LIBER QUARTERLY 21 (2), January 2012 — ISSN: 1435-5205. P213–225
http://liber.library.uu.nl/
Igitur publishing

Putting 600,000 Books Online: the Large-Scale Digitisation Partnership between the Austrian National Library and Google

Max Kaiser

Austrian National Library,
www.onb.ac.at,
max.kaiser@onb.ac.at,
Twitter: @maxkaiser

Abstract

In a public-private partnership with Google, the Austrian National Library is digitising its historical book holdings. Some 600,000 volumes from the sixteenth to the nineteenth centuries will be digitised and made available free of charge. The project demonstrates that public-private partnerships can be successful in enabling our heritage institutions to provide large-scale access to their holdings, provided that such partnerships are not exclusive and free access is ensured. The article outlines the preparatory phase and work flows established in the project.

Key Words: digitisation; Austrian National Library; Google; Google Books; public-private partnership

In cooperation with Google the Austrian National Library will digitise and put online all of its historical book holdings. The Austrian National Library is the central academic library in Austria with a history going back to the fourteenth century. The historical book holdings rank among the world's most eminent collections on account of the library's history as former court library of the Habsburg Empire. In the sixteenth century the earliest regulations were promulgated for cost-free deposit of books at the library to expand the holdings. These legal deposit regulations were gradually extended until the beginning of the nineteenth century. The collection

includes about 44,000 books from the sixteenth century and the 15,000 volumes of the famous library of Prince Eugene of Savoy. Apart from one of the largest German-language historical book collections, the Austrian National Library also owns important and large holdings from eastern and south-eastern Europe.

For many years the Austrian National Library has been carrying out ambitious digitisation projects and it has a steadily growing offer of digital services. For example, analogue reproduction services were replaced by digitisation on demand. Among the digital library services are portals for digitised historical newspapers (ANNO, to date around 6.5 million pages)[1] and legal texts (ALEX, to date around 3 million pages)[2], as well as a digital platform of the library's Picture Archives and Graphics Collections (Bildarchiv Austria)[3]. In cooperation with the Austrian Media Centre (Österreichische Mediathek) and the Phonogram Archive of the Austrian Academy of Sciences, the Austrian National Library has been digitising its analogue sound collections for several years. The legal obligations of the Austrian National Library include the collection of online publications and archiving the Austrian web space.

As the digital collections grow rapidly, the Austrian National Library must devise adequate strategies for long-term preservation in order to ensure access by future generations to these holdings. In 2008 a department for digital preservation was established which is responsible for the operational aspects of this complex challenge. The Austrian National Library is also actively engaged in EU-funded digital preservation research projects.

The partnership with Google — to date the largest public-private partnership in Austria's cultural sector — will take the Austrian National Library a big step closer to the strategic goal of comprehensive digitisation of its historical holdings. In the framework of *Austrian Books Online*[4] approximately 600,000 volumes — all in the public domain — from the beginning of the sixteenth to the second half of the nineteenth century will be digitised in the years to come years, totalling about 200 million pages. These will be available free of charge via *Google Books* and the Digital Library of the Austrian National Library.

Google Books

Two sources are feeding into *Google Books*, which was initiated in 2004. In the so-called 'partner programme', publishers provide their books to Google to digitise and make them available online. In the 'library programme' Google is currently cooperating with about 40 library partners throughout the world to digitise their books; among them are thirteen libraries in Europe.[5] In cooperating with Google, the Austrian National Library is following the example of renowned libraries such as the university libraries of Harvard, Michigan, Stanford and Oxford, and the New York Public Library, which have been working with Google since 2004. Shortly before the conclusion of the Austrian National Library's contract with Google, in March 2010, the Italian Cultural Ministry announced a cooperative agreement with Google in which the Italian National Libraries of Rome and Florence will participate. There were also announcements of partnerships with Google by the National Library of the Netherlands (Summer 2010), by the Czech National Library (Spring 2011), and by the British Library (June 2011). Besides the Austrian National Library, in the German-speaking countries the Bavarian State Library has signed a contract with Google, in Spring 2007. In that partnership, more than 500,000 books have already been digitised so far.

