Gregor Neuböck*

Kollaboration und Web 2.0

Eine neue Dimension der Bearbeitung innerhalb der Digitalen Landesbibliothek Oberösterreich

DOI 10.1515/bfp-2016-0065

Zusammenfassung: Die Digitale Landesbibliothek Oberösterreich (DLOÖ)¹ ist heute viel mehr als nur ein einfaches Retrodigitalisierungsprojekt. Mittlerweile umfasst das Angebot Web-2.0-Werkzeug Möglichkeiten zur Erstellung für Anmerkungen, zum Anlegen eigener Sammlungen oder zur Speicherung von Suchen. Mit einem eigenen Crowdsourcing-Modul können Volltexte verbessert, Annotationen in Bildern oder Transkriptionen von Handschriften durchgeführt werden. Nur exklusive bearbeitbare Datenfelder zur wissenschaftlichen Beschreibung von Handschriften durch Wissenschaftler runden das Angebot ab.

Schlüsselwörter: Crowdsourcing; Web 2.0; Digitale Landesbibliothek Oberösterreich; Erfahrungsbericht

Collaboration and Web 2.0. A New Dimension of Workings Within the Digitale Landesbibliothek Oberösterreich

Abstract: The Digital National Library of Austria (DLOÖ) is much more than just a simple retro-digitization project. We now include the possibility to work with the Web 2.0 tool for notes, or to help you create your own collections or for storing your searches. With an own crowdsourcing engine full text, can be improved, annotations in images or transcriptions of manuscripts are possible. Exclusively this programme gives access to editable data fields which can help to clarify scientific descriptions of manuscripts by scholars and scientists.

Keywords: Crowdsourcing; web 2.0; Digitale Landesbibliothek Oberösterreich; field report

1 Vorwort

Die „Digitale Landesbibliothek Oberösterreich“ ist ein Digitalisierungsprojekt das ursprünglich zur Retrodigitalisierung urheberrechtsfreier Bücher entwickelt wurde. Heute

beinhaltet dieses neben der Digitalisierung urheberrechtsfreier Bücher und der Verwaltung elektronischer Pflichtexemplare eine Fülle an Services und Angeboten, die sich an den interessierten Laien genauso wie an Wissenschaft und Forschung richten.

Die Oberösterreichische Landesbibliothek hat ganz bewusst einen Schwerpunkt auf die Entwicklung neuer Werkzeuge der DLOÖ im Bereich Datenvernetzung und -anreicherung gesetzt, da wir uns davon eine spürbare Verbesserung der Qualität unserer Metadaten erwarten. Um derartige Projekte gut managen zu können, ist es wichtig zu verstehen, dass sich ständig verändernde Prozesse systemimmanente Kennzeichen von Web-2.0-Projekten sind. So gesehen stellt dieser Erfahrungsbericht eine Momentaufnahme in der Entwicklung der DLOÖ dar. Unser vorrangiges Ziel war und ist die Verbesserungen von Usability und Qualität der Metadaten. Durch den Einsatz neuer technischer Möglichkeiten (z. B. Crowdsourcing, Linked Open Data,² named-entity recognition³) wollen wir diese Kriterien bestmöglich vorantreiben.

In diesem Artikel gebe ich Einblicke in die wichtigsten Entwicklungen der DLOÖ im Bereich Kollaboration/Web 2.0.

2 Das Crowdsourcing-Modul der DLOÖ

Die Software für das Crowdsourcing⁴-Modul basiert im Kern auf einer Erstentwicklung der Landesbibliothek Berlin, ist plattformunabhängig und in JAVA⁵ programmiert.

1 <http://digi.landesbibliothek.at>.

*Kontaktperson: Dipl. Päd. Gregor Neuböck, MAS, M.Sc., gregor.neuboeck@ooe.gv.at

2 Linked Open Data: im Web frei verfügbare Daten, die per URI (Uniform Resource Identifier) identifiziert und vernetzt werden.

