Der SSH Open Marketplace
Kontextualisiertes Praxiswissen für die Digital Humanities

Zarei, Alireza
alireza.zarei@gwdg.de
GWDG Göttingen

Seung-Bin, Yim
Seung-Bin.Yim@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Ďurčo, Matej
Matej.Durco@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Illmayer, Klaus
Klaus.Illmayer@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Barbot, Laure
laure.barbot@dariah.eu
DARIAH-EU

Fischer, Frank
frank.fischer@dariah.eu
DARIAH-EU; Higher School of Economics, Moskau

Gray, Edward
edward.gray@dariah.eu
DARIAH-EU

1. Was ist der SSH Open Marketplace?

Die internationale Digital Humanities-Gemeinde als "community of practice" hat bereits früh damit begonnen, Kataloge mit forschungsrelevanten Tools aufzubauen, um damit einen essentiellen Teil ihrer Praxis zu kartografieren. Diese Tool-Directorys "are frequently referred to as an important component of digital humanities infrastructure" (Dombrowski 2020).

Auf der Ebene einzelner Organisationen, die Übersichten ihrer eigenen und für ihre Stakeholder relevanten Tools anbieten, funktioniert das auch problemlos: DARIAH-DE etwa listet eigene Werkzeuge und Dienste auf (https://de.dariah.eu/en/dienste-und-werkzeuge), CLARIN-NL bietet einen Überblick über Werkzeuge der CLARIN-Infrastruktur (http://www.clarin.nl/node/404), DH Austria unterhält eine fokussierte, aber vielfältigere Liste (https://dha.acdh.oeaw.ac.at/en/know-more), es gibt eine französische Liste von Werkzeugen zur Korpusexploration (http://explorationdecorpus.corpusecrits.huma-num.fr/) und die Special Interest Group (SIG) der ADHO zu Digital Literary Stylistics betreibt einen einfachen Google-Spreadsheet, zu dem alle und jede*r beitragen können (https://dls.hypotheses.org/774).

Bei Projekten, die darauf abzielen, die gesamte Tool-Landschaft zu kartieren, ist es mit Listen nicht mehr getan. Datenbanken kommen zum Einsatz, womit auch die Kosten und der Wartungsaufwand steigen. Als Beispiele dienen das kanadische TAPoR (Text Analysis Portal for Research), das DiRT Directory (Digital Research Tools, vgl. Dombrowski 2014) oder TERESAH (Tools E-Registry for E-Social science, Arts and Humanities). Während es einen Konsens darüber gibt, dass diese Directorys von Nutzen sind, sind fast alle diese Projekte an einem Punkt gescheitert: am fehlenden Nachhaltigkeitskonzept, das sicherstellen würde, dass eine Plattform auch nach dem Auslaufen eines finanzierten Projekts weiter existieren kann und dafür Ressourcen bereitstehen (vgl. Barbot et al. 2020). Diese Diskrepanz zwischen mehr oder weniger erwiesener Nützlichkeit und langfristiger Unwartbarkeit hat Quinn Dombrowski das "Directory Paradox" genannt (Dombrowski 2021).

Im Rahmen des Horizon-2020-geförderten Projekts "Social Sciences & Humanities Open Cloud" (SSHOC), das eine Laufzeit von Januar 2019 bis April 2022 hat, wird der SSH Open Marketplace entwickelt, der einen Überblick nicht nur über digitale Tools und Services bietet, sondern auch über Trainingsmaterialien, Publikationen, Datensets und Workflows. Dabei werden diese untereinander kontextualisiert: Der Eintrag zu einem Tool verlinkt etwa Forschungspaper, die mithilfe dieses Tools entstanden sind, desweiteren passende Trainingseinheiten, zum Beispiel aus dem "Programming Historian", und, falls vorhanden, Forschungsdaten in einem für das Tool geeigneten Format. Die "Werkbänke der Digital Humanities" (Fischer et al. 2021) erscheinen auf diese Weise breit kontextualisiert. Das flexible Datenmodell und das Exponieren einer offenen API ermöglichen es, das gesammelte Praxiswissen für die Digital Humanities nutz- und erforschbar zu machen und eine neue, sich aktiv weiterentwickelnde Datenbasis dafür zu schaffen.

Der SSH Open Marketplace ist unter https://marketplace.sshopencloud.eu/ seit Januar 2022 als stabile Version verfügbar. Um den Inhalt der Plattform aktuell zu halten, gibt es ein Kurationskonzept, das die Community mit einschließt, sowie ein eigenes Arbeitspaket, welches eine nachhaltige Governance-Struktur für dieses Projekt erarbeitet, in deren Mittelpunkt die europäischen Forschungsinfrastrukturen DARIAH, CLARIN und CESSDA stehen. Der Marketplace soll nicht nur als praktische Hilfe im Forschungsalltag dienen, sondern auch helfen die Frage zu beantworten, welche Rolle Tools in der DH-Community eigentlich spielen, im Sinne einer "Tool Science" (vgl. Wolff 2015). Auch Lücken in der Softwareversorgung sollen so sichtbar werden. Alle Daten sind frei nachnutzbar, der Code steht unter einer Open-Source-Lizenz.

