LONG-TERM PRESERVATION AND WEB ARCHIVING – ACTIVITIES OF THE AUSTRIAN NATIONAL LIBRARY

Mag. Bettina Kann
Österreichische Nationalbibliothek
Digital Library Main Department - Head of Department
Long Term Preservation - Head of Department
Email: bettina.kann@onb.ac.at

Abstract

The collection and preservation of as much as possible of a country's published work plays a significant part in preserving a country's cultural memory. For this reason, almost every country in the world has formal legal requirements for publications to be supplied to their National Library ("legal deposit"), in order that this important contribution to our cultural and scientific heritage be preserved for as long as possible. It would make sense for this library copy requirement to encompass all forms of media and publication that were of importance in their time. Nowadays this includes an ever-growing number of online publications on the World Wide Web.

Key Words: *digitations, web archiving, Austrian National Library.*

The signs from the European Commission were clear: steps needed be taken in the field of archiving our digital heritage (1). What was expected of the member states was set out clearly by the European Council and the measures were linked to a concrete timetable, which, nonetheless, placed economic considerations to the fore. (2)

In Austria the obligation to offer and, if requested, to deliver library copies is regulated by the Media Act of 1981 (§ 43 et seq.). In the 2000 amendment to the Act (3) this legal deposit, which had until then been limited to print media products, was extended to include "other media products" (with the exception of audio-visual products). This was a response by the legislators to an urgent concern expressed by the libraries. Nevertheless, this amendment from the year 2000 only dealt with delivery of so-called offline publications (i.e. publications which are recorded on physical media, such as CD-ROMs, DVDs and similar).

This means that until recently none of the already numerous purely online publications had been deposited with either the Austrian National Library or any other library in Austria, meaning that no provision had been made with regard to long-term preservation and accessibility.

For this reason, the Austrian National Library participated in a working party moderated by the Federal Chancellery, which had as its goal the writing of an appropriate amendment to the Media Act. Together with representatives of the media industry and important associations they discussed not only the means of delivery, but also the subsequent use of the collected media products, all of which was put into a draft amendment by the Federal Chancellery. This draft amendment was the basis for the Amendment to the Media Act which came into force on 1st March 2009, whereby the Austrian National Library is empowered to collect online publications and build up an archive of Austrian websites. (4)

Until this point in time, the National Library was only able to collect, archive and make available online publications as a result of voluntary negotiations with the media owners. Working within these limits the Austrian National Library had already been collecting online publications from selected institutes and publishers since 2004.

The sheer number of digitally produced publications, their ephemeral nature, and the difficulty in delineating the resources to be archived because we are

dealing with dynamic and interactive forms of publication, all necessitate the drawing up of specific guidelines for the collection of online publications. The speed with which developments occur in the field of electronics mean that these guidelines will need to be reviewed on a regular basis and adapted to take account of new developments at the time.

An important criterion is the so-called *"Austriacum"*, which means that the publication must either be published in Austria or hosted on an Austrian server, or that it must have a demonstrable relevance to Austria. This framework is set out explicitly in the Media Act.

Examples of the online publications which the National Library will have the task of collecting are: e-journals, electronic dictionaries, e-books, e-prints, websites, electronic university theses, digital works by living authors and posthumous works, e.g. electronic manuscripts, preliminary drafts of literary works, private e-mail correspondence etc.

The following resources will not generally be collected and archived:

- Directories (lists of links), discussion lists, news groups and similar, application programmes (software), games, advertisements, event calendars.
- Personal home pages will only be collected in exceptional cases, and then only in the case of persons who are of public interest (e.g. authors).
- With the exception of websites, online publications will be archived in the purpose bought Exlibris DigiTool system, which conforms to the OAIS model (5). With regard to the format, the Austrian National Library prefers PDF/A or XML (with the associated DTDs and style sheets).

Web Archiving

With its web archiving the Austrian National Library has as its primary objective the collection and preservation of a significant proportion of the national

web space. As a valuable part of our cultural heritage this extensive content from the World Wide Web should continue to be available to interested users and scientists in the future, long after it has disappeared from the web.

The complex task of collecting the data will be carried out using a combination of different collection methods.

- *Domain Harvesting*

  Collecting everything which appears on one domain, such as, for example, the Austrian .at domain, is called domain harvesting. Using a compete list of all the sites registered at the domain, *nic.at*, appropriate software will be used to harvest and save all the .at websites (6). In addition, websites from other top-level domains will be harvested if they can be shown to have relevance to Austria. The selection of sites not belonging to the .at domain will be largely a manual, and therefore very time-consuming, process. For this reason, the Austrian National Library is working on developing automated processes to identify sites from outside the .at domain which have relevance to Austria. The massive data volumes (several terabytes) and the throughput times involved (estimated at several months for the .at domain), mean that domain harvesting can only be carried out within limits. For this reason we cannot aim to collect everything through domain harvesting, merely a representative cross section of what was in publication at that point in time.

- *Selected Harvesting*

  Since domain harvesting will be carried out infrequently, much content will be lost, particularly on dynamic websites which are frequently updated. For this reason, web curators will select important websites from particular areas, such as media, science, authorities etc., for which they will set appropriate harvesting intervals. So, for example, the websites of daily newspapers could be saved on a daily basis, so that all important content can be archived.

