Archival Cultural Heritage Online: Eine Virtuelle Forschungsumgebung im Spannungsfeld von Open Access, Nachhaltigkeit und Datenschutz

Lange, Felix
flange@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin

Wintergrün, Dirk
dwinter@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin

Wannenwetsch, Oliver
oliver.schmitt@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

Schoepflin, Urs
schoepfl@mpiwg-berlin.mpg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

Wie sich die langfristige wissenschaftliche Nutzbarkeit von großen digitalen Datenrepositorien sicherstellen lässt, ist eine in den letzten Jahren in der DH-Community und darüber hinaus intensiv diskutierte und noch nicht abschließend geklärte Frage. In den Digitalen Geisteswissenschaften werden in diesem Zusammenhang zur Zeit vorrangig Probleme der technischen Nachhaltigkeit und der Datenstandards diskutiert [Fornaro (2016)]. Im Hinblick auf Repositorien für die gegenwartsnah arbeitenden geistes- und sozialwissenschaftlichen Disziplinen wie die Zeitgeschichte sind aber auch komplexe datenschutz- und urheberrechtliche Anforderungen zu berücksichtigen. Eine im Archivwesen diskutierte Antwort auf diese Herausforderung ist es, die rechtliche Absicherung des Zugangs zu Digitalisaten in sog. „digitalen Lesesälen“ zu organisieren, die einen Zugriff ausschließlich in den Räumlichkeiten des jeweiligen Archivs zulassen [Plassmann (2016), S. 219]. Dabei wird aber das Ziel der Open-Access-Bewegung, wissenschaftliche Quellen und Forschungsergebnisse einer möglichst großen Fachöffentlichkeit zugänglich zu machen, verfehlt. In den Sozialwissenschaften hat die Brisanz dieser Frage bereits zur Gründung von Datenzentren geführt, die Fragen der technischen *und* der rechtlichen Datensicherheit in den Mittelpunkt stellen. Der vorliegende Beitrag stellt mit Archival Heritage Online – ArCHO eine digitale Forschungsinfrastruktur vor, die dazu dient, das Verlangen nach offener wissenschaftlicher Nutzung mit den rechtlichen Bedingungen für den nachhaltigen Zugang zu zeitgeschichtlichem Archivmaterial in Einklang zu bringen.

Der prototypische Anwendungsfall für ArCHO ist das seit 2014 laufende und auf zunächst fünf Jahre angelegte Forschungsvorhaben „Geschichte der Max-Planck-Gesellschaft“ (GMPG). Es untersucht die Geschichte der MPG von ihrer Gründung im Jahre 1948 bis zum Jahr 2002 und legt dabei den Schwerpunkt auf institutsu#bergreifende Fragestellungen zu Themenfeldern wie Periodisierungen, Innovationen, Internationalisierung, Forschung und Wirtschaft, Gender und Wissenschaft sowie Konkurrenz und Kooperation. Diese Themen lassen sich naturgemäß nicht allein durch kleinere Fallstudien bearbeiten, sondern erfordern thematisch und chronologisch breit angelegte Querschnittsuntersuchungen mit einer entsprechend umfänglichen Quellengrundlage. Aus diesem Grund wird im Laufe des Projektes ein großes digitales Textkorpus angelegt, dessen Schwerpunkt Digitalisate von mehreren Regalkilometern an Verwaltungsschriftgut aus der Generalverwaltung der MPG und einzelnen Instituten bilden. Desweiteren werden thematisch spezialisierte Datenbestände wie eine Patent- und eine Personendatenbank sowie ein digitales Korpus mit Veröffentlichungen der MPG aufgebaut. Mit dafür entwickelten oder angepassten Tools [Kruse et al. (2015)] lassen sich so beispielsweise Konjunkturen von Forschungsthemen, unterschiedliche professionelle Netzwerke zwischen WissenschaftlerInnen und wissenschaftliche Karrierewege erforschen. Im Sinne der guten wissenschaftlichen Praxis sollen die Arbeitsergebnisse, also sowohl die digitalisierten und annotierten Quellen als auch alle statistischen Auswertungen, mindestens zehn Jahre nach Projektende abrufbar bleiben. ArCHO als digitales Findmittel und Analyseplattform ist daher mit einem Fokus auf langfristiger Verfügbarkeit von Forschungsdaten konzipiert worden. Dabei wurde eine Nachhaltigkeitsstrategie entwickelt, die der noch ungeklärten Aufgabenteilung zwischen Forschungseinrichtungen, Gedächtnisinstitutionen sowie Daten- und Rechenzentren bei der Langzeitarchivierung geisteswissenschaftlicher Forschungsdaten Rechnung trägt. Denn diese Aufgabe kann angesichts der großen technischen Komplexität und des Wartungsaufwandes für Virtuelle Forschungsinfrastrukturen sowie der großen Menge an vorzuhaltenden Daten nicht allein Gedächtnisinstitutionen wie wissenschaftlichen Archiven überantwortet werden. Andererseits sind Rechen- und Datenzentren nur bedingt dazu in der Lage, neben dem Archivrecht auch komplexe spezifische Zugangsregeln für einzelne Datenrepositorien umzusetzen. Daher ermöglicht es ArCHO mit einem verlässlichen Zugangsmanagement,

dass die Forschungseinrichtung den Zugang selbst rechtssicher regeln kann.

