Anwendungen von DH-Methoden in der Erschließung und Digitalisierung von Kulturerbe. Ein Vorschlag zur Systematisierung

Franken, Lina
lina.franken@uni-hamburg.de
Universität Hamburg, Deutschland

Bevor Kulturerbe digital verfügbar und nachnutzbar ist, stehen komplexe Prozesse an, die als spezifische „invisible work“ (Star/Strauss 1999) bezeichnet werden können. Diese werden kaum außerhalb der engeren Community problematisiert: Wo finde ich in den analogen Katalogen, Zettelkästen, Findbüchern u.ä. die notwendigen Informationen zur Erfassung und Erschließung? Welche Sammlungsbestände sind besonders wichtig und deshalb zu digitalisieren? Ist eine Massen-Digitalisierung mit geringer Detailtiefe der Erfassung oder eine detaillierte Erschließung einiger Teilbestände sinnvoll? Wie können die Vorgehensweisen von unterschiedlichen Personen und Institutionen vereinheitlicht werden? Und wie können Informationen, die bereits digital vorliegen – etwa in Tabellen oder älteren Erfassungssystemen – an internationale Regelwerke und Standards zur Metadatenhaltung angepasst werden?

Kernanliegen des vorgeschlagenen Beitrags ist es, für die Sammlungserschließung und Digitalisierung – gerade in Museen, aber auch in Archiven und Bibliotheken – Methoden, Tools und Analyseperspektiven der Digital Humanities stärker als bisher zu nutzen. Dafür werden mögliche Synergien und exemplarische Anwendungen in einer ersten Exploration aufgezeigt. Dabei wird auf eigene Projekterfahrungen zurückgegriffen, die im Kontext des Digitalisierungsvorhabens „Digitales Portal Alltagskulturen im Rheinland“ (2013-2017) gesammelt wurden.[1] Diese sind bereichert um Perspektiven des DH-Forschungsverbundes an der Universität Hamburg, „Automatisierte Modellierung hermeneutischer Prozesse (hermA)“ (2017-2020).[2]

What (not) to do with 100.000 Pictures. Sammlungserschließung als Sisyphos-Arbeit?

Im LVR-Institut für Landeskunde lagern umfangreiche Bestände, die nach Sammlungsgeschichte gegliedert und nur rudimentär erschlossen sind. Vor allem die Fotodokumentationen sind für Erschließung und Digitalisierung prädestiniert, zeigen sie doch plurale Facetten immateriellen Kulturerbes seit etwa 1900[3]. Dazu kommen umfangreiche Materialien zu 31 schriftlichen Befragungen aus den 1970er bis 2000er Jahren sowie kleinere Einzelsammlungen. Die Fotobestände liegen in unterschiedlichen Negativformaten sowie als Dias chronologisch sortiert vor. Dazu sind Fotoabzüge auf Karteikarten geklebt, mit Metadaten versehen und thematisch abgelegt in einer über Jahrzehnte gewachsenen Systematik. Negative und Abzüge sind also unterschiedlich archiviert und nur in Einzelfällen einander zuzuordnen. Es bestehen auf Einzelplatzrechnern gepflegte Bestandslisten sowie eine Datenbank[4] mit Teilinventarisierung. Seit 2013 werden im Rahmen der LVR-weiten Digitalisierungsstrategie Digitalisate erstellt, mit angereichert und Metadaten abgelegt, um sie mittelfristig öffentlich zugänglich zu machen.[5] Die Systeme werden aktuell in die Datenbank digiCULT.web[6] übertragen, wodurch mit dem LIDO-Format größere Datenmengen standardkonform publiziert werden können.

Vergleichbare gewachsene Systeme gibt es in vielen Museen, Archiven und Bibliotheken. Gerade in kleineren Museen und Forschungseinrichtungen, deren Hauptaugenmerk auf der Ausstellung bzw. Forschung und weniger auf der Sammlungserschließung liegt, sind die Erfassungen unvollständig und in der Institutionsgeschichte von verschiedenen Akteuren mit spezifischen Wissenshintergründen und Systematiken erfolgt. Diese Systeme funktionieren vor allem aufgrund des Wissens von Archivar*innen und Forschenden (vgl. zu Wissenskonzepten Koch 2006): Sie wissen, welche Materialien wo liegen, welche Strukturen was auffindbar machen, was in welchen Kontexten verwendet wurde. Wenn Stellenwechsel oder Verrentungen anstehen, geht dieses Wissen verloren oder wird bruchstückhaft weitergegeben.