As a result of partnerships with publishers and libraries more than 15 million digitised books can now be searched and found via *Google Books* (http://books.google.com). Of those, about three million volumes are in the public domain. Unlike in the United States, Google digitises exclusively public domain works in Europe.

In December 2010 Google set up a sales and download platform for eBooks. In the 'Google eBookstore' around three million books are available to date, among those the public domain works from the library programme, which can be downloaded free of charge. In the USA more than 200,000 books from publishing partnerships are on offer for sale only. Google eBooks can be read in a web browser or on eBook readers. The digitised books of the Austrian National Library will become part of Google's free-of-charge eBook service.

Cornerstones of the Partnership

A project the size of *Austrian Books Online* requires a financial framework which would have been hard to accomplish for the Austrian National Library without a partner like Google. The costs of full-text digitisation of books are enormous.[6] More than half a million books are envisaged to be included in the project; without a partner like Google, such an effort would have been scarcely affordable.

While Google finances full-text digitisation, book transport and insurance, the Austrian National Library bears the costs of selection, preparation and re-shelving of the books. Other significant costs for the library include the necessary updates of metadata, quality control of the digitised items, data storage, and making the digitised items available through the Digital Library.

Apart from the costs the time factor plays an important role: until recently, digitisation projects in libraries typically accounted for not more than 5,000 to 10,000 digitised volumes per year. Consequently a project of the scope of *Austrian Books Online* would have lasted several decades.

In order to achieve the aim of comprehensive digitisation of the Austrian National Library's historical book holdings, the public-private partnership with Google suggested itself, as Google not only disposes of the necessary resources but also has many years of practical experience in mass digitisation. For both partners the project represents a 'win-win' situation: As a leading search enterprise Google supports the Austrian National Library's goal of making its historical holdings available online to a worldwide audience. On the other hand, through this partnership, Google obtains relevant multilingual content which contributes to its goal of making all books of the world findable and searchable.

Apart from improving the accessibility of the books, digitisation makes an important contribution to book conservation, because in future the originals will not have to be used as often. On top of that, digitisation is part of a preservation strategy for the holdings in case of a disaster. In this context the fire in the Viennese Hofburg in 1992 should be remembered, to which the Austrian National Library's Hall of State almost fell victim, and also the fire disaster in the Duchess Anna Amalia Library in Weimar in 2004.

The Austrian National Library considers its decision to enter into a public-private partnership (PPP) justified by 'The New Renaissance', a report which was published in January 2011 by a high-level expert group of the European Commission (the Comité des Sages)[7] and which has since provoked a lot of discussion. The report deals with the fundamental significance of digitisation for the democratisation of access to knowledge and culture. While the EU member states have declared themselves in favour of making European cultural heritage accessible through the joint European platform *Europeana*, which provides access to holdings from libraries, museums and archives, there has been limited availability of public funding for the large-scale digitisation projects that are necessary. The report, considering this situation and the high costs of digitisation, sees in PPP models an essential complement to funding provided by the public sector. However, the report stresses that free public access to the digitised items must be secured and that such partnerships should not be exclusive. The key suggestions by the Comité des Sages have recently been taken up by the European Commission in their updated Recommendation on the digitisation and online accessibility of cultural material and digital preservation[8] published in October 2011 which contains an annex with key principles for public-private partnerships.

Both the suggestions of the Comité des Sages and the recommendations by the European Commission are in line with the cornerstones of the agreement between the Austrian National Library and Google. In defining the general framework of the partnership it was essential for the library to clearly define its goals and non-goals. The most important points were already agreed by both partners in the early stages of their talks:

- Only public domain material may be digitised.
- Cooperation with Google is non-exclusive. The library is free at any time to digitise the same holdings with other partners.
- The library receives copies of all digitised items and can make them available online for non-commercial use.
- Both partners are obliged to make all of the digitised items available for online access free of charge. This obligation exceeds the duration of the partnership.
- The Austrian National Library can make their digitised items available through other platforms such as *Europeana* and provide them to research partners.