3 Named-entity recognition: Prozess, mit dem aus einem Text eindeutige Referenzen wie Eigennamen, Orte, Zeiten oder ähnliches herausgezogen und klassifiziert werden.

4 Crowdsourcing oder Crowdworking: angelehnt an den Begriff Outsourcing, bezeichnet es die Auslagerung von Arbeiten an die Allgemeinheit.

5 JAVA ist eine objektorientierte und plattformunabhängige Programmiersprache.

Abb. 1: Oberfläche Crowdsourcing-Modul

Im Januar 2014 startete die Entwicklung gemeinsam mit unserer Betreuungsfirma (Intranda), dabei wurden einerseits bereits vorhandenen Funktionen radikal verändert, andererseits neue Features in einem relativ rasanten Feedbackprozess entwickelt. Insbesondere die Bedienoberfläche (GUI⁶) wurde vollständig runderneuert. Unsere Anforderungen an Funktionalitäten und Usability wichen erheblich von denen der Landesbibliothek Berlin ab, sodass man angesichts der großen Umwälzungen eher von einer Neuentwicklung sprechen darf.

Ursprünglich wurde die Software zur Verbesserung des maschinengelesenen OCR⁷-Textes der Adressbücher Berlins entwickelt. Die schlechte Qualität der Vorlagen und die Vielzahl verschachtelter Tabellen bedingten besondere Anforderungen und spezielle Funktionalitäten.

Die Oberösterreichische Landesbibliothek hatte aufgrund der vorhandenen Materialheterogenität diametrale Anforderungen und Ziele. Wir hatten die Anforderung, dass Monographien, Bandserien, Zeitschriften, Zeitungen, alte Drucke, Karten, Inkunabeln und Handschriften mit Metadaten angereichert und Volltexte verbessert werden können und wollten Wortkoordinaten für korrektes Highlighting⁸ bei der Volltextsuche realisieren. Anders gesagt, wir suchten ein Werkzeug das viel breiter aufgestellt und

in unterschiedlichsten Einsatzszenarien verwendbar wäre. Trotz unseres international bescheidenen Bestandes (ca. 4500 Werke online; Stand Mai 2016) war uns rasch klar, dass auch wir diese Aufgabe hausintern nicht leisten werden können.

Schon heute haben wir durch die Crowd bei einzelnen Werken einen deutlich verbesserten Volltext, da Benutzer den Volltext von sich aus nachbearbeitet haben.⁹

In der neuen graphischen Oberfläche kann man Bildanzeige und Volltexte nebeneinander legen. Bewegt man den Mauszeiger über den Volltext, wird in der Bildanzeige um das zugehörige Wort ein roter Rahmen eingeblendet. So kann man größere Textmengen relativ rasch auf ihre Korrektheit überprüfen. Klickt man mit der Maus nun auf ein Wort, wird ein Eingabewerkzeug auf der Volltextseite eingeblendet. Mit diesem kann man ein Wort editieren, löschen (Symbol Mülltonne) oder mit den Pfeiltasten ein weiteres Wort davor/danach oder eine Zeile oberhalb/unterhalb einfügen.

Zusammenhängende sinnlose Zeichenabfolgen, wie sie immer wieder bei maschinengelesenem Volltext vorkommen, können nun komfortabel mit einem eigens entwickelten „Löschenmodus“ entfernt werden.

⁶ GUI: Graphical User Interface: grafische Benutzeroberfläche oder auch grafische Benutzerschnittstelle.

⁷ OCR: optical character recognition: Texterkennung oder Optische Zeichenerkennung.

⁸ Highlighting bezeichnet die farbliche Hervorhebung von bestimmten Suchbegriffen in den einzelnen Treffern.

⁹ <http://digi.landesbibliothek.at/viewer/fulltext/AC00565746/97/>, http://digi.landesbibliothek.at/viewer/fulltext/AC08513816_Alphabetische_Verzeichnis_Nr_V/32/, <http://digi.landesbibliothek.at/viewer/fulltext/279/147/>, <http://digi.landesbibliothek.at/viewer/fulltext/AC08858389/115/>.

