Computational Literary Studies Data Landscape Review

Börner, Ingo
ingo.boerner@uni-potsdam.de
Universität Potsdam

Charvat, Vera Maria
VeraMaria.Charvat@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)

Ďurčo, Matej
Matej.Durco@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)

Mrugalski, Michał
michal.mrugalski@hu-berlin.de
Humboldt-Universität zu Berlin

Odebrecht, Carolin
carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin

Literarische Werke und deren digitale Repräsentationen stellen auch in den Fachbereichen der Computational Literary Studies (CLS) das Fundament für epistemische Auseinandersetzungen und Diskurse. Die unterschiedlichen Prozessierungen und Visualisierungen wie digitale Editionen (Sahle 2013) oder Netzwerkanalysen (Trilcke 2013) von literarischen Werken aller Gattungen (Epik, Drama, Lyrik) erzeugen eine Vielzahl an heterogenen Daten, die immer flexibler und umfassender miteinander in Interaktion treten und kommunizieren. Diese Entwicklung stellt die Frage der Interoperabilität der Daten in den Mittelpunkt, wobei Linked Open Data (LOD; Heath & Bizer 2011) eine zentrale Rolle spielen.

Das übergeordnete Ziel von "Computational Literary Studies Infrastructure"[1] – ein von der EU finanziertes "Integrating Activitie for Starting Communities (IASC)"-Projekt – ist die Schaffung eines einheitlichen und einfachen Zugangs zu den besten europäischen und nationalen Infrastrukturen für die CLS-Community. In unserem Arbeitspaket *Data Selection and Curation* möchten wir Informationen über literarische Werke systematisch für die CLS-Community kompilieren, aufbereiten und konsolidieren, um die Zugangsparadigmen für literarische Daten signifikant zu rekonfigurieren und die Einhaltung der FAIR-Prinzipien (findable, accessible, interoperable, reusable; Wilkinson et al. 2016) erheblich zu verbessern.

Um die Auffindbarkeit und den forschungsorientierten Zugang zu literarischen Daten für die CLS-Community zu ermöglichen, ist eine Inventarisierung der CLS-Datenlandschaft erforderlich, die forschungsrelevante Kriterien für die Datenauswahl sowie deren Erfassung und Beschreibung anwendet. Mit dieser Inventarisierung, die wir in Form einer *Data Landscape Review* durchführen, kann die vorhandene Datenlandschaft als digitales Erbe für CLS-Kontexte erst umfassend sichtbar und als Vorlage für weitere Forschungsvorhaben zugänglich gemacht werden.

Dabei stellen wir uns unter anderem folgende Fragen: Welche Beschreibungsmerkmale sind für die Daten als Kollektion im Sinne einer eigenen epistemischen Einheit wesentlich? Welche Beschreibungsmerkmale sind in Bezug auf die literarischen Vorlagen und deren Aufbereitungen wichtig? Wie kann nach Kollektionen oder einzelnen Datensätzen im Sinne der *programmable corpora* (Fischer et al. 2019) recherchiert werden?

Die Ergebnisse unserer *Data Landscape Review* werden wir als Posterpräsentation mit Fokus auf den technischen Bericht zur Kartierung und Kontextualisierung der CLS-Daten vorstellen.

Aufbauend auf der Review werden die Ergebnisse in Form eines stetig wachsenden, interaktiven Online-Katalogs literarischer Corpora für die CLS-Community bereitgestellt. Dieser wird eine umfassende Übersicht über die verfügbaren Ressourcen inklusive ausführlicher beschreibender Metadaten liefern und die üblichen Abfrage- und Erschließungsmöglichkeiten mittels verschiedener Such- und Filtermechanismen bieten. Konzeptueller Ausgangspunkt für die strukturierte Sammlung der Informationen ist das Metamodell für Korpusmetadaten (MKM; Odebrecht 2018) – ein, generisches erweiterbares Beschreibungsmodell, für die zentralen Entitäten *Korpus*, *Dokument* und *Annotation* sowie ihre Beziehungen untereinander.