Die in ArCHO implementierte Zugangsverwaltung setzt auf eine starke Differenzierung von Nutzerrollen einerseits und von Bestandteilen einzelner Datensätze andererseits. Auf der Nutzerseite muss beispielsweise im Anwendungsfall GMPG unterschieden werden zwischen der wissenschaftlichen Öffentlichkeit, Forschern innerhalb des Forschungsvorhabens mit einem privilegierten Zugang zu den Aktenbeständen und einem Projektkollegium, das besonders sensible Datenbestände nach einer Einzelfallprüfung für die Forscher freigibt. Weitere Abstufungen von Zugangsrechten können sich aus spezifischen Aufgabenbereichen bei der Dateneingabe und -verwaltung ergeben [vgl. Neuroth et al. (2010), 16:14 ff.]. Die Aufgabe der Zugangsregelung muss auch nach Projektende weiter von dazu befugten Personen ausgeübt werden können und ist daher eine wichtiger Nachhaltigkeitsaspekt. Denn beispielsweise ist bei datenschutzrechtlich sensiblen Dokumenten mit Personenbezug, deren Sichtung durch Forscher der Einwilligung der betroffenen Personen bedarf, je nach konkreter rechtlicher Ausgestaltung diese Bewilligung an das Forschungsvorhaben und damit an dessen Laufzeit gebunden. Die Nutzungserlaubnis erlischt in diesen Fällen nach Projektende und entsprechend muss auch der digitale Zugang verwehrt werden. Auf der anderen Seite werden manche Akten erst nach Ende der Archivschutzfrist vollständig nutzbar, was in einer nachhaltigen Forschungsinfrastruktur ebenfalls berücksichtigt werden sollte.

Auf der Datenseite ermöglicht ArCHO eine starke Differenzierung von einzelnen zu einem Dokument gehörenden Daten mit dem Ziel, unter Einhaltung der rechtlichen Vorgaben möglichst viele Informationen für die Forschung zur Verfügung zu stellen. So sind bei einer Personalakte mit sensiblen Inhalten möglicherweise die Signatur, Laufzeit und Angaben zur inhaltlichen Klassifikation durch das haltende Archiv nicht schutzwürdig, wohl aber der Volltext und der Titel. Es kann also je nach Bestand jedes Metadatum und jedes Derivat des Digitalisates (OCR-Erfassungen u.a.) eine andere Schutzwürdigkeit haben. Die Gesamtzahl dieser Regeln, die zwischen beliebigen Typen von (Meta-)Daten unterscheiden, und die Vielzahl von abgestuften Nutzerrechten führen zu einer Matrix aus Nutzerrollen und Teildatensätzen, deren einzelne Werte sich stets ändern können. Sie wird technisch realisiert durch einen sogenannten *Policy Decision Point* (PDP). Dabei handelt es sich um ein außerhalb des eigentlichen Dokumentkorpus angesiedeltes und technisch eigenständiges Softwaremodul, das zwischen der Nutzer-Datenbank und dem Korpus vermittelt.

Die Umsetzung eines solchen Rechtemodells innerhalb einer ansonsten marktüblichen Webanwendung leistet den oben geschilderten Anforderungen aber noch nicht Genüge. Denn ein solches System wäre höchst verwundbar gegenüber Hacking-Angriffen. So ist denkbar, dass durch *Injection*-Attacken sensible Teile der Datenbank, und im schlimmsten Fall sogar die Zugangsverwaltung, ausgelesen werden. Weiterhin stellt der Download größerer Mengen an Dateien im Projektalltag ein gewisses Risiko der ungewollten Weiterverbreitung dar und ist angesichts der notwendigen hohen Dokumentqualität auch recht zeitaufwändig. Eine sinnvolle Alternative ist daher eine Viewer-Anwendung, welche Dokumente bereits serverseitig so gut aufbereitet, dass ein kompletter Download vermieden werden kann. Die Anforderungen solch vergleichsweise komplexer Anwendungen an die Client-Software (i. A. Browser) können jedoch im Laufe der Zeit zu Inkompatibilitäten führen und somit die Nachhaltigkeit der gesamten Anwendung gefährden.