Abb. 2: Löschenmodus

Das Löschen ganzer Seiten ist ebenfalls möglich, wurde allerdings bewusst eingeschränkt und bedarf spezieller Rechte.

Vom Volltext eines Werkes werden neben der ursprünglichen OCR-Datei alle weiteren Versionen des Volltextes als einzelne ALTO¹⁰-Dateien (XML-Format) gespeichert, sodass bei Bedarf jederzeit auf den ursprünglichen Volltext zurückgesetzt werden kann. Die spätere Nachnutzung in Fremdsystemen ist dadurch ebenfalls garantiert. Unsere OAI-PMH¹¹-Schnittstelle bietet dazu einen maschinenlesbaren Zugang in vielen verschiedenen Dateiformaten (dc, ese, mets, marcxml, epicur und lido) an.¹² Zusätzlich werden alle diese Dateiversionen regelmäßig archiviert, mit einem Zeitstempel und einer ID versehen, was die Zuordnung zu den Bilddateien der Bücher innerhalb des Dateisystems enorm erleichtert. Neben der Verbesserung von Volltexten besteht bei uns zusätzliche die Möglichkeit, Inhalte zu bearbeiten.

Annotationen können derzeit zu den Bereichen *Person*, *Einrichtung*, *Adresse* und *Kommentar* erfasst werden. Jeder Annotationstyp wird mit einem roten Rahmen und einem eigenen Symbol kenntlich gemacht, sodass den Benutzern sofort der Zusammenhang zwischen der jeweiligen Textstelle (dieser befindet sich unterhalb des Bildes), der Art der Annotation und dem zugehörigem Bildbereich klar wird.

In den Metadatenfeldern der einzelnen Annotationsarten können derzeit persistente Identifier, GND¹³-Daten, Namen, Geodaten, Orte, Straßen, Länder und Anmerkungen erfasst werden. Das Entwicklungspotential für diese Felder scheint enorm und kann heute noch gar nicht abgeschätzt werden. Man denke da z. B. nur an die Georeferenzierung von Karten.

Dass sich mittlerweile auch die Albertina in Wien im Zusammenhang mit Bildklebebänden für dieses Modul interessiert, spricht für das Potential dieses Werkzeuges.

Eine weitere wichtige Funktionalität stellt der Transkribiermodus dar. Bei Werken die über keinen Volltext (insbesondere Handschriften) verfügen, wird optional ein Transkribiermodus angeboten. In diesem werden keine Wortkoordinaten vergeben, stattdessen kann man in einem WYSIWYG-Editor Fließtext eingeben.

Da die wissenschaftliche Transkription per se schon eine enorme Herausforderung darstellt, wollten wir keine zusätzlichen Hürden einbauen, indem jedes einzelne Wort als Wortkoordinate erfasst werden muss, auch wenn dies optional möglich wäre.

Seit einiger Zeit entdecken immer mehr Benutzer dieses Werkzeug, und wenn dann Handschriften auf hohem Niveau bearbeitet werden, freut man sich als Betreiber umso mehr.¹⁴ Um die Zugangshürden zu unseren Werkzeugen so gering wie möglich zu halten, genügt schon ein OpenID¹⁵-Account, man kann sich allerdings auch ganz klassisch mit E-Mail und Passwort über ein Webformular anmelden.¹⁶

Im Januar 2015 wurde der Probetrieb aufgenommen und seit März des Jahres wird das Crowdsourcing-Modul zur Datenanreicherung im Vollbetrieb geführt.

3 Weitere Web-2.0-Werkzeuge

Der Zugang zu den nun folgenden Funktionalitäten erfordert die Registrierung via Anmeldeformular oder einen OpenID-Account.

