CROWDSOURCING AN DER OBERÖSTERREICHISCHEN LANDESBIBLIOTHEK

von Gregor Neuböck

Zusammenfassung: *Die Oberösterreichische Landesbibliothek betreibt im Rahmen der Digitalen Landesbibliothek Oberösterreich (DLOÖ; https://digi.landesbibliothek.at) seit mehr als fünf Jahren ein umfangreiches Crowdsourcingmodul. Das Modul besteht aus unterschiedlichen Teilen mit verschiedenen Funktionalitäten zur Datenanreicherung bzw. zur Verbesserung automatisiert erstellter Volltexte. Zu Beginn des Beitrags beschäftigt sich der Autor mit der Definition des Begriffs Crowdsourcing und dessen Einordnung in verwandte Begrifflichkeiten. Nach einer Aufzählung entscheidender Punkte für einen erfolgreichen Einsatz dieser Technik erfolgt eine Darstellung der Gründe für die Implementierung von Crowdsourcing innerhalb der DLOÖ. Im Hauptteil widmet sich der Autor ausführlich den einzelnen Modulteilen, erklärt Funktionalitäten und Nutzen sowie ihre bisherigen Einsatzgebiete und berichtet über zukünftige Entwicklungen und Ziele.*

Schlagwörter: *Oberösterreichische Landesbibliothek; Digitale Landesbibliothek Oberösterreich; Crowdsourcing*

CROWDSOURCING AT THE UPPER AUSTRIAN FEDERAL STATE LIBRARY

Abstract: *The Upper Austrian Federal State Library operates as part of the Digital Library of Upper Austria (DLOÖ; https://digi.landesbibliothek.at) and has had an extensive crowdsourcing module for more than five years. The module consists of different parts which have different functions, data enrichment and the improvement of automatically created full texts. At the beginning of the article, the author deals with the definition of crowdsourcing and its classification into related terms. After a list of decisive points for the successful use of this technique, a presentation of the reasons for the implementation of crowdsourcing within the DLOÖ is discussed. In the main part, the author elaborates in detail the individual module parts, explains functionalities and benefits as well as their previous areas of application and report on future developments and goals.*

Keywords: *Upper Austrian Federal State Library; Digital Library of Upper Austria; Crowdsourcing*

DOI: *https://doi.org/10.31263/voebm.v72i2.2834*

Inhalt

1. Einleitung

Crowdsourcing, ein Begriff, der seit mehr als zehn Jahren verstärkt Verwendung im Segment der Web2.0-Anwendungen findet (schon 2006 wurde der Begriff von Jeff Howe geprägt)[1], steht vereinfacht gesagt für freiwillige oder bezahlte Arbeit externer Kräfte. Der Begriff selber hat sich aus den Begriffen *Crowd* und *Outsourcing* entwickelt, also einer Art Auslagerung von Arbeit.

Crowdsourcing kann auch als Werkzeug gesehen werden, mit dem man den steigenden Personalbedarf, bedingt durch stetig wachsende Aufgaben im bibliothekarischen Umfeld, durch unentgeltliche Zuarbeit realisieren kann. Bei den eingesetzten Technologien handelt es sich, wie weiter oben schon erwähnt, um sogenannte Web2.0-Technologien.[2] Das bedeutet, dass derartige Technologien auf Techniken basieren, die keine lokalen Programme mehr benötigen, abgesehen von einem Webbrowser. Damit wird ein sehr einfacher Zugriff, unabhängig von Ort und Zeit, realisiert.

Mit Hilfe dieser zusätzlichen Arbeitskräfte können in digitalisierten Werken, insbesondere in automatisiert erstelltem Volltext, Fehler ausgebessert oder zusätzliche Daten erfasst werden. Die Crowd schafft also einen Mehrwert durch verbesserte, angereicherte Volltexte oder durch die Möglichkeit, zusätzlich Daten zu erfassen (z.B. werden in einem Bild der Name oder die Geodaten eines Berges eingetragen).