Wenn analoge Bestände digitalisiert werden sollen, geht es in der Vorbereitung (z.B. der Antragsstellung für Drittmittel) um hohe technische Qualität, Fragen des Workflows sowie der Bereitstellung, einzuhaltende Standards oder eigene Präsentationsoberflächen. Die konkreten Arbeitsschritte zur Umsetzung dieser Zielvorgaben werden oft erst in der Projektrealisierung entwickelt. Gerade die Museumsdatenbanken sind geprägt von einer gewachsenen Pluralität, die in dieser Form nicht als veröffentlichungswürdig gelten und strukturelle Nachbearbeitungen erfordern.

Für die hier exemplarisch stehenden Fotobestände des LVR-Instituts fiel die Entscheidung, sich nicht an den bestehenden (Ablage-)Systematiken zu orientieren. Neben anderen Argumenten war dabei die unvollständige und heterogene, thematisch abgelegte Erschließung per Karteikarte ausschlaggebend. Um eine Auswahl zur inhaltlichen Erschließung zu ermöglichen und analoge Vorarbeiten gering zu halten, wurden alle Negative und Dias digitalisiert (zur Problematik von Original und Kopie, die sich für Foto-Abzüge spezifisch stellt, vgl. Schönholz 2017). Im Zuge der Auftragsvergabe wurden Bestände erstmals gezählt und liegen als vollständige Digitalisate vor. Ein Meilenstein dieser Fleißarbeit waren die im Vergleich und Überblick erstmals digital verfügbaren Bildbestände.

Doch was macht man mit 100.000 Fotos, die außer einer Dateibenennung keinerlei Informationen mit sich bringen? Hier werden die eingangs aufgeworfenen Fragen konkret. Vergleichbar mit den Diskussionen um Distant Reading (Moretti 2007 und 2016; Crane 2006) stellt sich ob der Verfügbarkeit der Digitalisate die Frage, wie diese zielführend erschlossen werden können. Wie findet man relevante Bildbestände für eine Tiefenerschließung? Wo helfen die bestehenden Metadaten zielführend?

Maschinelle Unterstützung nutzen, aber wie?

Für diese Arbeitsschritte sind Methoden der Digital Humanities vielversprechend. Zwar wird zunehmend die Frage nach der Nutzung von Digitalisaten als Open Data für die Wissensproduktion[7] diskutiert, oft jedoch erst nach Abschluss der Erschließungsarbeiten. Nicht erst die publizierten Daten digitalen Kulturerbes können mit DH-Verfahren erforscht werden – diese sind bereits in der Erschließung enorm hilfreich. Zwei Verfahren aus dem Bereich des Machine Learning scheinen besonders erfolgsversprechend: maschinelle Bilderkennung sowie die Analyse bestehender Metadaten mittels Text Mining.

Die großen Mengen unerschlossener Fotos können mit maschineller Bilderkennung hinsichtlich ihrer Ähnlichkeit gruppiert werden, wie es etwa PixPlot realisiert:[8] Bildanordnungen im Vektorraum machen Schwerpunkte des Bestandes deutlich, außerdem lassen sich Subkorpora bilden, die mit Massenbearbeitung formal erschlossen werden können. In der Arbeitspraxis ist daneben das Identifizieren von Dubletten relevant: Wenn beispielsweise Abzüge des Archivs in der Vergangenheit abfotografiert wurden, existiert das Foto in unterschiedlichen Kopien im digitalisierten Bestand – manuell eine Suche nach der Nadel im Heuhaufen, automatisiert mit Bildvergleich gut zu identifizieren (vgl. die vielversprechenden Ansätze bei Schneider 2019). Auch ähnliche Aufnahmen, z.B. aus einer Bildserie, sind so zuzuordnen. Die inhaltliche Erschließung wird durch eine Ähnlichkeitssuche ebenfalls deutlich vereinfacht: Hat man etwa eine gute Aufnahme eines Gegenstandes, so lassen sich relativ eindeutig andere Abbildungen dessen im Bestand finden. Hier wäre jedoch eine menschliche Intervention (zumindest zu Beginn eines möglichen Active Learning-Verfahrens) aufgrund der feinen Unterschiede notwendig. Zudem sind zweifelsfrei die Trainingsdaten von großer Relevanz und sollten wo möglich aus bereits erschlossenen, vergleichbaren Kulturerbe-Datensätzen bestehen.