3.1 Kommentarfunktion

Im Jahr 2014 entwickelten wir die Möglichkeit, *plain text* in einem Feld unterhalb jedes Bildes einzutragen. Manche Nutzer können uns mit diesem Werkzeug¹⁷ z. B. auf Fehler

¹⁰ ALTO: Dateiformat, das zur Beschreibung der Volltexte auf einer Seite zusammen mit dem zugehörigen Layout dient. Wurde von der Library of Congress entwickelt.

¹¹ OAI-PMH (protocol for metadata harvesting): ist eine Schnittstelle zum Einsammeln und Weiterverarbeiten von Metadaten.

¹² <http://digi.landesbibliothek.at/viewer/oai/>.

¹³ GND: ist eine Normdatei für Personen, Körperschaften, Kongresse, Geografika, Sachschlagwörter und Werktitel, die vor allem zur Katalogisierung von Literatur verwendet wird.

¹⁴ <http://digi.landesbibliothek.at/viewer/fulltext/279/29/>, <http://digi.landesbibliothek.at/viewer/image/279/29/-/>, <http://digi.landesbibliothek.at/viewer/fulltext/177/167/>, <http://digi.landesbibliothek.at/viewer/image/177/167/-/>.

¹⁵ OpenID ist ein dezentrales Authentifizierungssystem für webbasierte Dienste. Der Benutzer muss sich nur einmal bei einer Institution registrieren und kann sich danach bei allen systemunterstützten Webseiten über diesen Zugang einloggen.

¹⁶ <http://digi.landesbibliothek.at/viewer/createAccount>.

¹⁷ <http://digi.landesbibliothek.at/viewer/image/krih/19>.

Abb. 3: Annotationsmodus

Abb. 4: Transkribiermodus

aufmerksam machen, andere wiederum generieren Mehrwerte¹⁸ für alle Beteiligten.

3.2 Bücherregalfunktion

Mit dieser Funktion können Benutzer eigene virtuelle Bücherregale erstellen und benennen, um diesen dann Bü-

cher zuzuweisen. So können „Lieblingsbücher“ oder thematisch ähnliche Büchersammlungen rasch auffindbar gemacht werden. Man kann sich eine eigene kleine virtuelle Bibliothek schaffen. Optional können diese auch der Allgemeinheit zugänglich gemacht werden (öffentliche Bücherregale).

¹⁸ <http://digi.landesbibliothek.at/viewer/image/296/1>.

3.3 Suchanfragen speichern

Auch das Abspeichern von Suchen ist mit der Anmeldung möglich. Komplizierte und aufwändige Suchanfragen können dann jederzeit wieder rasch aufgerufen werden, ohne die zuvor verwendete Syntax erneut eingeben zu müssen.

3.4 GND-Einbindung

Als weitere externe Quelle zur Datenanreicherung wird seit mehr als eineinhalb Jahren die GND-Anbindung genutzt, die uns zusätzliche lokale Suchmöglichkeiten nach unterschiedlichen Schreibweisen ermöglicht, da der gesamte Datensatz der GND von unserem Solr-Indexer indexiert wird. Demonstrieren lässt sich das z.B. an folgendem GND-Datensatz¹⁹ von Herder, indem man per copy and paste eine der asiatischen Schreibweisen seines Namens (约翰·哥特弗里特·赫尔德) in den Suchschlitz eingibt und fündig wird. Zu finden sind die GND-Einträge, sofern vorhanden, im Menüpunkt *Bibliographischen Daten*.

3.5 Übersichtsseite

Als Ergänzung zum Crowdsourcing-Modul haben wir eine optionale (wird nach Bedarf generiert) Übersichtsseite für unsere Handschriften entwickelt. Diese setzt sich aus einem Literatur- und einem Beschreibungsfeld sowie verschiedenen bibliographischen Daten zusammen.

