Digitization and digital preservation projects at the German National Library of Medicine (ZB MED)

Michelle Lindlar
German National Library of Medicine
Cologne, Germany

Abstract
Preservation of the holdings and thus guaranteeing accessibility has always been a key task of libraries. Whereas preservation techniques for analogue materials have been researched for decades and largely been proven in best practices, the field of digital preservation is relatively new, yet the pace is much faster. The digital revolution offers libraries new opportunities as in the case of digitization to improve access. At the same time, we are faced with an ever-growing amount of digital data, which needs to be managed and preserved in a fast-paced context of changing technology. This article describes the German National Library of Medicine's (ZB MED www.zbmed.de) activities in the fields of digitization and digital preservation.

Key words: digital preservation; digitization; information management; workflow.

CC MED and catalogue enrichment

The digitization of tables of contents

The German National Library of Medicine (ZB MED) (1) looks back on ten years of experience in digitization. The digitization of tables of contents of more than 1,200 life science and biomedicine journals is the backend core component of Current Contents Medicine (CC MED) (2) CC MED is an ongoing service which started in 2000 to improve access to journal article references in the field of life sciences and biomedicine. The selection focuses on journals published in Germany or in German, which are mainly not indexed in Medline. Through a workflow which guarantees end-to-end quality control, via OCR the article data are automatically extracted from the scans and added to the medical information portal MEDPILOT (3).

Based on the experiences with this project in the past ten years, ZB MED will be launching CC GREEN, a similar project for nutritional, environmental and agricultural science journals, later this year. CC GREEN article data will respectively be added to GREENPILOT (4), the virtual library for nutrition, the environment and agriculture.

In September 2005, ZB MED, USB Köln (University and City Library of Cologne) and HBZ (Library Service Centre of North Rhine-Westphalia) partnered the "180T" project. Within four months the tables of contents of more than 180,000 books were scanned and, using a workflow which guarantees multi-level quality control, processed as full text using text recognition and directly fed into various catalogue systems. This Catalogue Enrichment generates a significant added value as the means for a more targeted literature search. It allows for direct access to the table of content via a URL to the PDF out of the catalogue record, thus creating a "browsing" or "preview" feature. Moreover, the full-text searchable table of contents allow for an improved quality of search results.

Even though the planned 180.000 mark has been reached, catalogue enrichment continues. ZB MED and USB Köln have been joined by additional partner libraries and in 2009 the HBZ union catalogue contained over half a million enriched titles.

Address for correspondence: Michelle Lindlar, Projektleitung Retrodigitalisierung / Langzeitarchivierung, ZB MED German National Library of Medicine, Gleueler Str. 60, 50931 Cologne, Germany
Tel: +492214787060, Fax: +492214787102, E-mail: lindlar@zbmed.de

In-house digitization

It is a given fact that digitization projects enable libraries to better meet the user's expectation in access. Nevertheless, it is also a given fact that the user's expectations in digital libraries are almost unrealistically high. When building digital libraries we need to be aware not only of the changing technology in the form of hard facts, but also of the "ways of the web" and users' needs and expectations. ZB MED is currently adapting a workflow for in-house digitization projects. In addition to questions regarding selection of material, scanning parameters and metadata, the question of the presentation of the works and the flexibility of the chosen platform plays a key role. Smaller boutique digitization projects, starting with a collection of anthropological works, will pave the way towards mass digitization.

Digital preservation

Whereas digitization is a form of preservation through a surrogate form where the analogue original still exists, libraries of today are faced with an ever growing amount of material that has no analogue counterpart: "digital-born material". Digital-born material is more and more an essential part of our holdings – most commonly in the form of e-journals, electronic thesis and dissertations and digital grey literature, but also in the form of non-textual materials such as audio-visual materials or even primary data such as simulations or databases. Recognizing the prevailing threat that digital-born data faces if no action is taken, the UNESCO General Conference adopted the *Charter on the Preservation of Digital Heritage* in 2003. The Charter defines "digital heritage" as "unique resources of human knowledge and expression". Access to these resources needs to be safeguarded by developing and implementing strategies: "Continuity of the digital heritage is fundamental. To preserve digital heritage, measures will need to be taken throughout the digital information life cycle, from creation to access. Long-term preservation of digital heritage begins with the design of reliable systems and procedures which will produce authentic and stable digital objects" (5).

Unlike its printed counterpart, digital information cannot be accessed directly and depends on three factors: firstly, the carrier material, such as a CD-ROM; secondly, a reading device, such as a CD-ROM drive; and lastly a software to render the information and make it readable, such as Adobe Acrobat Reader. Considering the various digital carrier materials in libraries' holdings – e.g. CD-ROMs, Audio CDs, DVDs and hard drives – and the multitude of different file formats and different versions of software with which these files were produced, the impact of digital preservation may become clear. Carrier material deteriorates, reading devices are no longer available due to obsolescence and software may no longer run on the current operational system.

The Goportis Digital Preservation Team

In 2009 Goportis (6), the Leibniz Library Network for Research Information consisting of the German National Library of Medicine (ZB MED), the German National Library of Science and Technology (TIB) and the German National Library of Economics (ZBW) declared digital preservation a joint task of utmost importance. The Goportis Network has committed itself to a sustainable approach to permanent accessibility of digital data by sharing resources, knowledge and experience. A Goportis digital preservation team was put into place to develop a joint digital preservation strategy. The digital preservation strategy will be stipulated in a Goportis Digital Preservation Policy.

