netherlands), and Maria Guercio (ISTBAL, University of Urbino). At each of these nodes a content editor supports their work, and Peter McKinney serves as a co-coordinator to the project. An Advisory Committee with experts from various organisations, institutions, and companies from all over Europe give advice and support to ERPANET.

[1] ERPANET is a European Commission funded project (IST-2001-32706). See www.erpanet.org for more details and available products.

Chapter 2: Scope of the Case Studies

While theoretical discussions on best practice call for urgent action to ensure the survival of digital information, it is organisations and institutions that are leading the drive to establish effective digital preservation strategies. In order to understand the processes these organisations are undertaking, ERPANET is conducting a series of case studies in the area of digital preservation. In total, sixty case studies, each of varying size, will investigate awareness, strategies, and technologies used in an array of organisations. The resulting corpus should make a substantial contribution to our knowledge of practice in digital preservation, and form the foundation for theory building and the development of methodological tools. The value of these case studies will come not only from the breadth of companies and institutions included, but also through the depth at which they will explore the issues.

ERPANET is deliberately and systematically approaching disparate companies and institutions from industry and business to facilitate discussion in areas that have traditionally been unconnected. With these case studies ERPANET will broaden the scope and understanding of digital preservation through research and discussion. The case studies will be published to improve the approaches and solutions being developed and to reduce the redundancy of effort. The interviews are identifying current practice not only in-depth within specific sectors, but also cross-sectorally: what can the publishing sector learn from the aeronautical sector? Eventually we aim to use this comparative data to produce intra-sectoral overviews.

This cross-sectoral fertilisation is a main focus of ERPANET as laid out in its Digital Preservation Charter.[2] It is of primary importance that disparate groups are given a mechanism through which to come together as best practices for digital preservation are established in each sector.

Aims

The principal aims of the study are to:

- build a picture of methods and match against context to produce best practices;
- accumulate and make accessible information about practices;
- identify issues for further research;
- enable cross-sectoral practice comparisons;
- enable the development of assessment tools;
- create material for training seminars and workshops; and,
- develop contacts.

Potential sectors have been selected to represent a wide scope of information production and digital preservation activity. Each sector may present a unique perspective on digital preservation. Organisational and sectoral requirements, awareness of digital preservation, resources available, and the nature of the digital

[2] The Charter is ERPANET's statement on the principles of digital preservation. It has been drafted in order to achieve a concerted and co-ordinated effort in the area of digital preservation by all organisations and individuals that have an interest and share these concerns.
http://www.erpanet.org/charter.php.

object created place unique and specific demands on organisations. Each of the case studies is being balanced to ensure a range of institutional types, sizes, and locations.

The main areas of investigation included:

- perception and awareness of risk associated with information loss;
- understanding how digital preservation affects the organisation;
- identifying what actions have been taken to prevent data loss;
- the process of monitoring actions; and,
- mechanisms for determining future requirements.

Within each section, the questions were designed to bring organisational perceptions and practices into focus. Questions were aimed at understanding impressions held on digital preservation and the impact that it has had on the respective organisation, exploring the awareness in the sector of the issues and the importance that it was accorded, and how it affected organisational thinking. The participants were asked to describe, what in their views, were the main problems associated with digital preservation and what value information actually had in the sector. Through this the reasons for preserving information as well as the risks associated with not preserving it became clear.

The core of the questionnaire focused on the actions taken at corporate level and sectoral levels in order to uncover policies, strategies, and standards currently employed to tackle digital preservation concerns, including selection, preservation techniques, storage, access, and costs. Questions allowed participants to explore the future commitment from their organisation and sector to digital preservation activities, and where possible to relate their existing or planned activities to those being conducted in other organisations with which they might be familiar.