Welches Potential Web2.0-Technologien besitzen, zeigt z.B. das Megaprojekt Wikipedia. Wer allerdings Crowdsourcing effektiv einsetzen möchte, ist gut beraten, sich im Vorfeld mit dem Zusammenwirken von Menschen in Netzwerken eingehend auseinanderzusetzen.[3]

Im Zentrum unserer Überlegungen sollten immer die Crowdworker stehen, aber wie können diese dazu motiviert werden, ihre Arbeitskraft zur Verfügung zu stellen? Finanzielle Anreize stellen nur ein Mittel zur Verbesserung der Motivation, viel wichtiger aber ist es, immaterielle Anreize in den Vordergrund zu stellen. Zugehörigkeit zu einer Gruppe sowie Anerkennung von dieser sind als wesentliche Motivatoren anzusehen. Wechselseitige Anerkennung und Aufbau einer Community sind ebenso zu nennen wie extrinsische Faktoren, also z.B. die Bereitschaft etwas beizutragen, weil man dadurch später vielleicht beruflich profitieren könnte[4]. Zusätzlich sollte ein besonderer Fokus auf im Ruhestand befindliche Expertinnen und Experten gelegt werden. Diese haben nach Meinung des Autors ein großes Potential und könnten bei geeigneten Rahmenbedingungen als hochqualifizierte Crowdworker gewonnen werde.

2. Ausgangslage

Im Rahmen der DLOÖ wird überwiegend urheberrechtsfreie Literatur der Oberösterreichischen Landesbibliothek digitalisiert. Der Schwerpunkt liegt dabei auf den sogenannten „Obderennsia“[5], jener identitätsstiftenden Literatur, die meist als Pflichtabgabe in den Bestand der Bibliothek gelangte oder wie im Fall der Handschriften und Inkunabeln als Folge der Klosterauflösungen Josephs des II. der Bibliothek zugeschlagen wurden. Zusätzlich digitalisieren wir seltene Stücke aus dem Bestand unserer alten Drucke (ca. 30.000 vor 1850), Handschriften (ca. 1.500, davon 350 mittelalterliche Vollhandschriften) und Inkunabeln (ca. 845 Drucke und somit sechstgrößte Sammlung Österreichs).

Gleichwohl sich die Bibliothek in der Digitalisierung von Anfang an für einen Weg umfangreicher Datenerfassung entschieden hat, ist zusätzlich ein optimaler Volltext unumgänglich, um exzellente Suchanfragen zu ermöglichen. Umfangreiche Datenerfassung bedeutet für uns, dass wir in jedem Buch neben allen Kapiteln und Unterkapiteln, alle vorhandenen weiteren Strukturen wie z.B. Abbildungen, Vorwörter, Einleitungen, Briefe, Gedichte, Tabellen, Bibliographien, Sachregister,... erfassen. In Kombination mit Fremddaten wie z.B. der GND erzeugen wir so eine beeindruckende Datendichte, die ihrerseits wiederum Grundvoraussetzung

für granulare Suchanfragen ist. So kann in der DLOÖ z.B. nach Abbildungen mit dem Titel „Gmunden“ oder nach Karten mit dem Titel „Perg“ gesucht werden.

Der Volltext der OCR[6] kann nach dem derzeitigen Stand der Technik niemals perfekt sein, insbesondere bei Frakturschriften aber auch bei verschachtelten Tabellen kommt es zu mehr oder weniger guten Erkennungsraten. Suchanfragen führen so oftmals nicht zum gewünschten Erfolg. Schon vor vielen Jahren war uns diese Problematik bewusst und so suchten wir nach Möglichkeiten diese Fehler zu eliminieren, bzw. auszubessern.
Wir entwickelten ein Crowdsourcingmodul (es handelt sich programmiertechnisch um eine eigene Einheit unseres Goobi-Viewers) auf Basis einer früheren Entwicklung der Berliner Landesbibliothek, erweiterten dieses um zusätzliche Funktionalitäten und orientierten uns bei der Entwicklung an einer möglichst intuitiven Bedienung.

3. Voraussetzungen

Um in der DLOÖ mitzuarbeiten, muss man sich einmalig unter https://digi.landesbibliothek.at/viewer/user/ registrieren (siehe Abbildung 1). Entweder man meldet sich mit einem eigenen Google-Account an oder man erstellt ein lokales Konto durch einen Klick auf den Link „Neues Benutzerkonto erstellen“. Nach der Bestätigung eines zugesendeten Links kann man als „Crowdworker“ mitarbeiten. Für die Crowd-Experts (siehe weiter unten im Kapitel „Übersichtsseite“) sind besondere Rechte erforderlich, die auf Anfrage von einem Administrator der DLOÖ exklusiv vergeben werden müssen.