- *Event Harvesting*

  A special form of selected harvesting is event harvesting, whereby content relating to particular events can be archived. Classic topics for event harvesting would be, for example, elections or sporting events (e.g. EURO 2008™). Numerous websites are only available for the duration of the event, and event harvesting can therefore be regarded as a valuable addition to domain and selected harvesting. In any case, when you consider that the average life of a website is 44 days, it is clear that there is always a risk that a website will have disappeared before the next "routine" harvesting occurs.

The Austrian National Library is using a combination of all three strategies in order to preserve as complete and expressive a reflection of Austrian web space as possible. A few event harvestings have already been carried out: EURO 2008™, and both the National and European elections in 2008. Taking into account the so-called deduplication (files which are already available will not be saved again, merely referenced), 31 million files with a total size of 350GB were collected as a result of these harvestings. The first Austrian domain harvesting was begun in September 2009, and the first throughput, with a limitation of 10 MB per website, was completed at the end of December. Websites which have not yet been harvested will be harvested in subsequent harvestings with higher limits. The first throughput has so far resulted in the collection of 895,445 domains, with a total size of 1.4 terabytes and approximately 78 million files.

Image 1: The websites of the political parties during the National Council elections in 2008

To carry out its web archiving activities the Austrian National Library uses the NetArchive Suite developed in Denmark, and Wayback Machine developed by Internet Archive. The Austrian National Library's membership of the *International Internet Preservation Consortium (IIPC)* (7) enables it to benefit from a worldwide exchange with institutions who are leading the way in web archiving, as well as to participation in numerous working groups and projects.

Digitisation of analogue sound carriers and digital long-term preservation

The Austrian National Library owns a collection of approximately 22,000 analogue sound documents, with a total playing time of around 30,000 hours, which is seriously endangered by the fragility of the carrier materials. As well as a small number of wax cylinders and self-cut discs, the collection includes a large number of shellacs and vinyl discs. The vast majority, however, comprises tapes and cassettes, which contain live-recordings from concerts (from events organised

by the Austrian National Library), recordings of interviews and speeches by writers, and field recordings by folk-music researchers. Many of these recordings are unique copies and count as part of Austria's cultural heritage.

Since 2007 the Austrian National Library has been carrying out a digitisation programme, planned to run over many years, to preserve the information on these endangered sound carriers. The programme's objective is the long-term preservation of digital data, since it will, in the foreseeable future, become impossible to play the analogue sound carriers. The programme was planned by the Austrian National Library in conjunction with the Phonogrammarchiv of the Austrian Academy of Sciences, and the Austrian Media Centre (Österreichische Mediathek). The wax cylinders were digitised by the Phonogrammarchiv, and the tapes are currently being digitised by the Austrian Media Centre. They are being digitised into 24bit/96khz Broadcast Wave format. From these master files, which are ideal for long-term preservation, mp3 files (128kBit/s, 48kH) are generated. On legal grounds, however, these can only be listened to inside the Austrian National Library. In addition to these digitised files, the corresponding technical and administrative metadata (including those which document the digitisation process itself) are also being stored.

Strategies and Services for digital long-term preservation

In discussing digital long-term preservation, I must tackle the question of the long-term usability of the collected material. There seem, at present, to be two main strategies which enable us to guarantee this: migration and emulation, or a combination of the two as the case may be. The Austrian National Library tends to favour migration, which means the wholesale copying from one data carrier to another, including, probably at the same time, migration from one data format into another. What is particularly important in this is the so-called "technology watch", namely recognising at what point in time a particular technology becomes obsolete and must be replaced with another.

The advantages of migration:

- Libraries can draw on significant past experience.
- The archived documents are always up-to-date and can be used with current applications programmes.
- Repeated migration means that the archived objects are permanently subject to quality control.

The risks of migration:

- Conversion tools usually have to be created for each migration project, since the source and target formats and their compatibility is constantly changing.
- There is a real danger of data being lost or corrupted.
- Migration is expensive.

It is clear that the not all institutions can afford the development expense involved in implementing controlled and structured migration. This is one of the reasons why the Austria National Library has become a partner in the EU funded *PLANETS* project.

The aim of the *PLANETS* (8) (Preservation and Long-term Access through Networked Services, running until May 2010) project is to develop services and tools which will support institutions in assuring long-term access to digital, cultural and scientific resources.

The project will enable organisations to make informed decisions about long-term preservation strategies and reduce the costs of preservation actions through increased automation and scaleable infrastructure.

The individual projects include:

- "Preservation Planning" services that empower organisations to define, evaluate, and execute plans for long-term preservation
- "Characterisation" services and tools for the automatic and technical characterisation of digital objects

- "Preservation Actions" services and tools for the transformation or emulation of digital objects.
- An "Interoperability Framework" to integrate tools and services in a distributed service network
- A "Testbed" to provide a consistent basis for the objective evaluation of the different tools and services.
- A comprehensive programme to disseminate the results of the projects.

The Austrian National Library is leading the "Testbed" subproject and co-ordinating the development of software and hardware environments for the evaluation of tools and services for digital long-term preservation.

References

1. EU Commission "On the digitisation and online accessibility of cultural material and digital preservation" dated 24.8.2006: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:236:0028:01:EN:HTML
2. "Council Conclusions on the Digitisation and Online Accessibility of Cultural Material, and Digital Preservation", 30.10.2006: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52006XG1207(01):EN:HTML
3. BGBl. I, 75/2000
4. BGBl. I, 8/2009
5. http://public.ccsds.org/publications/archive/650x0b1.pdf
6. Situation as at 28. Dec. 2009: ca. 908.596 registered .at Domains, source: http//www.nic.at
7. http://www.netpreserve.org
8. http://www.planets-project.eu