Three people within each organisation are targeted for each study. In reality this proved to be problematic. Even when organisations are identified and interviews timetabled, targets often withdrew just before we began the interview process. Some withdrew after seeing the data collection instrument, due in part to the time/effort involved, and others (we suspect) dropped out because they realised that the expertise was not available within their organisation to answer the questions. The perception of risks that might arise through contributing to these studies worried some organisations, particularly those from sectors where competitive advantage is imperative, or liability and litigation issues especially worrying. Non-disclosure agreements that stipulated that we would neither name an organisation nor disclose any information that would enable readers to identify them were used to reduce risks associated with contributing to this study. In some cases the risk was still deemed too great and organisations withdrew.

Chapter 3: Method of Working

Initial desk-based sectoral analysis provides ERPANET researchers with essential background knowledge. They then conduct the primary research by interview. In developing the interview instrument, the project directors and editors reviewed other projects that had used interviews to accumulate evidence on issues related to digital preservation. Among these the methodologies used in the Pittsburgh Project and InterPARES I for target selection and data collection were given special attention. The Pittsburgh approach was considered too narrow a focus and provided insufficient breadth to enable full sectoral comparisons. On the other hand, the InterPARES I data collection methodology proved much too detailed and lengthy, which we felt might become an obstacle at the point of interpretation of the data. Moreover, it focused closely on recordkeeping systems within organisations.

The ERPANET interview instrument takes account of the strengths and weaknesses from both, developing a more focused questionnaire designed to be targeted at a range of strategic points in the organisations under examination. The instrument[3] was created to explore three main areas of enquiry within an organisation: awareness of digital preservation and the issues surrounding it; digital preservation strategies (both in planning and in practice); and future requirements within the organisation for this field. Within these three themes, distinct layers of questions elicit a detailed discovery of the state of the entire digital preservation process within participants' institutions. Drawing on the experience that the partners of ERPANET have in this method of research, another important detail has been introduced. Within organisations, three categories of employee were identified for interview: an Information Systems or Technology Manager, Business Manager, and Archivist / Records Manager. In practice, this usually involved two members of staff with knowledge of the organisation's digital preservation activities, and a high level manager who provided an overview of business and organisational issues. This methodology has allowed us to discover the extent of knowledge and practice in organisations, to understand the roles of responsibility and problem ownership, and to appreciate where the drive towards digital preservation is initiated within organisations.

The task of selecting the sectors for the case studies and of identifying the respective companies to be studied is incumbent upon the management board. They compiled a first list of sectors at the very beginning of the project. But sector and company selection is an ongoing process, and the list is regularly updated and complemented. The Directors are assisted in this task by an advisory committee.[4]

[3] See http://www.erpanet.org/studies/index.php. We have posted the questionnaire to encourage comment and in the hope that other groups conducting similar research can use the ideas contained within it to foster comparability between different studies.

[4] See www.erpanet.org for the composition of this committee.

Chapter 4: Project Gutenberg

http://www.gutenberg.net

Project Gutenberg produces free electronic versions of literature and reference works that are in the public domain. As the project has only a few paid staff members,[5] the majority of eBooks are scanned and edited by volunteers. Available via the Internet since 1994, Project Gutenberg is the oldest producer of freely accessible, electronic books (eBooks). From 1971 until 1997 over 1,100 eBooks were created. In the first eleven weeks of 2004 alone, three hundred new eBooks have been generated. There are now over 13,380 eBooks available and the production of eBooks is constantly increasing. Project Gutenberg is dedicated to making these resources available to the general public in a form that the vast majority of the computers, programs and people can easily read (ASCII). However, most texts are available in a wide range of formats for users to select.

New features have been added recently to Project Gutenberg's core services. Specifically, the new Radio Gutenberg[6] makes audio and video files accessible to the public for download as well as broadcasts on their two radio channels. Gutenberg Music[7] makes digitised sheet music accessible. This project focuses only on the preservation of the eBooks.

The Project Gutenberg Literary Archive Foundation (PGLAF) is a recognised charitable organization by the US Internal Revenue Service.

[5] Paid staff are financed through donations. http://www.gutenberg.net/donate.

[6] Radio Gutenberg http://www.gutenberg.net/audio/.

[7] Gutenberg Music http://www.gutenberg.net/music/.