2. Das Datenmodell

Dem SSH Open Marketplace liegt ein umfassendes, aber pragmatisches Datenmodell zugrunde, das auf generische Konzepte baut (vgl. Barbot et al. 2019b). Zu den grundlegenden Eigenschaften des Modells gehören (siehe auch Abb. 1):

- die fünf genannten Hauptentitäten: Tools & Services, Training Materials, Publications, Datasets, Workflows (Abfolgen von Arbeitsschritten)
- flexibel typisierte Relationen zwischen den Einträgen (zur Kontextualisierung)

- die detaillierte Versionierung aller Änderungen (Einträge werden automatisiert auf Konsistenz geprüft, können aber auch händisch angelegt und kuratiert werden)
- Actors (etwa Autor*innen und/oder Programmierer*innen) werden als eigene Entitäten modelliert und mit eigenen Identifiern versehen (etwa ORCID)

Eine der zentralen dynamischen Eigenschaften des Datenmodells ist “activity”. Diese Eigenschaft klassifiziert die Einträge danach, in welcher Aktivität im Rahmen des Forschungsdaten-Lebenszyklus sie relevant sind. Die erlaubten Werte entsprechen dabei den Konzepten der TaDiRAH-Taxonomie (https://vocabs.dariah.eu/tadirah/), wie z.B. "Scanning" oder "POS-Tagging". Weitere Beispiele für dynamische Properties sind den bibliografischen Angaben entlehnte Attribute für Publications ("conference", "journal", "year").

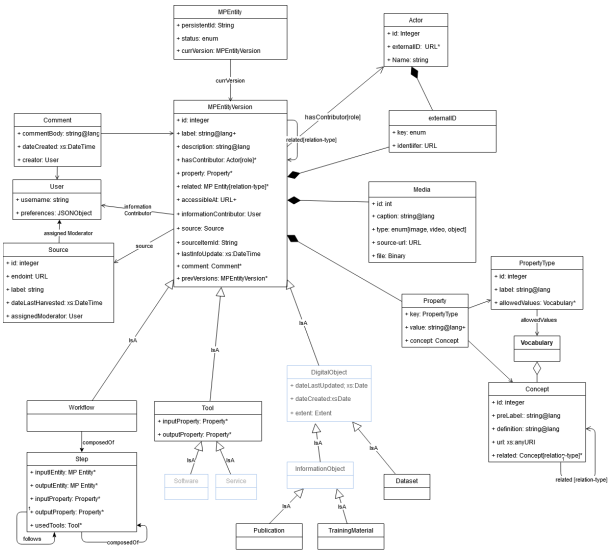

Abb. 1: Das Datenmodell des SSH Open Marketplace.

Die im Marketplace vorhandenen Daten werden über die API bereitgestellt, die online mithilfe von Swagger dokumentiert ist. Bis zum Projektende sollen die Daten auch im Sinne des LOD-Paradigmas im RDF-Format ausgeliefert werden. Als Target-Ontologie kommen die SSHOC Reference Ontology, Wikidata und Schema.org infrage, Mehrfachmappings sind denkbar und sinnvoll. Die Daten werden auch über einen SPARQL-Endpunkt abfragbar sein.

3. Überblick über die Inhalte des SSH Open Marketplace

Bisher (November 2021) haben über 5.000 Einträge ihren Weg in die Datenbank gefunden, die sich wie folgt auf die fünf Kategorien aufteilen:

- Tools & Services: 1671
- Training Materials: 321
- Publications: 2993
- Datasets: 305
- Workflows: 30

Anders als die Vorgängerprojekte, die oft von vorn angefangen haben, setzt der Marketplace darauf, bereits vorhandene Daten wiederzuverwenden, zu aktualisieren und anzureichern. Die häufigsten Quellen für die bezogenen Daten sind bisher:

- dblp computer science bibliography: 2837 Publikationen
- TAPoR: 1373 Tools
- SSK (Standardization Survival Kit): 29 Workflows und 370 Arbeitsschritte, letztere über die entsprechende Zotero-Bibliothek des SSK
- Humanities Data: 290 Datensets
- The Programming Historian: 83 Trainingseinheiten
- CLARIN Language Resource Switchboard: 56 Tools
- EOSC Marketplace: 15 Services
- SSHOC Service Catalogue: 13 Services

Dank unserer flexiblen Ingestion-Pipeline können zukünftig weitere Quellen eingespeist werden, geplant sind DARIAH Campus, die CLARIN Resource Families, das SSH Training Discovery Toolkit und SSHOC Training Material, die CESSDA Training Resources, Methodi.ca sowie die Daten nicht mehr gepflegter Projekte wie TERESAH.