Text Mining-Verfahren würden in GLAM-Datenbanken nicht nur aus dem Museumsbereich präzisiere Suchabfragen und gerade in bestehenden Erfassungen systematische Funde ermöglichen. Ansätze wie facettierte Suchmasken mit Linked Open Data, wie sie in Plattformen zur Datenpräsentation zunehmend realisiert werden,[9] wären im Backend enorme Arbeitserleichterungen. Schon banale Automatisierungen wie Rechtschreibkorrektur und Vereinheitlichungen der Formalerschließung sind momentan in der Regel nicht vorgesehen. Sie kosten viel Zeit und Konzentration, sobald nicht simple Ersetzungen vorzunehmen sind. Gleichzeitig fehlt den Entwickler*innen der eingesetzten Datenbanken die zielgenaue Kollaboration mit entsprechender DH-Forschung zu konkreten Tools und Verfahren sowie der Testung von verschiedenen Funktionen für eine hohe Qualität der Ergebnisse, die vor der Übernahme in die Infrastruktur erfolgen muss. LOD wird zudem aktuell nur in Ausnahmen direkt in die Erfassungssysteme eingebunden – erst so könnte die Arbeit an Ontologien und konkreten Datensätzen gezielt verbunden werden. Dazu kommt die Notwendigkeit, die Erfassungen besser zu vernetzen; auch in Fällen, in denen dies erst nach der Veröffentlichung möglich oder notwendig wird. Hier sollte eine Öffnung für fortlaufende kollaborative Ergänzung und Korrektur von Daten in Verbindung mit der Erfassung von Paradaten (McIlvain 2013) geschaffen werden.

Viel Zeit wird mit der Erzeugung von metacrap (Doctorow 2001) verbracht. Die messy Metadaten, die für viele Erfassungen – oft im Backend, aber auch publiziert – bestehen, sind durch Tools einfach zu identifizieren und automatisch zu beheben: Museum Analytics[10] beispielsweise ermöglicht es, große Mengen von Museumsdaten zu analysieren, ist allerdings für publizierte Metadaten vorgesehen. Gerade in der Migration zwischen Datenbanksystemen sowie für den internen Gebrauch zwecks Qualitätskontrolle, Bestandssichtung und Entscheidung über Nachbearbeitungen vor der Veröffentlichung erlaubt dieses einen anderen Blick auf die Bestände. Das Tool Breve[11], das Tabellen visualisiert, könnte ergänzende Funktionen übernehmen. Die Entwicklungsvorhaben des Verbundprojektes GND4C[12] oder des Projekts Qrator[13] sind richtungsweisend, leider aber noch nicht verfügbar, und entsprechende Konferenzworkshops zu DH für Gedächtnisinstitutionen (Döhl/Voges 2019) erfreulich.

Erste Ansätze zur Nutzung von DH-Analyseverfahren werden auch von Museumsseite diskutiert, etwa hinsichtlich Netzwerkanalyse der Sammlungsbestände

(Werner 2019) oder Möglichkeiten der Visualisierung (Mayr/Windhager 2019), bilden dort jedoch (noch) die absolute Ausnahme. Dies spiegelt sich auch in den Programmen der entsprechenden Tagungen wie „Museums and the Internet“[14] oder denen der Fachgruppe Dokumentation des Deutschen Museumsbundes[15]. Falls entsprechende Ansätze bereits genutzt werden, so geschieht dies wiederum weitestgehend als „invisible work“ (s.o.) ohne Darstellung in der (Forschungs)Öffentlichkeit. In weiten Teilen der entsprechenden GLAM-Community wird gerade von Museumsseite die DH noch zu wenig als möglicher Kooperationspartner wahrgenommen, um entsprechende Workflows und Implementierungen zu konzipieren.