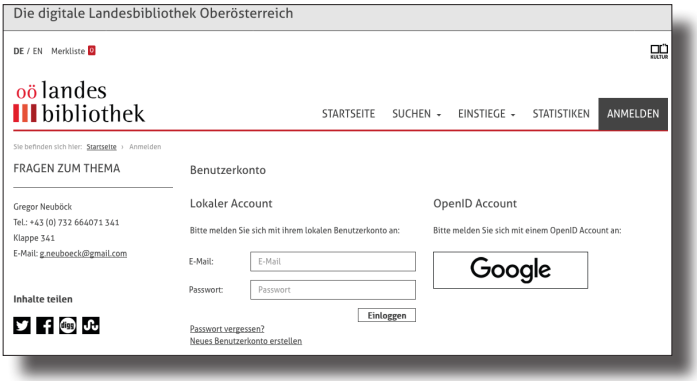

Abb. 1: Registrierung für die Mitarbeit in der DLOÖ

4. Volltexte bearbeiten

Alle Drucke im Workflow der DLOÖ laufen durch einen automatisierten OCR-Schritt. Im OCR-Schritt werden neben dem Volltext auch Wortkoordinaten[7] erfasst, auf Basis derer das sogenannte Highlighting, also die farbige Hervorhebung von Suchtreffern, realisiert ist. Um eine langfristige Nutzung der Volltexte zu garantieren, verwenden wir als Speicherformat ALTO (Version 2.1)[8]. Für jede Benutzerin/jeden Benutzer wird eine eigene ALTO-Datei angelegt, sodass jederzeit die ursprüngliche Textversion wiederhergestellt werden kann. Möglicher Missbrauch, also die bewusste Eingabe falscher Daten, kann somit über die Versionierungs-Historie rasch behoben werden.

Abhängig von Schriftart, Druckqualität und Satz kommt es zu erheblichen Qualitätsunterschieden bei der OCR. Das Modul „Volltexte bearbeiten" bietet die Möglichkeit diese Unterschiede auszugleichen. Um den Modus „Volltexte bearbeiten" zu starten, muss man sich zuerst anmelden. Möchte man bei einem Werk den Volltext bearbeiten, klickt man links von der Bildanzeige im Kasten „CROWDSOURCING" auf den Link „Am Digitalisat mitarbeiten". Der Link „Meine letzten Aktivitäten" führt zu einer Liste der zuletzt bearbeiteten Seiten, stellt also einen Schnelleinstieg ins Crowdsourcing dar (siehe Abbildung 2).

Abb. 2: Schnelleinstieg ins Crowdsourcing

Beim erstmaligen Aufruf einer Seite im Crowdsourcingmodul gelangt man zuerst ins Hauptmenü des Crowdsourcings. Man hat nun die Auswahl zwischen „Volltexte bearbeiten" und „Inhalte erfassen" (siehe Abbildung 3).

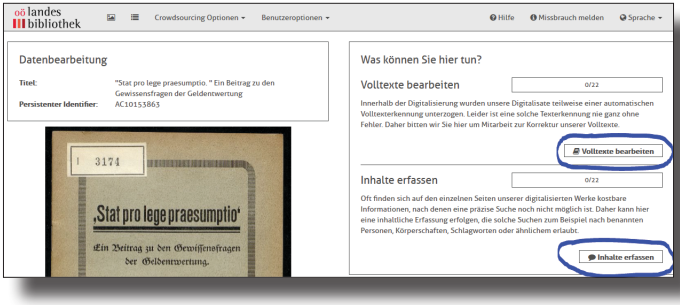

Abb. 3: Hauptmenü Crowdsourcing

Handelt es sich um ein Werk ohne Volltext (z.B. eine Handschrift), erhält man nach dem Klick auf „Volltexte bearbeiten“ eine weitere Auswahl mit „Neues Wort einfügen“ oder „Transkribieren“ (siehe Abbildung 4).