- The library is fully autonomous in decisions regarding which books are to be digitised in the framework of the project.
- The logistics and digitisation processes are supervised and evaluated by the library's Conservation Institute.
- The library can terminate the partnership in case it does not meet its expectations.

Transparent communication to the public of the key points of the public-private partnership proved to be essential. Accompanying the press conference to announce the project in June 2010 the Austrian National Library put online comprehensive 'Frequently Asked Questions' with regards to the main aspects of the partnership. Those FAQs, made available in German and English, are constantly being updated.[9] In addition to 'classical' outreach work, social media channels such as Twitter are also included in the communication strategy.[10] *Austrian Books Online* enjoys a lot of attention in the media and among the public and has been received almost entirely positively, which can be seen as a confirmation of the project's successful positioning as a model case of a public-private partnership.[11]

Project Preparation

In addition to the historical book holdings of the general collection, including the volumes in the library's State Hall, other holdings will be digitised as well: the public domain books of the Department of Maps, the Department of Rare Books and Manuscripts, the Department of Music, and of the library of the Theatre Museum. The book holdings of the Fidei Commiss Library, i.e., the former private library of the House of Habsburg-Lothringen, are also part of the project. While there are metadata records for all the other holdings, the latter collection will be catalogued in its entirety for the first time as part of *Austrian Books Online*.

The project is divided into seven work packages that cover key processes: book logistics, metadata and catalogues, conservation and restoration, data download and quality control, online access, IT infrastructure, and project management. To date about seventy staff members are involved in the project, of whom about twenty work exclusively for *Austrian Books Online*, deal-

ing with book logistics, book preparation, restoration, data download and quality control, software development and project management.

A project the size of *Austrian Books Online* requires considerable preparation. From June to December 2010 a preparatory project laid the ground work for the operational project phase. The project in one way or another concerns every department of the library, hence it was important to integrate it into the organisational processes and to interlink it with other running projects. It was necessary to plan in detail the required personnel resources and to carry out the necessary organisational changes in advance.

In order to secure broad acceptance of the required organisational changes, effective internal communication of the vision and goals of the project was essential. Several departments had to re-evaluate their workflows and provide resources to the project, and the priority of some internal projects had to be re-assessed. Another important part of the preparatory project phase included consultation and coordination with other library partners of the *Google Books* project in Europe and the USA in order to benefit from their experiences.

A lot of attention was paid to implementing the logistics workflows for pulling books and preparing them for digitisation. For cost and efficiency reasons it is not possible to select books individually. Thus, literally, books are being digitised shelf by shelf. Books are only excluded from the project because of their (large) format, their state of conservation, their deviating format (e.g., folded maps), or their special value.

Project Implementation

After a successful test delivery to Google and a first data download at the end of 2010 the massive pulling of books was begun, and finally in Spring 2011 digitisation commenced. The first digital books are already available via *Google Books.*

The logistics are complex because huge quantities of books have to be moved. A particular challenge is the baroque State Hall: the historical book shelves

include a gallery, and many books are stored in up to three rows behind one another. A hydraulic lift is being used to enable pulling those books.

Preparation of the books for digitisation is carried out in a newly adapted-preparation area in the library. The books are checked by members of the processing team and are supplied with barcodes. Barcodes are necessary on the one hand for tracking each volume in the logistics and digitisation workflows and on the other hand for linking the resulting digital items with the metadata records in the electronic catalogue. For each book certain updates of the metadata record are necessary. Amongst other things each barcode has to be linked to a metadata record and for multi-volume works separate metadata records have to be created for each of the volumes. Finally each book has to be checked out from the integrated library system before being sent to the digitisation centre.

For carrying out these essential preparatory tasks, the processing team has eight minutes per volume. More complex cataloguing tasks which cannot be carried out in the timeframe available are dealt with by a separate team. This applies to cases where several works are bound together in one volume — which might affect up to a couple of dozen different works. In order to prepare these volumes for digitisation it is necessary to identify each work, create metadata records for each work and link each of them to a barcode. This is a complex and time-consuming process.