Daher realisiert ArCHO auf der Ebene der Middleware mit Containern und Virtualisierung Architekturprinzipien, wie sie (aus zum Teil sehr verschiedenen Gründen) in der Diskussion um nachhaltige wissenschaftliche Software zur Zeit eine große Rolle spielen. In der konkreten Implementierung wird erreicht, dass der Bildschirm des Nutzers einen per RDP-Protokoll bereitgestellten Virtuellen Desktop zeigt, der jeweils Einzelansichten von Dokumentseiten wiedergibt. Diese können nicht ohne Weiteres heruntergeladen werden. Auch ein programmatischer Zugriff auf die Datenbank ist nicht möglich, daher können Angreifer keinen massenhaften Abzug sensibler Daten erreichen. Außerdem wird die Webanwendung als solche technisch nachhaltig gemacht. Denn da sie sich in einem sehr stark abgeschlossenen System befindet, ist die technische Konfiguration des Client-Rechners zumindest mittelfristig fast ohne Belang.

Die geschilderte Kombination von Virtualisierung, Middleware-Containern und der feingranularen Zugangsverwaltung ist eine pragmatische Antwort auf das ungelöste Problem des rechtssicheren Zugangs zu schutzwürdigen digitalisierten Archivalien. Sie bietet eine Alternative zur räumlichen Zugangsbeschränkung auf Archivlesesäle. ArCHO soll dazu beitragen, die nachhaltige Nutzbarkeit von Daten, die in Forschungsprojekten erhoben wurden, über Orts- und Disziplingrenzen hinweg zu ermöglichen und damit eines der wesentlichen Versprechen der Digitalisierung in den Geisteswissenschaften einzulösen.

Fußnoten

1. Als Beispiel einer über die Geisteswissenschaften hinausgehenden, internationalen Initiative sei die Arbeit der Research Data Alliance genannt: https://rd-alliance.org/ [letzter Zugriff 20. August 2016].
2. S. hierzu die „Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities“, die auch vom MPI für Wissenschaftsgeschichte unterstützt wird: https://openaccess.mpg.de/Berlin-Declaration [letzter Zugriff 20. August 2016].
3. Z. B. das „GESIS Secure Data Center“: http://www.gesis.org/en/services/data-analysis/data-archive-

service/secure-data-center-sdc/ [letzter Zugriff 26.11. 2016].
4. ArCHO befindet sich zum Zeitpunkt der Abfassung im Stadium eines Prototypen und wird in der Projektlaufzeit zu einem generischen Service erweitert.
5. http://gmpg.mpiwg-berlin.mpg.de [letzter Zugriff 20. August 2016].
6. Vgl. die Empfehlung 7 der Denkschrift „Sicherung guter wissenschaftlicher Praxis“ der DFG: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf [letzter Zugriff 20. August 2016].
7. Dieses für die meisten historischen Quellenkorpora absurd klingende Szenario ist im Bereich der Zeitgeschichte durchaus als denkbar anzunehmen.
8. Beachte z. B. die thematische Ausrichtung des FORGE-2016-Workshops: https://www.gwiss.uni-hamburg.de/gwin/ueber-uns/forge2016.html [letzter Zugriff 20. August 2016].
9. Die Desktop-Virtualisierung wird mit Apache Guacamole realisiert: https://guacamole.incubator.apache.org [letzter Zugriff 20. August 2016].

Bibliographie

Fornaro , Peter R. / Rosenthaler, Lukas (2016): „File Formats for Archiving: Stability and Persistence Issues“, in: *DH2016: Conference Abstracts* 507–508.

Kruse, Sebastian / Schmaltz, Florian / Stiller, Juliane / Wintergru#n, Dirk (2015): „Herausforderung ‚Big Data’ in der historischen Forschung“, in: *DHd 2015: Von Daten zu Erkenntnissen* 171–174 https://dhd2015.uni-graz.at/de/nachlese/book-of-abstracts [letzter Zugriff 20. August 2016].

Neuroth, Heike / Oßwald, Achim / Scheffel, Regine / Strathmann, Stefan / Huth, Karsten (2010): *nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Göttingen: Niedersächsische Staats- und Universitätsbibliothek Göttingen http://www.nestor.sub.uni-goettingen.de/handbuch/index.php [letzter Zugriff 20. August 2016].

Plassmann, Max (2016): „Archiv 3.0? Langfristige Perspektiven digitaler Nutzung“, in: *Archivar. Zeitschrift für Archivwesen* 3: 219–223 http://www.archive.nrw.de/archivar/hefte/2016/Ausgabe_3/Archivar_3_2016.pdf [letzter Zugriff 20. November 2016].