Abb. 4: Bearbeitung von Werken ohne Volltext

Im Modus „Volltexte bearbeiten“ sieht man auf der linken Seite die Bildansicht und rechts den erfassten Volltext. Klickt man auf ein Wort, erscheint im Textbereich ein Navigationswerkzeug „Bearbeiten-Werkzeug“. Gleichzeitig werden Rahmen um das ausgewählte Wort (auf Basis der Wortkoordinaten) auf der linken (Bildseite) und rechten (Textseite) Seite angezeigt. Wörter können korrigiert, gelöscht oder zusätzlich eingefügt werden (siehe Abbildung 5).

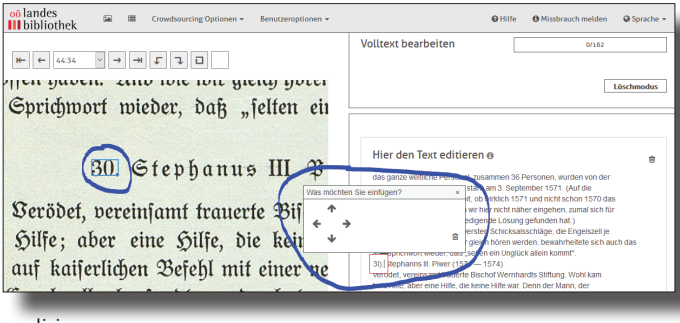

Abb. 5: Text editieren

Im Modus „Transkribieren“ hat man einen WYSIWYG-Editor[9] zur Verfügung (siehe Abbildung 6).

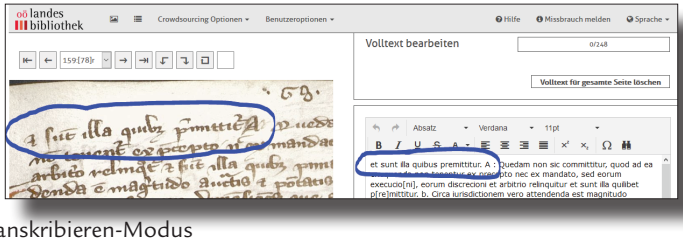

Abb. 6: Transkribieren-Modus

Dieser trägt wesentlich zur Vereinfachung des Eingabemodus bei, allerdings werden in diesem Modus keine Wortkoordinaten erfasst. Als Folge davon kann bei den Suchtreffern aus transkribiertem Text kein Highlighting angezeigt werden.

5. Inhalte erfassen

Mit dieser Funktion können unterschiedliche Arten von Metadaten zu einzelnen Bildern erfasst werden. Bei der Arbeit im betreffenden Modul befindet sich auf der linken Seite der Bildbereich. Auf der rechten Seite erscheint der Datenerfassungsbereich. Derzeit können Daten zu den Themen „Person“, „Einrichtung“, „Adresse“ und „Kommentar“ aufgenommen werden. Alle erfassten Daten werden nach dem Speichern indexiert und stehen schon kurz danach bei Suchanfragen zur Verfügung (siehe Abbildung 7).

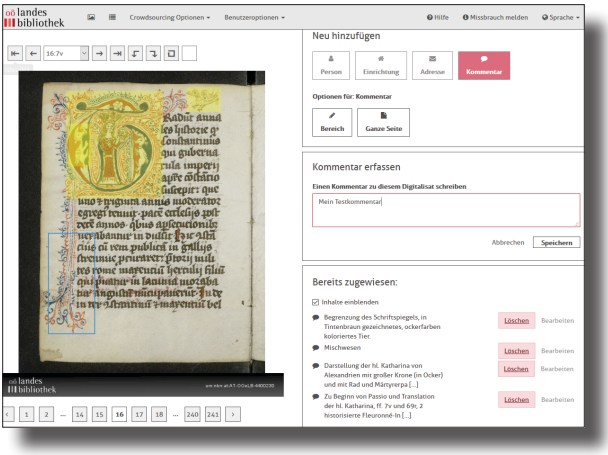

Abb. 7: Indexierung erfasster Daten

6. Kommentarfunktion

Eine einfache Kommentarfunktion wird nach dem Einloggen unterhalb jedes Bildes angeboten. Dieses Tool kann dazu genutzt werden, kurze Kommentare zu einem Bild als „plain text“ einzugeben. In einem Fall wurden wir von einem Benutzer/einer Benutzerin z.B. auf einen fehlerhaften Scan bei einer Zeitschrift hingewiesen (siehe Abbildung 8).