The 100,000 volumes of the Fidei Commiss Library constitute a special case. First they are being catalogued by a team of four staff members. Cataloguing these holdings, which is a prerequisite for digitisation, has the additional benefit that they will now be searchable in the electronic catalogue for the first time.

As part of the preparatory process the books are checked for their conservational condition and then cleared for scanning by members of the Institute for Conservation. When necessary, certain book protection measures are taken by the team, such as consolidating the book spine, protecting the binding or fixing loose pages.

Scanning takes place in a Google digitisation centre in Germany. The frequency of deliveries is planned in advance. The procedures for book transportation, storage and digitisation were agreed with the Institute for Conservation and

the Austrian Federal Office for Monuments. As each book shipment sent to the digitisation centre in Germany constitutes an exportation of national cultural possessions abroad, all shipments must be permitted in advance by the Austrian Federal Office for Monuments.

After returning from the digitisation centre each volume is inspected separately, checked into the integrated library system and finally returned to the book shelves. On average, each book which is being digitised as part of *Austrian Books Online* is unavailable to users for three months.

Digital Items and the Digital Library

The project also has a complex IT component — up to 95,000 digitised items per day need to be downloaded from Google and processed in an automated way. ADOCO (Austrian Books Online Download and Control), an application developed by the Austrian National Library, is used for data download, control and management. Data are provided by Google via a machine interface. The image and OCR files are of the same quality as those which are provided to users via *Google Books*.

Quality control is particularly demanding. Many of the libraries participating in *Google Books* forego quality checks altogether because of the quantity of data involved in this mass digitisation project. The Austrian National Library has decided to carry out quality control on the basis of automated jobs and representative sample checks. This requires a rearrangement of quality assurance processes. While quality control in previous digitisation projects has mainly been carried out manually, in this case the detection of individual errors is not the aim, but the IT-assisted discovery of clusters of likely systematic errors. Error candidates filtered out in this process can then be checked manually in a second step. For automated quality control the Austrian National Library can build on the prototype results of European research projects in the areas of digital preservation and mass digitisation.[12] To date there are few established practices, software solutions or best practice examples, so joint research projects will be essential.[13]

Storage and backup of the digital copies requires a significant extension of the Austrian National Library's mass storage system. A feasibility study was

carried out and it was decided not to outsource data storage but to build in-house storage capacity. The JPEG-2000 master files of the digital copies are stored redundantly; the access copies will be generated on-the-fly via an image content server.

In order to ensure citability and permanent identification of the digitised items a URN (Uniform Resource Name) from the namespace NBN (National Bibliography Number), which is administrated by the national libraries, will be generated for each book, respectively each title.[14]

Beginning in mid-2012 the Austrian National Library will gradually make the digitised copies available through its Digital Library. In a first implementation phase users will be able to find the digitised books through the Austrian Nation Library's 'Quick Search' service which has been available since Spring 2011. For accessing the digitised items a book viewer will be implemented which will provide functionalities like page turning and zooming. It will also be possible to download single or multiple pages or complete books as PDFs. In the second implementation phase in 2013 full-text search will also become available. The implementation of apps for mobile devices (e.g., iPhone, iPad, Android devices) is planned for 2013. The Austrian National Library will also make the digitised books available through other platforms like *Europeana* or the online portal of the European national libraries *The European Library*.

Outlook

In particular the availability of full-text search will provide new possibilities for research.[15] Users will no longer be limited to a search in the metadata records of the Austrian National Library's electronic catalogue, but will be able to search the tables of contents, indices or the full text of the books for relevant words or phrases. In this way they will discover works which they possibly would not have found otherwise.

In the medium term the Austrian National Library is planning to enrich the full texts with additional information and to offer additional services apart from mere full-text search. Using technologies like named entity recognition, for example, will allow identifying and annotating place names and person

names in texts and indexing them for search. These data can also be linked with ontologies, thesauri and lexica, or with other data sets available in the web (linked data).