4. Extraktionstask

Wie im vorangegangenen Kapitel beschrieben, speist sich der Marketplace aus verschiedenen Katalogen, deren Inhalte auf das Datenmodell gemappt und in die Kurationspipeline geschoben werden. Daneben werden aus dem Volltext wissenschaftlicher Publikationen und Übungsmaterialien erwähnte Tools extrahiert, um diese besser kontextualisieren zu können (Tool → mentionedIn → Publication). Forschungspapiere geben oft explizit Aufschluss über die Verwendung spezifischer Tools, Methoden und Datensätze, bieten daher Erfahrungswerte aus dem Forschungsalltag.

Die Systemarchitektur beinhaltet eine eigene "extraction"-Komponente, die in Volltexten von Publikationen Tools und Services identifiziert und mit entsprechenden Einträgen im Marketplace zusammenbringt und entsprechende Relationen anlegt.

Den Anfang bildete ein Experiment: die exemplarische Extraktion von Tools, die in den Beiträgen der jährlichen ADHO-Konferenzen erwähnt werden. Dafür wurde ein eigenes Kommandozeilenwerkzeug namens ToolXtractor entwickelt, das auf dem Erkennen von Zeichenketten basiert, die einer Positivliste entnommen werden (vgl. Barbot et al. 2019a und Fischer/Moranville 2020). Basierend auf der Tool-Liste des TAPoR-Projekts wurden in den Proceedings der Konferenzjahre 2015–2019 insgesamt 1.498 Erwähnungen gezählt, die auf 238 individuelle Tools zurückgingen. Die 15 am häufigsten genannten Tools im gewählten Korpus waren Gephi, Omeka, stylo, MALLET, Excel, D3.js, NLTK, WordPress, Drupal, TextGrid, CollateX, GeoNames, TXM, Solr und die Voyant Tools.

Nach diesen ersten Einblicken in die Trends der Tool-Nutzung innerhalb der DH-Forschung haben wir unseren Ansatz erweitert, um auch bisher noch nicht katalogisierte Tools zu finden. Wir haben dafür einen Datensatz geistes- und sozialwissenschaftlicher Publikationen entsprechend annotiert und mittels Transfer-Learning ein eigenes NER-Modell (Named Entity Recognition) trainiert, das Ergebnisse mit hoher Präzision und hohem Recall liefert.

Innerhalb der Extraktionspipeline (Abb. 2) werden über die API des SSH Open Marketplaces zunächst die bereits erfassten Publikationen abgerufen. Das erwähnte NER-Modell wird dann auf jeden Satz der entsprechenden Volltexte angewendet und liefert eine

Liste möglicher Tools zurück. Es wird überprüft, ob es für diese Tools bereits Einträge im Marketplace gibt; in diesem Fall wird eine Relation zwischen dem Tool und der Publikation hinzugefügt. Alle anderen extrahierten potenziellen Tools, die noch keinen Eintrag im Marketplace haben, werden in die Kurationspipeline eingespeist, wo sie entsprechend der Richtlinien bearbeitet werden.

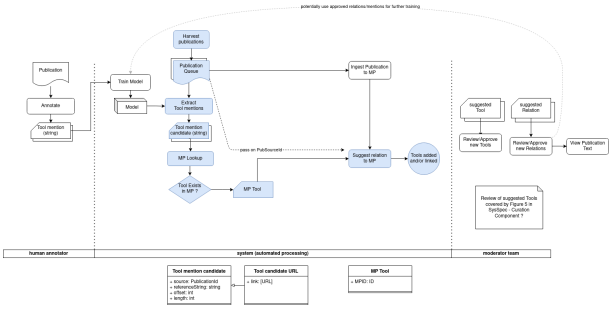

Abb. 2: Extraktionspipeline.

Eine Herausforderung bei der Datenextraktion stellen die verschiedenen Datenformate von DH-Publikationen dar. Je nach Konferenz bzw. Publikationsorgan und Jahr der Veröffentlichung findet sich ein bunter Mix aus PDF-, TEI- und HTML-Dateien. Teils sind Dateien aus bestimmten Konferenzjahren gar nicht mehr in zitierbarem Zustand verfügbar (etwa die Paper der DH2015 in Sydney), ein Missstand auch im Sinne des Konferenzmottos "Kulturen des digitalen Gedächtnisses". Ein weiteres Problem stellen Journale mit beschränktem Zugriff dar, etwa *Digital Scholarship in the Humanities* (DSH).