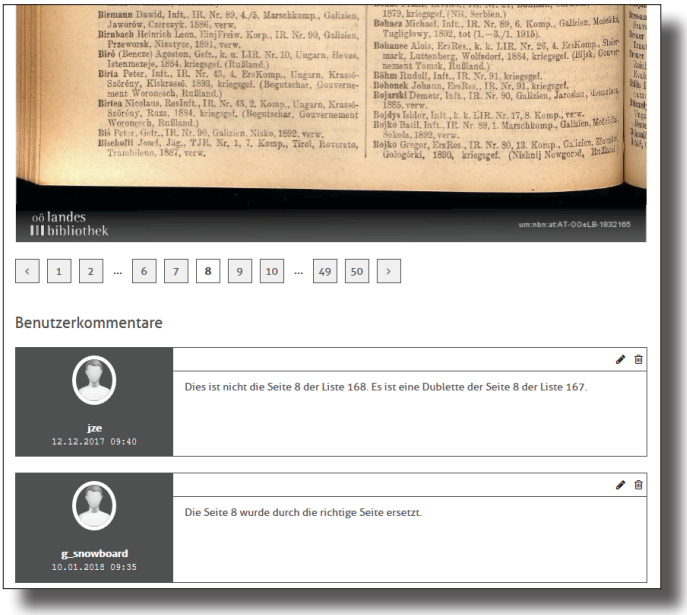

Abb. 8: BenutzerInnenkommentare

7. Übersichtsseite

Der Bereich Übersichtsseite stellt eine besondere Form des Crowdsourcings innerhalb der DLOÖ dar. Er beinhaltet wissenschaftliche Beschreibungen aus den unterschiedlichen Fachbereichen der Handschriftenkunde, die auch Kodikologie genannt wird. Da es sich um inhaltlich sehr hohe Anforderungen handelt, benötigen wir für diesen Bereich hochspezialisierte Expertinnen und Experten. Wir bezeichnen diese Personen deswegen auch als Crowd-Experts (überwiegend wird es sich bei diesen um aktive oder ehemalige Wissenschaftlerinnen/Wissenschaftler handeln).

Auf der Übersichtsseite können die fertigen Beschreibungen im Dateiformat RTF, PDF, DOXC und HTML hochgeladen werden. Nach dem Upload erfolgt automatisiert eine einheitliche HTML-Darstellung der eingespielten Dateien. Es können mehrere Handschriftenbeschreibungen nebeneinander vorhanden sein. Diese können aus unterschiedlichen Fachbereichen (Paläographie, Kunstgeschichte...) stammen bzw. auf Basis unterschiedlicher Merkmale einer Handschrift (Einband, Lagen, Marginalien,...) nebeneinander existieren.

Da es sich, wie zuvor bereits erwähnt, um einschlägiges Expertenwissen handelt, sind zur Bearbeitung der Übersichtsseite Sonderrechte erforderlich. Durch ein ausgeklügeltes Rechtemanagement ist der Zugriff ausschließlich auf die eigenen Handschriftenbeschreibungen möglich.

Für die Übersichtsseite werden persistente Identifier vergeben, so können Wissenschaftlerinnen und Wissenschaftler später auf diese Seite referenzieren, was die Attraktivität aus Sicht der Crowd-Experts wesentlich erhöht.

In Zukunft wird es möglich sein, Handschriftenbeschreibungen im Format TEI[10] herunterzuladen (siehe Abbildung 9).