Chapter 5: Details and circumstances of the Interviews

Michael Hart, Founder and Director of Project Gutenberg and Dr. Greg Newby, CEO of the Project Gutenberg Literary Archive Foundation completed the questionnaire and participated in email communications between March and April 2004.

Chapter 6: Analysis

This section presents an analysis of the data collected during the case study. It is organised to mirror the sequence of topics in the questionnaire.

- Perception and Awareness of Digital Preservation
- Preservation Activity
- Compliance Monitoring
- Digital Preservation Costs
- Future Outlook

Perception and Awareness of Digital Preservation

Project Gutenberg is not only one of the earliest web sites on the internet but it is also one of the earliest digital libraries in existence. They have been active in creating eBooks for over thirty years and are aware of the social benefits to be gained through preserving these resources for public access. Project Gutenberg ensures that all eBooks are available in plain text and other open formats to avoid obsolescence. The eBooks are uploaded to two main servers[8] and can then be mirrored by over thirty sites worldwide. The combination of open formats and many copies should ensure that access to these digitised literary works is preserved for the long-term.

The Main Problems

The major long-term problem lies in ensuring that copyright laws are respected for all of the digitised works made accessible by Project Gutenberg. Mirror sites exist in many countries around the world and, as such, ensuring that copyright laws are respected in each can be difficult. However, no eBook will be posted to the main site in the U.S. without gaining copyright clearance. Recent extensions to copyright laws in the U.S. and Europe have presented new challenges for the Project Gutenberg team. This is because no new works will be released to the public domain until 2018. Hart believes that these extensions to copyright laws benefit ‘very few copyright holders at the expense of universal access to literature and knowledge’.[9] These changes will impact the amount of research that needs to be done before an eBook can be digitised and made available.

Asset Value and Risk Exposure

Project Gutenberg exists to make literature and reference materials freely accessible to the general public in a digitised format. As mentioned above, Michael Hart believes that free access to literary works is vital for enabling the sharing of knowledge, art, music and culture. As such there is no inherent financial risk associated with the loss of the material, however, the significance of the collection has a value that should not be underestimated and would be a great loss.

[8] The two main servers are located at ibiblio: the public’s library and digital archive (ftp.ibiblio.org) and the Internet Archive (ftp.archive.org).

[9] From an interview with Michael Hart: The Second Gutenberg http://promo.net/pg/upi_interview_05_02.html.

Regulatory Environment

Project Gutenberg must adhere to U.S. laws involving operation as a not-for-profit corporation. However, these regulations are not sector specific. Project Gutenberg must be exceedingly careful to respect U.S. copyright laws regarding the works that they digitise and make available over the Internet. However, once a publication has been verified as being in the public domain, there are no other legal restrictions affecting Project Gutenberg.

Preservation Activity

Policies and Strategies

Project Gutenberg scans literary works and employs OCR technology to create eBooks. In some cases, eBooks are typed in by hand. The eBooks are then edited by a team of volunteer proof-readers. There are procedures and guidelines available online for volunteers to consult when scanning and editing texts for Project Gutenberg to ensure that all eBooks follow a standard format. Once the eBook has been produced, it is uploaded to two main servers. The eBook is made accessible via the official Project Gutenberg website, the Internet Archive site and over thirty mirror sites around the world. As there are no access or distribution issues, Project Gutenberg encourages users to save copies of the eBooks to CD or DVD.

Project Gutenberg believes that by generating a multitude of versions – those stored on the main servers, on local servers (through mirror sites) and those downloaded to CD and DVD – will ensure that the bit stream of the literary work is preserved for access. This embodies the philosophy of the LOCKSS strategy. LOCKSS 'uses the caching technology of the web to collect pages of journals as they are published, allowing libraries to take physical custody of selected electronic titles they purchase'[10]. LOCKSS was inspired by the words of Thomas Jefferson who said "let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident". [11]