Die vorgestellten Zugänge könnten im Ergebnis der Implementierung nicht nur aufzeigen, welche Daten in einem Bestand enthalten sind, sondern auch, welche Leerstellen in der Erfassung noch geschlossen werden sollten. Von einer linearen Durchsicht, Überarbeitung und Freigabe der Datenbank-Einträge kann mit entsprechender Tool-Unterstützung – und einhergehender Interoperabilität! – zu einer gezielten Nachbearbeitung von Teilbeständen oder Vereinheitlichung einzelner Metadatenfelder übergegangen werden. So bleibt mehr Zeit für eine inhaltliche Erschließung und Analyse sowie die dringend notwendige epistemologischen Reflexionen dieser Prozesse.

Fazit: DH-Verfahren in Sammlungsdatenbanken

Was fehlt in der Gesamtschau momentan? Vor allem die Öffnung von Erschließungssystemen für die dargestellten Methoden sowie die Öffnung der entsprechenden Communities zueinander.

Erforderlich ist dabei der frühzeitige Einbezug von bestehenden Tools und Analyseverfahren – lange vor der Veröffentlichung der Datensätze. Gerade in der Exploration weitestgehend nicht erfasster Bestände zur Vorbereitung der Erschließung und in der Qualitätskontrolle von Metadaten liegen große Potentiale, die noch zu wenig genutzt werden. Wenn gleichzeitig bereits tiefenerschlossene Bestände als Trainingsdaten genutzt werden, können die Ansätze auch unabhängig von der Nutzung konkreter Datenbanken gegenseitigen Mehrwert sowohl in der Methodenentwicklung als auch in der Erschließung bringen.

Eine öffentliche Finanzierung und Weiterentwicklung von den entsprechenden Tools ist dabei dringend notwendig. Statt weiter in gewinnorientierte Software zu investieren, sollten Genossenschaften und Vereine gegründet und ausgebaut werden. Eine Unterstützung der Toolentwicklung durch entsprechend kompetente DHler*innen ist vielversprechend. So wäre etwa ein Hackathon zur Erweiterung von Erschließungssystemen eine Möglichkeit, um über das Tagesgeschäft hinausgehende Innovationen umzusetzen. Entsprechend erweiterte Datenbanken sollten viel häufiger auch in der Forschung verwendet werden, die aktuell noch viel zu oft in Form von Excel-Sheets (weiter-)arbeitet, obwohl Datenbanken mit erweiterten Funktionen existieren. Mit einfachen Import/Exportfunktionen innerhalb der Tools und Verbindungen zu Analyseverfahren könnten Forschungsumgebungen geschaffen werden, in denen kollaboratives Arbeiten gleichzeitig die Datensätze anreichert und Forschungsfragen beantwortet.

Fußnoten

1. Projektergebnisse unter https://alltagskulturen.lvr.de/. Das DFG-geförderte Projekt wurde durch die Autorin koordiniert, von Dagmar Hänel geleitet und maßgeblich auch durch den wissenschaftlichen Dokumentar im Projekt, Christian Baisch, vorangetrieben.
2. Das von Gertraud Koch und Heike Zinsmeister geleitete Verbundprojekt befragt DH-Perspektiven auf Möglichkeiten zur Modellierung hermeneutischer Prozesse. Vgl. https://www.herma.uni-hamburg.de/ .
3. Vgl. Sammlungsbeschreibungen unter https://alltagskulturen.lvr.de/de/sammlungen .
4. Faust Software, vgl. https://www.land-software.de/ .
5. Genutzt wird eine Weiterentwicklung von MediaFiler, vgl. https://mediafiler.com/en .
6. digiCULT.web als entitätsbasierte Online-Datenbank ist vorrangig für Museumsbestände entwickelt und baut auf CIDOC-CRM auf. Vgl. https://www.digicult-verbund.de/de/digicultweb . Zu LIDO vgl. http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/ .
7. Vgl. etwa die Schwerpunktsetzung „Open Data – now what?“ der Sharing is Caring-Konferenz 2019. http://sharecare.nu/programme/ . Danke an Samantha Lutz für den Hinweis.
8. Vgl. https://github.com/YaleDHLab/pix-plot . Vgl. auch Leonard 2019.
9. Vgl. etwa Suchfacetten der DDB, https://www.deutsche-digitale-bibliothek.de/ , mittlerweile auch der Europeana, https://www.europeana.eu/ .
10. https://www.max.gwi.uni-muenchen.de/ .
11. http://hdlab.stanford.edu/breve/ .
12. Ziel 3 des Projektes ist die „Bereitstellung von Schnittstellen und Werkzeugen zur Unterstützung nicht-bibliothekarischer Anwendungskontexte.“ Vgl. https://wiki.dnb.de/pages/viewpage.action?pageId=134055796 . Danke an Axel Vitzthum für den Hinweis.
13. Vgl. https://qurator.ai/ . GLAM-Institutionen mit digitalem Kulturerbe sind hier ein Anwendungsfall.
14. Bisher waren entsprechende Beiträge bei der Tagung absolute Ausnahme. Vgl. das Archiv unter https://mai-tagung.lvr.de/de/startseite.html .
15. Das Archiv der Tagungsprogramme und Vortragsfolien lässt nicht auf entsprechende Diskussionen schließen. Vgl. https://www.museumsdokumentation.de/?