Risk assessment and normalization

The preservation team is currently developing a risk assessment plan, which serves as a systematic guideline for the three institutions in conducting adequate and regular evaluations of their digital holdings. These evaluations cover aspects such as size of the collection by type (e.g. electronic thesis and dissertations), condition, file formats, hardware and software dependencies and suggested preservation measures. An adequate inventory of the digital holdings is the necessary basis for a decision regarding possible normalization of digital holdings. Whereas the risk assessment answers the question *What do we have in our holdings*, normalization addresses the question *What do we want to have in our holdings?* For instance, one form of normalization could require all electronic thesis and dissertations to be in PDF/A-1 format. PDF/A is an ISO standard file format, based on Adobe Systems Inc. PDF Reference Version 1.4. The benefit of PDF/A is that it strives for device independence and self documentation. This is achieved by embedding all information necessary to display the file the same way in years to come – such as for example fonts, colour information and graphics – in the file itself. However, self documentation also means that the file format has certain limits. Audio and video content is forbidden in a PDF/A file and, moreover, it is not permitted for a PDF/A document to be reliant on information from external sources, such as hyperlinks.

It thus becomes clear that we need to know the digital content that we hold and the structure and content of material which we want to acquire in great detail to be able to make decisions regarding possible normalization. Furthermore, a thorough understanding of benefits as well as drawbacks of file formats is also essential.

Preservation action

The tasks of risk assessment and decisions regarding normalizations show that digital preservation is as much an institutional and organizational question as it is a technical one. Nevertheless, technical measurements and procedures need to be implemented in order to preserve digital information and to guarantee accessibility to digital information over time.

Digital information needs to be *monitored* to measure the risk it faces. This happens through risk assessment, as mentioned above. Additionally, the process can be automated for larger amounts of files through a monitoring system. Action to be taken for files at risk can include (but are not limited to) "refreshing", "emulation" or "migration". Refreshing means copying files from one carrier, like a CD-ROM, to a new carrier of the same type. Emulation refers to a way to replicate an old system, like WordPerfect in an MS-DOS environment, within your current work environment, for example on a Windows XP machine. Migration means converting digital documents from a file format "at risk" to a different file format – like migrating Microsoft Word .doc files to PDF. Goportis is planning a joint proof of concept implementation. The proof of concept will include the integration of existing workflows for digital materials in a preservation system. An example for this is a workflow for electronic thesis and dissertations, where bibliographic metadata from the catalogue system will be used by the preservation system in addition to technical metadata. Along with the access copy for the user and a digital preservation master file all metadata will be stored as a logical unit. Monitoring systems as described above will be tested on normalized as well as complex materials, such as audio-visual digital information.

Conclusion

The fast paced changes in technology is best described by the following quote by digital preservation expert Jeff Rothenberg: "Digital information lasts forever – or five years, whichever comes first" (7). The complexity of guaranteeing long-term accessibility to digital data requires a commitment to a constant technology watch. We need to be aware of changes in dependencies such as file formats, hardware, but also trends in scientific communication and publications to be able to handle our digital holdings now and in the future. At the same time, action needs to be taken as soon as possible to prevent the loss of digital information that we already store within our holdings. The German National Library of Medicine meets this challenge through digitization as well as digital preservation projects.

Submitted 12.03.2010 *Accepted* 19.03.2010

References

1. ZB MED. German National Library of Medicine [Internet]. 2010 March 12. Available from: http://www.zbmed.de
2. ZB MED. CC MED – Current Contents Medicine [Internet]. 2010 February 02. Available from: http://www.zbmed.de/projekt_ccmed.html
3. ZB MED. MEDPILOT [Internet]. 2010 March. Available from: http://www.medpilot.de
4. ZB MED. GREENPILOT [Internet]. 2010 March. Available from: http://www.greenpilot.de
5. UNESCO. Charter on the Preservation of Digital Heritage [Internet]. 2003 October 15. Article 5. Available from: http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html
6. GOPORTIS. Leibniz Network for Research Information [Internet]. 2010 March. Available from: http://www.goportis.de
7. Rothenberg, J. Ensuring the Longevity of Digital Documents [Internet]. 1999 February 22. p.2 Available from: http://www.clir.org/pubs/archives/ensuring.pdf

Further Reading

Arms CR. and Fleischhauer C. Sustainability of digital formats: planning for Library of Congress collections [Internet]. 2007 May 21. Available from: http://www.digitalpreservation.gov/formats/index.shtml

Dempsey L, Lavoie B. Thirteen ways of looking at... digital preservation. D-Lib Magazine [Internet]. 2004 July/August 10:(7/8). Available from: http://dlib.org/dlib/july04/lavoie/07lavoie.html

Digital Preservation Europe. What is digital preservation? [Internet]. 2006 April 28. Available from: http://www.digitalpreservationeurope.eu/what-is-digital-preservation/

Rothenberg J. Jeff Rothenberg's digital longevity papers [Internet]. 2009 November 17. Available from: http://www.panix.com/~jeffr/Prof/digilong.html