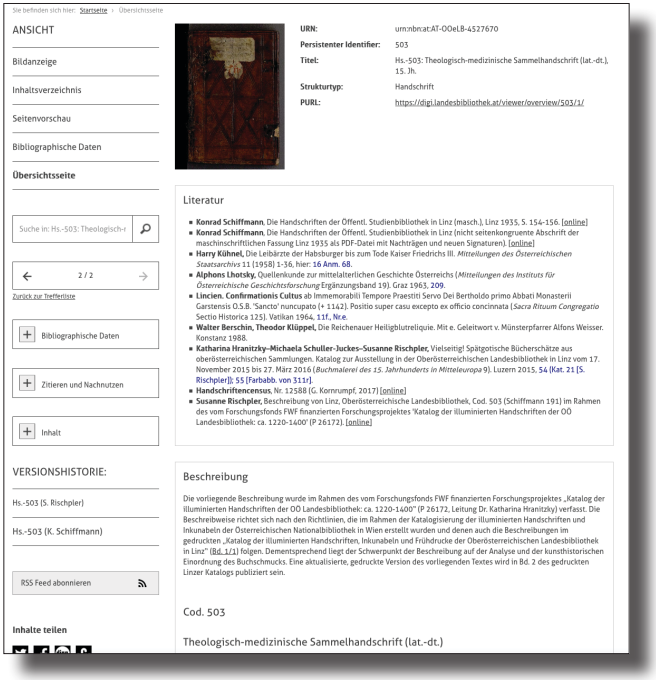

Abb. 9: Handschriftenbeschreibungen in TEI

8. Ein wenig Statistik

Innerhalb der DLOÖ gibt es ca. 1.400 registrierte Nutzerinnen und Nutzer. Bisher wurden von diesen Benutzerinnen und Benutzern an die 50.000 Seiten bearbeitet und an die 450 Kommentare abgegeben. Wir konnten durch diese intensive Nutzung wichtige Erfahrungen machen, was Performance und Usability angeht und haben auf Basis dieser schon viele kleine Verbesserungen durchgeführt.

Im Rahmen eines hauseigenen Projektes zur Korrektur der Verlustlisten Österreich-Ungarns (https://digi.landesbibliothek.at/viewer/resolver?urn=urn:nbn:at:AT-OOeLB-1723425) wurden bisher 2.6 Mio. Namen kontrolliert und bei Bedarf korrigiert. Wie wertvoll dieses genealogische Werkzeug für die Community ist, zeigen 2.500–10.000 Seitenaufrufe pro Tag.

Mittlerweile existieren auch schon 25 wissenschaftliche Handschriftenbeschreibungen (siehe z.B. https://digi.landesbibliothek.at/viewer/overview/20/1/). In diesen steckt oft jahrelange Forschungstätigkeit, weswegen Wissenschaftlerinnen und Wissenschaftler natürlich großen Wert auf die Nachhaltigkeit der verwendeten Systeme legen[11].

9. Ein kleiner Ausblick

Das derzeitige Crowdsourcingmodul wird momentan hinsichtlich Usability gründlich überarbeitet. Der Einstieg ins Crowdsourcing soll durch technische Verbesserungen erleichtert werden, aber auch eine möglichst intuitive Bedienung steht im Zentrum der aktuellen Entwicklung. In Zukunft möchten wir für einen verstärkten Aufbau einer Crowd-Community Präsenzveranstaltungen abhalten und eine geeignete Social Media Kampagne starten. Die erwähnte Möglichkeit wissenschaftliche Handschriftenbeschreibungen als TEI herunterzuladen steht ganz oben auf unserer Agenda.

10. Fazit

Crowdsourcing hat grundsätzlich ein enormes Potential für Bibliotheken.

Fakt ist, dass die Betreuung und Entwicklung dieser modernen Systeme die Situation eines immerwährenden Beta-Stadiums erzeugt, was dafür sorgt, dass sich Softwareprodukte immer wieder erneuern, also viel langle-

biger sind. Es ist aber auch dafür zu sorgen, dass für diese Prozesse ausreichend personelle und finanzielle Ressourcen zur Verfügung gestellt werden.

Die Verknüpfung mit den sich derzeit exponentiell entwickelnden „Neuronalen Netzwerken" bietet ein riesiges Entwicklungspotential für die Zukunft und wird uns mit Sicherheit viele weitere, heute noch gar nicht abschätzbare, spannende Anwendungen ermöglichen.

Durch aktive Teilnahme der Bibliothekarinnen und Bibliothekare an dieser Entwicklung wird dieses Potential in geeignete Bahnen gelenkt werden. Innerhalb der Bibliotheks-Community herrscht noch immer vielerorts eine gehörige Portion Skepsis gegenüber dem Mehrwert, den Crowdsourcing generieren kann. Ein möglicher Missbrauch dieser Technologien, also die bewusste Eingabe falscher Metadaten, wird aus der Erfahrung des Autors heraus zu sehr in den Fokus der Diskussion gerückt, auch wenn man diesen niemals ganz ausschließen können wird. Wir hatten bisher noch keinen einzigen Missbrauchsfall, also die bewusste Eingabe falscher Daten.