Employing computer assisted methods to large full-text corpora like the ones of *Google Books* or of the Hathi Trust[16] Research Center, enables totally new research queries in the humanities and social sciences. Linguists, for example, now have access to a very large corpus of material for etymological analyses. Literary scholars can resort to new ways for researching literary influences; historians have access to new tools for analysing source material. An example of the emerging data-centric research paradigm is the 'Digging into Data Challenge' which was first announced in 2009 by a consortium of research funding agencies including NSF in the USA and JISC in the UK.[17]

In 2010, Google released their N-Gram Viewer[18] which allows carrying out frequency analyses of words or phrases within the text corpus of *Google Books*. First results of the new research method of 'Culturomics'[19] were published by Jean-Baptiste *et al*. in December 2010.[20]

It is to be expected that the data pool which is being set up from the historical book holding of the Austrian National Library will also be the basis for new source and data-centred research projects in the (digital) humanities and social sciences.

In the next few years *Austrian Books Online* will make one of the most important historical book holdings available to a worldwide public. The project will significantly improve the availability of the works, and thus make a contribution to the democratisation of knowledge and culture. Further information on *Austrian Books Online,* including detailed 'Frequently asked Questions', can be found at the project website (www.onb.ac.at/austrianbooksonline/), which is available in English as well.

Notes

[1] http://anno.onb.ac.at/.

[2] http://alex.onb.ac.at/.

[3] http://www.bildarchivaustria.at/.

[4] http://www.onb.ac.at/austrianbooksonline/.

[5] An up-to-date list of partners can be found in the 'Frequently Asked Questions' of Austrian Books Online: http://www.onb.ac.at/ev/austrianbooksonline/faq.htm.

[6] Cf. Nick Poole: The Costs of Digitising Europe's Cultural Heritage. A Report for the Comité des Sages of the European Commission. November 2010, http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/annexes/digiti_report.pdf.

[7] The New Renaissance. Report of the 'Comité des Sages'. Reflection Group on Bringing Europe's Heritage Online. January 2011, http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf.

[8] Commission Recommendation of 27.10.2011 on the digitisation and online accessibility of cultural material and digital preservation, http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/recom28nov_all_versions/en.pdf.

[9] http://www.onb.ac.at/austrianbooksonline/faq.htm.

[10] http://twitter.com/#!/ABooksOnline.

[11] A commentator in the Austrian daily paper *Der Standard* wrote: 'Partners like the Austrian National Library show that a sensible cooperation that protects the interests of both sides is possible. A library is only as useful as the access to its holdings, and that will be — thanks to Google — dramatically improved in the coming years.' (*Der Standard*, 17.6.2010, p. 12).

[12] Planets: http://www.planets-project.eu, IMPACT: http://www.impact-project.eu, SCAPE: http://www.scape-project.eu.

[13] See e.g. AQuA — Automated Quality Assurance Project: http://wiki.opf-labs.org/display/AQuA/Home.

[14] For URN NBNs in general see the reports of the persID initiative: http://www.persid.org/initiative.html.

[15] Compare Gregory Crane: What Do You Do with a Million Books? In: *D-Lib Magazine*, March 2006, http://www.dlib.org/dlib/march06/crane/03crane.html.

[16] http://www.hathitrust.org/.

[17] 'The idea behind the Digging into Data Challenge is to address how 'big data' changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials used by scholars in the humanities and social sciences — ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data or cell phone records — what new, computationally-based research methods might we apply?'

[18] http://www.diggingintodata.org/.http://ngrams.googlelabs.com/. N-Grams are the result of segmenting a text in *n* fragments and are being employed e.g. in computer linguistics.

[19] http://www.culturomics.org/.

[20] Jean Baptiste Michel *et al*.: Quantitative Analysis of Culture Using Millions of Digitized Books. In: *Science*, Vol. 331 (2011), Nr. 6014, S. 176–182, online 16.12.2010, http://www.sciencemag.org/content/331/6014/176 (DOI: 10.1126/science.1199644).