Selection

Project Gutenberg aims to make digitised versions of popular literature and reference materials in the public domain freely accessible to the general public. As copyright expires, publications can be freely replicated and distributed. Many of these works are out of print. By digitising the out of print works, Project Gutenberg feels that they are saving the publications from 'obscurity and ultimate oblivion'.[12] Basically, all of the texts can be classified into three categories: light literature (such as *Alice in Wonderland),* heavy literature (such as Shakespeare and Dante) and references (such as Roget's Thesaurus). Mathematical and scientific works are also made available including *The Human Genome*. There are no real restrictions to what Project Gutenberg will make accessible. As long as the material is in the public domain, they can be digitised and submitted to Project Gutenberg. However, Project Gutenberg aims to benefit the widest possible audience and therefore prioritise the

[10] http://lockss.stanford.edu/projectdescbrief.htm.

[11] Jefferson, Thomas. [1791] 1984. Thomas Jefferson to Ebenezer Hazard, Philadelphia, February 18, 1791. In Thomas Jefferson: Writings: Autobiography, Notes on the State of Virginia, Public and Private Papers, Addresses, Letters, edited by Merrill D. Peterson. New York: Library of America (taken from LOCKSS website http://lockss.stanford.edu/)

[12] From an interview with Michael Hart: The Second Gutenberg http://promo.net/pg/upi_interview_05_02.html.

digitisation of popular literature and reference materials rather than extremely specialised works. Project Gutenberg already have texts in over thirty-one languages and are especially keen to increase their multilingual holdings.

Preservation

Project Gutenberg already has numerous plain text files that are twenty to thirty years old. In that time, many file formats have come and gone while plain text is still readable on virtually all computers. The use of plain text will also help to insure against future obsolescence. All Project Gutenberg eBooks are created as plain ASCII text files. This means that people with 'Apples and Ataris all the way to the old homebrew Z80 computers'[13] as well as Mac and UNIX users are all able to read the text files. Any open format can be submitted but the Project Gutenberg team will also generate plain ASCII[14] text files. Project Gutenberg encourages users to creat new formats from the plain text files to suit their individual needs. Once the eBook has been generated and edited by volunteers, it is uploaded to two main servers.

Project Gutenberg uses the unique eBook number as the file name. Therefore, if the eBook is the 10001 plain text file created, it will be named 10001.txt. Project Gutenberg will accept as many open file formats as volunteers are willing to submit, but will also generate a plain text version. Additional versions in other formats will be named accordingly but with different file extensions (e.g., html, pdf, xml). Each eBook has its own subdirectory that contains all versions of the eBook.

Project Gutenberg has volunteers representing a wide range of sectors (cultural heritage, government and higher education). Through these affiliations, they keep up to date with digital preservation developments. Project Gutenberg staff have ties with many organisational leaders and informal collaborations on best practices are common.

Access

The eBooks are catalogued by Project Gutenberg volunteers to include the author, the author's dates of birth and death, language, eBook number, and the Library of Congress classification to enhance online searching capabilities. As the publications that Project Gutenberg aims to make accessible are already in the public domain, restricting access is not really an issue. Project Gutenberg is mirrored in over thirty sites around the world. As such, they cannot accurately estimate the number of downloads that take place across all of the mirrored sites, but state that the equivalent of one million eBooks are downloaded each month from the main central server[15]. In an effort to increase accessibility by non-English users, eBooks can be generated and submitted in any language.

Project Gutenberg uses Dublin Core to describe their electronic resources to enable resource discovery.

Compliance Monitoring

There are no external requirements that Project Gutenberg must meet. However, Distributed Proof-readers[16] work to edit and ensure that the eBook content is as

[13] http://promo.net/pg/history.html.

[14] American Standard Code for Information Interchange (ASCII)

[15] From an interview with Michael Hart: The Second Gutenberg http://promo.net/pg/upi_interview_05_02.html.

[16] Distributed Proofreaders http://www.distributedproofreaders.net/c/default.php.

accurate as possible. The eBook goes through two rounds of proofreading where it may be examined by hundreds of volunteers. Once the eBook has been proofread, it goes to the post-processing stage. ‘The ultimate goal of post-processing is to create a plain text eBook with consistent formatting throughout, which contains as few errors as possible, and which accurately reflects the intentions of the author.’[17] Project Gutenberg citations - for example in the Online Computer Library Center (OCLC)- appear as their own editions and, as such, do not comply with any particular paper edition. In some cases Project Gutenberg editions are listed as the only edition in existence.