Die Bibliothekarinnen und Bibliothekare sind plötzlich in der Situation, nicht mehr unumschränkte Herrscher der Datenhoheit zu sein. Crowdworker und Crowdworkerinnen sind nun beinahe gleichberechtigt bei der Dateneingabe. Diese Situation bedarf eines grundsätzlichen Paradigmenwechsels innerhalb der Kollegenschaft. Erfordert verstärkt Kenntnisse im Bereich der neuen Medien, Datenbanken, Cloudsysteme und Programmierung.

Ein sorgsamer Umgang mit Web 2.0-Technologien und ein Einsatz mit Bedacht und Überlegung schafft Vertrauen und ermöglicht so die Entwicklung neuer Anwendungen mit großem Potential für Bibliotheken in der Zukunft.

Dipl.-Päd. Gregor Neuböck MAS MSc
Oberösterreichische Landesbibliothek, Digitale Bibliothek
E-Mail: gregor.neuboeck@ooe.gv.at

Literatur

Analyzed Layout and Text Object. https://de.wikipedia.org/wiki/Analyzed_Layout_and_Text_Object (1. 4. 2019)

Franzke, Cordula (2017): Repositorien für Forschungsdaten am Beispiel der Digital Humanities im nationalen und internationalen Vergleich. Potentiale und Grenzen. Perspektive Bibliothek 6(1), S. 2–33. https://doi.org/10.11588/pb.2017.1.42164

Kollman, Tobias; Markgraf, Daniel (2018): Crowdsourcing. In: Gablers Wirtschaftslexikon. https://wirtschaftslexikon.gabler.de/definition/crowdsourcing-51787/version-274938

Papsdorf, Christian (2009): Wie Surfen zur Arbeit wird. Crowdsourcing im Web 2.0. Frankfurt/New York: Campus Verlag.

Text Encoding Initiative. https://de.wikipedia.org/wiki/Text_Encoding_Initiative (1. 4. 2019)

Vilain, Michael; Wegner, Sebastian (2018): Crowds, Movements & Communities. Potentiale und Herausforderungen des Managements in Netzwerken. Baden-Baden: Nomos Verlagsgesellschaft.

Texterkennung. https://de.wikipedia.org/wiki/Texterkennung (18. 3. 2019)

WYSIWYG. https://de.wikipedia.org/wiki/WYSIWYG (18. 3. 2019)

1 Papsdorf, Christian (2009): Wie Surfen zur Arbeit wird. Crowdsourcing im Web 2.0. Frankfurt/New York: Campus Verlag, S. 25.

2 Kollman, Tobias; Markgraf, Daniel (2018): Crowdsourcing. In: Gablers Wirtschaftslexikon. https://wirtschaftslexikon.gabler.de/definition/crowdsourcing-51787/version-274938

3 Vilain, Michael; Wegner, Sebastian (2018): Crowds, Movements & Communities. Potentiale und Herausforderungen des Managements in Netzwerken. Baden-Baden: Nomos Verlagsgesellschaft, S. 5.

4 Papsdorf, S. 127.

5 Obderennsia bezeichnet landeskundliche Literatur des Landes Oberösterreich

6 OCR (optical character recognition) bezeichnet die automatisierte Texterkennung: https://de.wikipedia.org/wiki/Texterkennung

7 Wortkoordinaten werden dazu benutzt dem erkannten Volltext die Fundstelle innerhalb eines Bildes zuzuweisen, auf Basis dessen eine farbliche Hervorhebung von Suchtreffern im Bild möglich wird.

8 ALTO (Analyzed Layout and Text Object): https://de.wikipedia.org/wiki/Analyzed_Layout_and_Text_Object

9 WYSIWYG-Editor (What You See Is What You Get) bietet eine grafische Oberfläche zur Formatierung von Text: https://de.wikipedia.org/wiki/WYSIWYG

10 TEI (Text Encoding Initiative) ist ein Format zum Austausch und zur Kodierung von Text: https://de.wikipedia.org/wiki/Text_Encoding_Initiative

11 Franzke, Cordula (2017): Repositorien für Forschungsdaten am Beispiel der Digital Humanities im nationalen und internationalen Vergleich. Potentiale und Grenzen. Perspektive Bibliothek 6(1), S. 2f. https://doi.org/10.11588/pb.2017.1.42164