5. Ausblick

Der SSH Open Marketplace baut auf den Erfahrungen vieler Vorgängerprojekte auf. Durch Nutzerbefragungen und frühes Community Engagement haben wir versucht, eine inklusive Plattform zu schaffen. Die Daten sind aber auch maschinenlesbar und mit anderen Linked-Data-Projekten verknüpft, sodass der Marketplace Teil eines größeren Ökosystems ist, aus dem neue Impulse und Daten kommen. Es wird in Zukunft darauf ankommen, die Entwicklungen im Feld der Digital Humanities genau zu beobachten und zusätzliche relevante Quellen in die Ingestion-Pipeline aufzunehmen, die das Fach disziplinär weiter begreifen (vgl. dazu den Aufsatz von Luhmann/Burghardt 2021).

Der Marketplace kann und soll nicht andere Arten von Repositorien ersetzen, etwa OpenAIRE oder Dataverse, es ist keine Hosting-Plattform für Daten oder Forschungspaper. Der Fokus liegt auf der Kontextualisierung: Es werden Inhalte bevorzugt, die eine wertige Relation zwischen Tools & Services, Training Materials, Publications, Datasets und Workflows herstellen.

Im Idealfall kann der Marketplace dabei helfen, die Rolle dieser Hauptentitäten zu verdeutlichen und zu stärken. Die Bedeutung offen zugänglicher Datensets etwa, die den FAIR-Prinzipien genügen, ist innerhalb der Digital Humanities immer noch gering, was sich unter anderem darin zeigt, dass die einzige dedizierte Sammlung von DH-relevanten Datensets das Privatprojekt eines einzelnen Forschers ist (Humanities Data, https://humanitiesdata.com/).

Fördernachweis

Der SSH Open Marketplace wird vom Europäischen Forschungsrat (ERC) im Rahmen des Forschungs- und Innovationsprogramms Horizon 2020 (Fördervereinbarung Nr. 823782) entwickelt.

Bibliographie

Barbot, Laure / Fischer, Frank / Moranville, Yoann / Pozdniakov, Ivan (2019a): "Which DH Tools Are Actually Used in Research?" In: *weltliteratur.net*, 6. Dezember 2019. (URL: https://weltliteratur.net/dh-tools-used-in-research/)

Barbot, Laure / Moranville, Yoann / Fischer, Frank / Petitfils, Clara / Ďurčo, Matej / Illmayer, Klaus / Parkoła, Tomasz /Wieder, Philipp / Karampatakis, Sotiris (2019b): *SSHOC D7.1 System Specification – SSH Open Marketplace* (Version 1.0). Zenodo.

Barbot, Laure / Dombrowski, Quinn / Fischer, Frank / Rockwell, Geoffrey / Spiro, Lisa (2020): "Who Needs Tool Directories? A Forum on Sustaining Discovery Portals Large and Small." In: *DH2020: »carrefours/intersections«. 22–24. Juli 2020. Book of Abstracts*. University of Ottawa. (URL: https://dh2020.adho.org/wp-content/uploads/2020/07/126_WhoneedstooldirectoriesAforumonsustainingdiscoveryportalslargeandsmall.html)

Dombrowski, Quinn (2014): "What Ever Happened to Project Bamboo?" In: *Literary and Linguistic Computing*, Vol. 29, Issue 3, September 2014, S. 326–339, *doi:10.1093/llc/fqu026* .

Dombrowski, Quinn (2021): "The Directory Paradox." In: Anne McGrail et al. (Hg.): *Debates in the Digital Humanities: Institutions, Infrastructures at the Interstices*. University of Minnesota Press (erscheint demnächst).

Fischer, Frank / Moranville, Yoann (2020): "DH Tools Mentioned in 'The Programming Historian'." In: *weltliteratur.net*, 17. Jan uar 2020. (URL: https://weltliteratur.net/dh-tools-programming-historian/)

Fischer, Frank / Burghardt, Manuel / Luhmann, Jan / Barbot, Laure / Moranville, Yoann / Zarei, Alireza (2021): "Die Werkbänke der Digital Humanities: Zur Rolle von Tools und Software für die Forschungsarbeit." In: *vDHd2021: "Experimente"*, Zenodo, *doi:10.5281/zenodo.4639228* .

Luhmann, Jan / Burghardt, Manuel (2021): "Digital humanities – A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape." In: *Journal of the Association for Information Science and Technology*, 1–24, doi:10.1002/asi.24533.

Wolff, Christian (2015): "The Case for Teaching 'Tool Science'. Taking Software Engineering and Software Engineering Education beyond the Confinements of Traditional Software Development Contexts." In: *2015 IEEE Global Engineering Education Conference (EDUCON)*, Tallinn, pp. 932–938, *doi:10.1109/EDUCON.2015.7096085* .