lan=de&q=Who%20is%20who/FG%20Dokumentation%20im%20DMB/Tagungsarchiv .

Bibliographie

Crane, Gregory (2006): "What Do You Do with a Million Books?" In: D -*Lib Magazine* 12/3. http://www.dlib.org/dlib/march06/crane/03crane.html.

Doctorow, Cory (2001): "Metacrap. Putting the Torch to Seven Straw-Men of the Meta-Utopia." https://people.well.com/user/doctorow/metacrap.htm.

Döhl, Frédéric / Voges, Ramon (2019): „Erklärt und ausprobiert – Digital Humanities für Gedächtnisinstitutionen." Workshop im Rahmen der Tagung *Zugang gestalten 2019*. https://zugang-gestalten.org/dokumentation-2019/.

Koch, Gertraud (2006): „Die Neuerfindung als Wissensgesellschaft. Inklusionen und Exklusionen eines kollektiven Selbstbildes." In: Hengartner, Thomas; Moser, Johannes (Hg.): *Grenzen & Differenzen. Zur Macht sozialer und kultureller Grenzziehungen*. Leipzig, S. 545–559.

Leonard, Peter (2019): "Large images dataset overtime: PixPlot new features." In: *Culture Analytics Workshop: Time Series, Digital Humanities 2019*. https://dev.clariah.nl/files/dh2019/boa/1079.html und https://github.com/CultureAnalytics/DH2019.

Mayr, Eva / Windhager, Florian (2019): „Vor welchem Hintergrund und mit Bezug auf was? Zur polykontexturalen Visualisierung kultureller Sammlungen". Vortrag im Rahmen der Tagung *Objekte im Netz. Wissenschaftliche Sammlungen im digitalen Zeitalter*. Folien unter http://objekte-im-netz.fau.de/projekt/sites/default/files/2019-11/Mayr%26Windhager_PolyContext.pdf.

McIlvain, Eileen (2013): "Paradata. What is Paradata?" In: *NSDL Documentation Wiki*. https://wiki.ucar.edu/display/nsdldocs/Paradata.

Moretti, Franco (2007): "Graphs, Maps, Trees. Abstract Models for Literary History." London, New York.

Moretti, Franco (2016): "Distant Reading." Göttingen.

Schneider, Stefanie (2019): „Über die Ungleichheit im Gleichen. Erkennung unterschiedlicher Reproduktionen desselben Objekts in kunsthistorischen Bildbeständen." In: *DHd 2019. Digital Humanities im deutschsprachigen Raum 2019. Konferenzabstracts*, S. 92–94. https://doi.org/10.5281/zenodo.2596095.

Schönholz, Christian (2017): „Jede Kopie ein Original!. Aspekte eines kulturellen Größenverhältnisses." In: Koch, Gertraud (Hg.): *Digitalisierung. Theorien und Konzepte für die empirische Kulturforschung*. Konstanz/München 2017, S. 157–182.

Star, Susan Leigh / Strauss, Anselm L. (1999): "Layers of Silence, Arenas of Voice. The Ecology of Visible and Invisible Work." In: *Computer Supported Cooperative Work* 8/1-2, S. 9–30.

Werner, Claus (2019): „Die Sammlung als Graph. Gephi als Tool der Sammlungsevaluation. Deutsches Bergbau-Museum Bochum." Vortrag im Rahmen der Tagung *Museums and the Internet (Mai-Tagung) 2019*. https://mai-tagung.lvr.de/media/mai_tagung/pdf/2019/MAI-2019-Werner.pdf.