Die Übersichtsseite dient der wissenschaftlichen Tieferfassung von Handschriften und kann nur mit Sonderrechten (Benutzer müssen sich auf der Webseite anmelden und erhalten danach von uns die notwendigen Rechte zugewiesen) bearbeitet werden, einerseits um neueste wissenschaftliche Erkenntnisse zu Handschriften direkt neben den Digitalisaten zur Verfügung²⁰ zu stellen, andererseits um weitere Quellen (Literaturangaben, Links zu anderen Webseiten²¹) und Informationen besser einbinden und vernetzen zu können.

In den nächsten Monaten werden wir Daten eines Projektes der Österreichischen Akademie der Wissenschaften zur Buchmalerei spätgotischer Handschriften der Oberösterreichischen Landesbibliothek unter der Federführung von Frau Dr. Katharina Hranitzky sukzessive in unser Portal einarbeiten.

¹⁹ http://digi.landesbibliothek.at/viewer/metadata/1055/1/LOG_0000.

²⁰ http://digi.landesbibliothek.at/viewer/overview/279/1/LOG_0000.

²¹ <http://manuscripta.at/m1/lib.php?libcode=AT5600> oder <http://www.handschriftencensus.de/hss>.

Abb. 5: Übersichtsseite

Im *Beschreibungsfeld* steht ein WYSIWYG-Editor²² zur Verfügung, so kann der eingegebene Text strukturiert und formatiert werden. So können innerhalb eines Fließtextes Links auf besondere Buchmalereien gesetzt werden. Das System wurde so konfiguriert, dass diese Links sich in einem Popup-Fenster öffnen, um die beschriebenen Initia- len in Ruhe betrachten, das Popup danach schließen und anschließend im Fließtext der Hauptseite weiterlesen zu können.

Die Übersichtsseite beinhaltet zusätzlich ein Feld *Literatur*. In dieses können einschlägige Quellen zu einer bestimmten Handschrift angegeben werden. Unterhalb des Beschreibungsfeldes wird ein automatisch generiertes Logfile angezeigt. Durch dieses wird sofort ersichtlich, wer, wann, welche Beiträge zu einer bestimmten Handschrift geleistet hat.

Somit entwickelt sich im Laufe der Zeit eine Bearbeitungs-Chronologie, was aus wissenschaftlicher Sicht wünschenswert und notwendig ist.

4 Was die nähere Zukunft bringt?

In den nächsten Wochen werden wir eine neue Version unseres Viewers installieren. Mit dieser erhalten wir ein völlig neues Bildmanagement, welches große Bilder rasch in sehr guter Auflösung zur Verfügung stellt und sich gleichzeitig an den Standard *International Image Interoperability Framework* (IIIF) hält. Die IIIF-Gemeinschaft dient der Normierung der Datenbereitstellung, führt zur Verbes-

²² WYSIWYG-Editor: Akronym für *What you see is what you get* und bezeichnet einen Editor bei dem man am Bildschirm sofort das Layout der späteren Ausgabe sieht.

serung der Usability, vereinfacht den Datenaustausch und die internationale Zusammenarbeit in der Forschung.

Named-entity recognition wird ebenfalls mit der neuen Version in Grundzügen realisiert und in Zukunft können E-Pubs²³ aus unseren Volltexten hergestellt werden, so müssen nur geringe Datenmengen von Benutzern heruntergeladen werden, wenn es vorrangig um den Text eines Buches geht.

Im Laufe des Herbsts planen wir die Implementierung der automatisationsunterstützten Handschriftenerkennung aus dem EU-Projekt *Transkribus*²⁴ um Dr. Günther

Mühlberger in die DLOÖ. Von diesem Werkzeug erwarten wir uns einen weiteren Schub in der wissenschaftlichen Tiefenerfassung von Handschriften.

Dipl. Päd. Gregor Neuböck, MAS, M.Sc.

Amt der Oö. Landesregierung
Direktion Kultur
Landesbibliothek
Schillerplatz 2
A-4021 Linz
Österreich
gregor.neuboeck@ooe.gv.at

²³ E-Pub ist ein Akronym für *electronic publication* und ist ein offener Standard für E-Books.

²⁴ <https://transkribus.eu/Transkribus>.