Digital Preservation Costs

A registered charity, Project Gutenberg relies on donations to pay their few dedicated staff members and for operational costs. Nearly 100 percent of the operational budget is focused on preservation. In terms of storage costs, the project founder believes that as disk drives become larger and cheaper, the price of putting eBooks on computers will become negligible [18].

Future Outlook

Project Gutenberg has already been implemented in Australia and Europe. Project Gutenberg of Canada is being founded in the near future. Project Gutenberg also hopes ‘to also create such projects in Africa, Asia, and other regions. In particular, they hope to create projects by which e-books can reach the masses via digital radio links to solar-powered PDAs. In addition, Project Gutenberg will be adding more multimedia e-books: paintings, sculptures, music, audio e-books, movies, etc., along with a wider variety of text formats.’[19]

Project Gutenberg will continue digitising literary works and aims to offer over 10,000,000 eBooks in over 100 languages by the time they celebrate their 50th anniversary in 2021. Project Gutenberg aims to enable the migration on request of their plain text files. This would mean that the plain text version could be generated in any type of file requested on the fly. This is currently in test mode. Project Gutenberg is also investigating creating the eBooks as born XML to allow easier creation of other formats on demand.[20]

[17] Post-proofing FAQ http://www.distributedproofreaders.net/c/faq/post_proof.php.

[18] From an interview with Michael Hart: The Second Gutenberg http://promo.net/pg/upi_interview_05_02.html.

[19] Michael Hart quoted in Project Gutenberg Progresses by Paula J. Hane, Information Today Volume 21 No. 5 http://www.infotoday.com/it/may04/hane1.shtml.

[20] Project Gutenberg Progresses by Paula J. Hane in Information Today Volume 21 No. 5 http://www.infotoday.com/it/may04/hane1.shtml.

Chapter 7: Conclusions

As the first and largest collection of eBooks, Project Gutenberg has been preserving electronic publications and making them accessible for over thirty years. By adhering to strict guidelines regarding the format of the eBook (plain text) for access and readability, Project Gutenberg has also ensured that their electronic resources can be preserved and migrated easily to other formats as needed. By uploading the eBooks to two main servers and by mirroring the Project Gutenberg database on sites around the world they ensure that backup versions of the eBooks are readily available if necessary. This multi-distributed approach is similar to the preservation strategy Lots of Copies Keeps Stuff Safe (LOCKSS) that is gaining worldwide interest. The combination of open formats and the proliferation of copies downloaded around the world should ensure that Project Gutenberg eBooks currently in existence and indeed any new eBooks created, are still accessible far into the future.

Appendix 1: Example Record for an eBook

Project Gutenberg Bibliographic Record

Data	
Title:	Of Human Bondage
Author:	Maugham, W. Somerset (William Somerset)
Language:	English
Subject:	Fiction
LoC Class:	Language and Literatures English literature
Release Date:	Oct 1995
Etext number:	351

Files

File Type	Download	File Size
Plain text	ibiblio.org select mirror P2P network	1.38 MB
Plain text (zipped)	ibiblio.org select mirror P2P network	579 KB

CONTACT DETAILS

ERPANET Coordinator
George Service House
11 University Gardens,
University of Glasgow
Glasgow, G12 8QQ,
Scotland

Tel: +44 141 330 4568
Fax: +44 141 330 3788
Coordinator@erpanet.org

ERPANET STAFF

directors
Seamus Ross, Principal Director
Niklaus Bütikofer, Co-Director
Mariella Guercio, Co-Director
Hans Hofman, Co-Director

coordinator
Peter McKinney

editors
Andreas Aschenbrenner
Georg Büchler
Joy Davidson
Prisca Giordani
Francesca Marini
Maureen Potter

www.erpanet.org

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK