

DIGITALISIERUNG UND LANGZEITARCHIVIERUNG

Max Kaiser

Austrian Books Online

Das *Google-Books*-Projekt an der Österreichischen Nationalbibliothek

In Kooperation mit Google digitalisiert die Österreichische Nationalbibliothek ihren historischen, urheberrechtsfreien Druckschriftenbestand. Rund 600.000 Bände vom sechzehnten bis neunzehnten Jahrhundert werden digitalisiert. Das Projekt „Austrian Books Online“ zeigt, dass Public Private Partnerships erfolgreich zum Ziel von Kulturerbeinstitutionen beitragen können, umfassenden Zugang zu ihren Beständen zu ermöglichen. Die Nicht-Exklusivität solcher Partnerschaften und der freie Zugang zu den Digitalisaten ist hierfür Voraussetzung. Der Artikel erläutert die Rahmenbedingungen der Partnerschaft, zeichnet den Projektaufbau und die Workflows nach und gibt einen Ausblick auf die Pläne zur Zugänglichmachung der digitalen Ressourcen.

In Zusammenarbeit mit dem Internetunternehmen Google digitalisiert die Österreichische Nationalbibliothek ihren gesamten historischen Druckschriftenbestand. Diese Kooperation ist die bislang größte österreichische Public Private Partnership im Kulturbereich. Im Projekt *Austrian Books Online*¹ werden rund 600.000 ausschließlich urheberrechtsfreie Werke vom Anfang des 16. Jahrhunderts bis zur zweiten Hälfte des 19. Jahrhunderts mit insgesamt rund 200 Millionen Seiten digitalisiert und über *Google Bücher* sowie die Digitale Bibliothek der Österreichischen Nationalbibliothek kostenfrei zugänglich gemacht. Der historische Buchbestand der Österreichischen Nationalbibliothek zählt zu den bedeutendsten Sammlungen weltweit.

Google Bücher, von Google Ende 2004 initiiert, besteht aus zwei Programmen: Im sogenannten „Partnerprogramm“ kooperiert Google mit mehr als 10.000 Verlagen weltweit, die ihre Bücher zur Digitalisierung und Onlineveröffentlichung zur Verfügung stellen. Den BenutzerInnen wird in *Google Bücher* eine begrenzte Anzahl von für die Suche relevanten Buchseiten gezeigt. Über die Anzahl der Seiten, die pro Buch angezeigt werden, entscheidet der Verlag. Bei Interesse können die BenutzerInnen das Buch über einen Link zur Website des Verlags oder zu einem Onlinehändler erwerben.

Im „Bibliothekenprogramm“ kooperiert Google mit rund 40 renommierten Bibliotheken weltweit in der Digitalisierung, darunter bislang 13 Bibliotheken in Europa.² Neben der Österreichischen Nationalbibliothek arbeiten in Europa auch die Nationalbibliotheken von Italien, den Niederlanden, der

¹ <http://www.onb.ac.at/austrianbooksonline>

² Eine aktuelle Liste der Partner findet sich in den „Frequently Asked Questions“ von *Austrian Books Online*: <http://www.onb.ac.at/austrianbooksonline/faq.htm>.

Tschechischen Republik sowie die British Library mit Google zusammen. Projektpartner ist auch die Bayerische Staatsbibliothek, die bereits mehr als 500.000 Bücher mit Google digitalisiert hat.

Als Ergebnis der Partnerschaften mit Verlagen und Bibliotheken können über *Google Bücher* (<http://books.google.at>) bereits mehr als 20 Millionen digitalisierte Bücher gefunden und durchsucht werden. Davon sind rund drei Millionen Bände urheberrechtsfrei und werden vollständig im Volltext angezeigt. Anders als in den USA digitalisiert Google in Europa mit Bibliotheken ausschließlich urheberrechtsfreie Werke.

Rahmenbedingungen der Kooperation der Österreichischen Nationalbibliothek mit Google

Die Kooperation zwischen der Österreichischen Nationalbibliothek und Google wurde der Öffentlichkeit im Juni 2010 vorgestellt. Begleitend zur Ankündigungspressekonferenz stellte die Österreichische Nationalbibliothek eine umfangreiche Projektwebsite mit ausführlichen „Fragen und Antworten“ zu den wesentlichen Aspekten der Partnerschaft ins Netz.³ Diese FAQs, die deutsch und englisch vorliegen, werden ständig aktualisiert.

Für das Ziel einer umfassenden Digitalisierung des historischen Buchbestandes der Österreichischen Nationalbibliothek lag die Partnerschaft mit Google nahe, verfügt das Unternehmen doch nicht nur über die erforderlichen Ressourcen, sondern auch über vieljährige praktische Erfahrung in der Massendigitalisierung. Die Kosten für die Volltextdigitalisierung von Büchern sind sehr hoch und betragen typischerweise zwischen 50 und 100 Euro pro Band.⁴ Ohne Kooperation mit einem Partner wie Google wäre ein Projekt der Größenordnung von *Austrian Books Online* für die Österreichische Nationalbibliothek daher kaum leistbar gewesen. Google kommt für die Kosten von Volltextdigitalisierung, Transport und Versicherung auf, die Österreichische Nationalbibliothek trägt den nicht unerheblichen Aufwand für die logistischen Prozesse innerhalb der Bibliothek, die erforderlichen Anpassungen der Metadaten und für die Qualitätskontrolle, Speicherung, Sicherung und Bereitstellung ihrer Digitalisate.

Die Kooperation zwischen der Österreichischen Nationalbibliothek und Google wird durch klar definierte Rahmenbedingungen geleitet, die im übrigen den wesentlichen Empfehlungen des im Jänner 2011 veröffentlichten

³ <http://www.onb.ac.at/austrianbooksonline/faq.htm>.

⁴ Vgl. dazu 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.

Berichts *The New Renaissance*⁵ einer hochrangigen EU-Expertengruppe sowie den Eckpunkten der Empfehlung der EU-Kommission *Zur Digitalisierung und Online-Zugänglichkeit kulturellen Materials und dessen digitaler Bewahrung*⁶ vom Oktober 2011 entsprechen:

- Es dürfen ausschließlich urheberrechtsfreie Werke digitalisiert werden.
- Die Kooperation mit Google ist nicht exklusiv. Die Österreichische Nationalbibliothek kann daher denselben Buchbestand jederzeit auch mit anderen Partnern digitalisieren.
- Die Entscheidung darüber, welche Bücher für die Digitalisierung in Frage kommen, trifft ausschließlich die Österreichische Nationalbibliothek.
- Die Logistik- und Digitalisierungsprozesse werden durch das Institut für Restaurierung der Österreichischen Nationalbibliothek begleitet und evaluiert.
- Die Österreichische Nationalbibliothek erhält Kopien sämtlicher Digitalisate inklusive der Volltexte und kann diese zur nichtkommerziellen Nutzung über ihre Digitale Bibliothek zugänglich machen. Die Einschränkungen in Bezug auf die nichtkommerzielle Nachnutzung durch Dritte sind zeitlich begrenzt.
- Google und die Österreichische Nationalbibliothek sind verpflichtet, die vollständigen Digitalisate inklusive der Volltexte kostenfrei zugänglich zu machen. Diese Verpflichtung besteht auch über die Dauer der Kooperation hinaus.
- Die Österreichische Nationalbibliothek kann ihre Digitalisate auch über andere Plattformen wie das europäische Kulturportal *Europeana* zugänglich machen und an Forschungspartner weitergeben.
- Die Österreichische Nationalbibliothek kann die Kooperation beenden, sollte sich herausstellen, dass sie nicht den Erwartungen entspricht.

Neben der Verbesserung der Zugänglichkeit der digitalisierten Werke leistet das Projekt *Austrian Books Online* auch einen wichtigen Beitrag zur Konservierung der Bücher, da in Zukunft ja nicht immer die wertvollen Originale benützt werden müssen. Zudem ist die Digitalisierung des historischen Buchbestands der Österreichischen Nationalbibliothek Teil der Bestandssicherung für den Katastrophenfall.

⁵ The New Renaissance. Report of the „Comité des Sages”. Reflection Group on Bringing Europe’s Heritage Online. Jänner 2011. http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf. Deutsche Übersetzung: http://ec.europa.eu/information_society/activities/digital_libraries/doc/executivesummary/final_renaissance_de.pdf.

⁶ http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/recom28nov_all_versions/de.pdf.

Projektdurchführung

In einem Vorprojekt in der zweiten Jahreshälfte 2010 wurden die Voraussetzungen für die operative Projektphase geschaffen. Wichtig war insbesondere die Einbettung des Projekts, das fast sämtliche Abteilungen des Hauses berührt, in die Organisationsprozesse der Österreichischen Nationalbibliothek und die Vernetzung mit den sonstigen laufenden Projekten. Geplant werden mussten die für die operative Phase erforderlichen Personalressourcen und organisatorischen Maßnahmen. Den größten Raum nahm in der Vorbereitung die Planung der logistischen Abläufe ein.

Es werden keine Bücher individuell für die Digitalisierung ausgewählt. Vielmehr wird buchstäblich Regal für Regal digitalisiert. Nicht Teil des Projekts sind lediglich Bücher, die aufgrund ihres fragilen Zustands, ihres Formats oder ihres besonderen Wertes für Digitalisierung nicht in Frage kommen. Vor der Digitalisierung wird jeder einzelne Band auf seine Eignung evaluiert. Digitalisiert werden neben den historischen Buchbeständen inklusive der Werke im Prunksaal der Österreichischen Nationalbibliothek auch die urheberrechtsfreien Bücher der Kartensammlung, der Sammlung von Handschriften und Alten Drucken, der Musiksammlung und des Theatermuseums. Teil des Projekts sind auch die Bestände der Fideikommissbibliothek, d.h. der ehemaligen Privatbibliothek des Hauses Habsburg-Lothringen. Handschriften und Manuskripte werden in diesem Projekt nicht digitalisiert.

Derzeit sind an der Österreichischen Nationalbibliothek rund 70 Personen mit dem Projekt befasst, von denen mehr als 20 MitarbeiterInnen ausschließlich für *Austrian Books Online* arbeiten, und zwar in den Bereichen Buchlogistik, Buchvorbereitung, Metadaten und Katalogisierung, Konservierung und Restaurierung, Datendownload und Qualitätskontrolle, Softwareentwicklung, IT-Infrastruktur und Projektmanagement.

Ende 2010 wurde mit den großflächigen Bücheraushebungen begonnen und schließlich im Frühjahr 2011 die Digitalisierung gestartet. Bis Ende 2011 wurden rund 50.000 Bände digitalisiert, die zum größten Teil bereits in digitaler Form über *Google Bücher* abrufbar sind.

Die logistischen Prozesse sind komplex und aufwändig. Derzeit finden die Bücheraushebungen vor allem im barocken Prunksaal der Österreichischen Nationalbibliothek statt, der eine besondere logistische Herausforderung darstellt. So sind die rund 200.000 Bände teilweise in drei Reihen hintereinander aufgestellt. Um die Effizienz der Bücheraushebung und -rückstellung zu erhöhen, wird mit einer Hebebühne gearbeitet, die allerdings aus Platzgründen in der Galerie nicht zum Einsatz kommen kann.

In einem für das Projekt neu adaptierten Manipulationsbereich werden die Bücher vor der Digitalisierung durch ein Bearbeitungsteam überprüft und mit Barcodes versehen. Barcodes sind für das Tracking im Logistik- und Digitalisierungsprozess und für die spätere Verlinkung der Digitalisate mit den

entsprechenden Metadaten im elektronischen Katalog erforderlich. Der historische Buchbestand der Österreichischen Nationalbibliothek ist zwar katalogisiert, jedoch müssen für dieses Projekt Anpassungen an den Katalogisaten vorgenommen werden. So muss jeder neue Barcode mit einem Metadatensatz verknüpft werden. Besonders aufwändig sind mehrbändige Werke: Hier wird für jeden einzelnen Titel ein eigener neuer Datensatz („Stücktitel“) angelegt. Das Projekt führt daher auch zu einer signifikanten Katalogverbesserung. Komplexere Katalogisierungsfälle, die sich nicht im festgelegten zur Verfügung stehenden Zeitrahmen von derzeit maximal acht Minuten pro Band durchführen lassen, werden von einem eigenen Team bearbeitet. Dies betrifft beispielsweise die sogenannten Adligate, also Bände, bei denen in früheren Jahrhunderten mehrere Werke zusammengebunden wurden. In einem zeitintensiven Prozess muss jedes einzelne Werk identifiziert, katalogisiert und einem Barcode zugeordnet werden.

Ein Sonderfall sind die rund 100.000 Bände der bislang noch nicht katalogisierten Fideikommissbibliothek. Die Katalogisierung wird durch ein Team von vier MitarbeiterInnen durchgeführt. Somit wird dieser wichtige Bestand nicht nur digitalisiert, sondern erstmals auch im elektronischen Katalog nachgewiesen und somit suchbar sein.

Zu den vorbereitenden Arbeiten gehört auch die abermalige Überprüfung und Freigabe der Bände durch das Institut für Restaurierung. Wenn erforderlich, werden durch das Restaurierungsteam Sicherungsarbeiten vorgenommen, etwa Verstärkung von Buchrücken, Sicherung von Einbänden oder Fixieren von losen Blättern.

Die Volltextdigitalisierung erfolgt in einem Digitalisierungszentrum von Google in Deutschland. Google verwendet eine innovative Scan-Technologie, die eine äußerst buch-schonende Digitalisierung erlaubt. Sämtliche Prozesse wurden mit dem Institut für Restaurierung der Österreichischen Nationalbibliothek und dem Österreichischen Bundesdenkmalamt abgestimmt. Da es sich um temporäre Ausfuhr von Kulturgut ins Ausland handelt, müssen alle Lieferungen vorab durch das Bundesdenkmalamt genehmigt werden. Die Bücher werden nach dem Rücktransport abermals überprüft, wieder in das Bibliothekssystem eingebucht und schließlich in ihre Regale zurückgestellt. Derzeit sind Werke, die im Rahmen des Projekts *Austrian Books Online* digitalisiert werden, für LeserInnen durchschnittlich drei Monate nicht benutzbar.

Digitalisate und Digitale Bibliothek

Pro Jahr müssen derzeit mehr als 15 Millionen digitalisierte Seiten automatisiert von Google übernommen und verarbeitet werden. Die Digitalisate haben dieselbe Qualität wie jene, die über *Google Bücher* angeboten werden und setzen sich aus Bilddateien, Volltextdateien (OCR-Ergebnisse) und Strukturdaten zusammen. Google stellt die Daten über eine Softwareschnittstelle be-

reit. Für Datendownload, -kontrolle und -ablage wird die von der Österreichischen Nationalbibliothek entwickelte Applikation ADOCO (*ABO Download and Control*) eingesetzt.

Anspruchsvoll ist insbesondere die Qualitätssicherung. Viele der an *Google Bücher* teilnehmenden Bibliotheken verzichten aufgrund der großen Datenmengen in diesem Massendigitalisierungsprojekt ganz auf Qualitätskontrolle. Die Österreichische Nationalbibliothek hat sich dafür entschieden, Qualitätskontrolle auf Basis automatisierter Routinen und repräsentativer Stichproben durchzuführen. Ziel ist nicht das Auffinden einzelner Fehler, sondern EDV-unterstütztes Entdecken von Clustern möglicher systematischer Probleme. Diese solcherart herausgefilterten Digitalisate können dann in einem zweiten Schritt manuell überprüft werden. Bei der automatisierten Qualitätskontrolle kann die Österreichische Nationalbibliothek auf Ergebnissen von EU-Forschungsprojekten zu digitaler Langzeitarchivierung und Massendigitalisierung aufbauen.⁷

Die Speicherung von Digitalisaten und OCR-Daten erfordert eine signifikante Erweiterung des Massenspeichersystems der Österreichischen Nationalbibliothek. Die JPEG-2000 Masterfiles werden redundant abgelegt, die Access-Versionen für die BenutzerInnen werden jeweils „on-the-fly“ über einen Image Content Server generiert.

Um die Zitierbarkeit und permanente Identifikation der digitalisierten Werke sicherzustellen, wird für jeden Buchtitel bzw. Band ein URN („Uniform Resource Name“) aus dem von Nationalbibliotheken verwalteten Namensraum NBN („National Bibliography Number“) vergeben.⁸ In Zusammenarbeit mit der Österreichischen Bibliothekenverbund und Service GmbH (OBVSG) wird derzeit ein URN-Resolver implementiert, der ab Ende 2012 im Österreichischen Bibliothekenverbund zum Einsatz kommen wird. Hierbei kann auf Software der Deutschen Nationalbibliothek, die bereits seit mehreren Jahren einen Persistent Identifier Resolution Service betreibt, aufgebaut werden.⁹

Ab Ende 2012 wird die Österreichische Nationalbibliothek die Digitalisate sukzessive über ihre Digitale Bibliothek zugänglich machen. In einer ersten Ausbaustufe werden BenutzerInnen die digitalisierten Bücher über die seit Frühjahr 2011 zur Verfügung stehende *QuickSearch*-Suchmaschine der Österreichischen Nationalbibliothek finden. Für die Bereitstellung der Digitalisate wird derzeit ein Book-Viewer entwickelt, der u.a. Funktionalitäten wie Blät-

⁷ Planets: <http://www.planets-project.eu>; IMPACT: <http://www.impact-project.eu>; SCAPE: <http://www.scape-project.eu>.

⁸ Zu URN NBNs allgemein siehe aktuell die Reports der persID Initiative: <http://www.persid.org/initiative.html>.

⁹ Zur URN-Strategie der Deutschen Nationalbibliothek siehe: <http://www.persistent-identifier.de/?link=3352>.

tern und Zoomen enthalten wird. Es wird auch möglich sein, einzelne Seiten bzw. komplette Bücher als PDFs herunterzuladen. Im Rahmen des Projekts digitalisierte historische Tageszeitungen werden über das bewährte Interface des Zeitungsportals der Österreichischen Nationalbibliothek ANNO (*Austrian Newspapers Online*¹⁰) bereitgestellt werden.

In einer zweiten Ausbaustufe, voraussichtlich ab Ende 2013, wird auch eine Volltextsuche in den Inhalten der Bücher möglich sein. Auch die Entwicklung spezieller Applikationen für mobile Endgeräte (z.B. iPhone, iPad, Android-Geräte) ist für einen späteren Zeitpunkt geplant. Die Österreichische Nationalbibliothek wird die digitalen Bücher auch über andere Plattformen, wie das europäische Kulturportal *Europeana* oder das Portal der europäischen Nationalbibliotheken *The European Library*, zugänglich machen.

Ausblick

Vor allem die in naher Zukunft mögliche Volltextsuche im digitalisierten historischen Buchbestand der Österreichischen Nationalbibliothek wird neue Möglichkeiten für die Forschung eröffnen.¹¹ Die BenutzerInnen werden bei der Suche nach relevanten Büchern nicht mehr auf den Katalog der Österreichischen Nationalbibliothek eingeschränkt sein, sondern werden auch die Inhaltsverzeichnisse, Register oder den Text der Bücher nach Wörtern oder Phrasen durchsuchen können und so auf Werke stoßen, die sie andernfalls möglicherweise nicht gefunden hätten. Mittelfristig plant die Österreichische Nationalbibliothek, die digitalen Volltexte mit zusätzlichen Informationen anzureichern und auf dieser Basis neben einer Volltextsuche weitere Services anzubieten. So sollen beispielsweise Ortsnamen und Personennamen in Texten identifiziert werden. Diese können dann für die Suche speziell indiziert und mit anderen Ressourcen („Linked Data“) verlinkt werden.

Die Anwendung neuer computerunterstützter Methoden auf große Volltext-Corpora ermöglicht neue wissenschaftliche Fragestellungen in den Geistes-, Kultur- und Sozialwissenschaften. Durch die Integration der digitalisierten Ressourcen in digitale Forschungsinfrastrukturen stehen diese für datenzentrierte Fragestellungen zur Verfügung. Ein Beispiel für das neu entstehende, datenzentrierte Forschungsparadigma ist der „Digging into Data Challenge“, der von einer internationalen Kooperation von Forschungsförderungseinrichtungen seit 2009 ausgeschrieben wird.¹² Ein Werkzeug für die als „Culturo-

¹⁰ <http://anno.onb.ac.at>.

¹¹ Vgl. Gregory Crane. What Do You Do with a Million Books? In D-Lib Magazine, März 2006. <http://www.dlib.org/dlib/march06/crane/03crane.html>.

¹² „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

mics¹³ bezeichneten Analysemethoden auf Grundlage einer großen Datenbasis („Big Data“) stellt Google selbst seit Ende 2010 mit seinem „Books Ngram Viewer“¹⁴ zur Verfügung, der es ermöglicht, Häufigkeitsanalysen von Wörtern oder Phrasen im Textkorpus von *Google Books* durchzuführen.

Austrian Books Online wird in den nächsten Jahren den bedeutenden historischen Buchbestand der Österreichischen Nationalbibliothek einem weltweiten Publikum unabhängig von Ort und Zeit zugänglich machen. Das Projekt wird die Verfügbarkeit der Werke signifikant verbessern und somit einen Beitrag zur Demokratisierung von Wissen und Kultur leisten. Aktuelle Informationen zu *Austrian Books Online* inklusive ausführlicher „Fragen und Antworten“ sind auf der Projektwebsite ([http://www.onb.ac.at/austrianbooks online/](http://www.onb.ac.at/austrianbooksonline/)) zu finden.

searches, sensor data or cell phone records – what new, computationally-based research methods might we apply?” <http://www.diggingintodata.org>.

¹³ <http://www.culturomics.org>. Vgl. Jean Baptiste Michel u.a. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. In *Science* 331/6014, 176–182. Online seit 16.12.2010; <http://www.sciencemag.org/content/331/6014/176> (DOI: 10.1126/science.1199644).

¹⁴ <http://ngrams.googlelabs.com>. N-Gramme sind das Ergebnis der Zerlegung eines Textes in Fragmente und finden Anwendung u.a. in der Computerlinguistik.