Während das Modell selbst abstrakt definiert ist, erarbeiten wir eine kongruente/entsprechende Ontologie im OWL-Format (OWL, 2012), welche eine Repräsentation der Daten in RDF (Resource Description Framework)[2] ermöglicht. Die Formalisierung als OWL-Ontologie gestattet darüber hinaus auch, Äquivalenzen zu bereits bestehenden Ontologien und Schemata im Sinne des LOD-Paradigmas explizit zu machen. Hier sind insbesondere Ansätze zur Text- und Publikationseinordnung wie FRBR (IFLA, 1998) und Dublin Core (ISO standard 15836) zu nennen. Neben Äquivalenzen auf der Schema-Ebene wird der Datensatz um Verweise/Verlinkungen zu externen Referenzressourcen wie zum Beispiel die Normdateien GND (Gemeinsame Normdatei)[3], VIAF (Virtual International Authority File)[4], WikiData[5] und GeoNames[6] angereichert. Diese sind unabdingbar, um semantische Interoperabilität zwischen Datensätzen herzustellen. Die in RDF serialisierten Daten werden selbstverständlich regelmäßig als geschlossener Datensatz ("Dump"), sowie über einen SPARQL[7]-Endpoint verfügbar gemacht. Die Ontologie sowie eine erste proof-of-concept Version des Online-Katalogs werden wir bei der Tagung präsentieren.

Ebenso ist zu berücksichtigen, dass dieser Katalog Teil von einem komplexen Gefüge an Ressourcen, Providern und Disseminationskanälen bzw. Aggregatoren ist. Die Position des Katalogs in diesem Gefüge und seine Beziehung zu verwandten Aggregatoren wie CLARIN VLO (Virtual Language Observatory)[8], Europeana[9] oder OpenAIRE (Open Access Infrastructure for Research in Europe)[10] müssen noch im Detail erarbeitet werden. Der grundlegende Ansatz wird dabei aber sein, die Information über mehrere Kanäle möglichst breit zu streuen/disseminieren und dafür auch Mappings der Metadaten in die erforderlichen Metadaten-Formate, wie CMDI (Component Metadata Initiative)[11] für VLO bzw. EDM (Europeana Data Model)[12] für Europeana bereitzustellen.

Die *Data Landscape Review* und der Online-Katalog werden den Forschenden Zugriff zu einer breiten Palette an Ressourcen, die über mehrere europäische Anbieter distribuiert sind, ermögli-

chen und mit Beschreibungen und Informationen auch einen umfassenden, domänenspezifischen Überblick über diese Ressourcen bieten.

Fußnoten

1. Website des Projekts CLS INFRA (No. 101004984): https://clsinfra.io/ (letzter Zugriff 24.11.2021)
2. RDF W3C Recommendation: https:/www.w3.org/TR/rdf-primer/ (letzter Zugriff 24.11.2021)
3. GND: https://gnd.network (letzter Zugriff 24.11.2021)
4. VIAF: http://viaf.org/ (letzter Zugriff 24.11.2021)
5. Wikidata: https://www.wikidata.org/ (letzter Zugriff 24.11.2021)
6. GeoNames: http://www.geonames.org/ (letzter Zugriff 24.11.2021)
7. SPARQL steht für SPARQL Protocol and RDF Query Language; SPARQL W3C Recommendation: https://www.w3.org/TR/sparql11-overview/ (letzter Zugriff 24.11.2021)
8. CLARIN VLO: https://vlo.clarin.eu/ (letzter Zugriff 24.11.2021)
9. Europeana: https://www.europeana.eu/de (letzter Zugriff 24.11.2021)
10. OpenAIRE: https://www.openaire.eu/ (letzter Zugriff 24.11.2021)
11. CMDI: https://www.clarin.eu/cmdi (letzter Zugriff 24.11.2021)
12. gesammelte Dokumentationen zum EDM: https://pro.europeana.eu/page/edm-documentation (letzter Zugriff 24.11.2021)

Bibliographie

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama", in: Fischer, Frank / Akimova, Marina / Orekhov, Boris (eds.): *Digital Humanities 2019. Conference Abstracts*, Utrecht University, Moscow, https://dev.clariah.nl/files/dh2019/boa/0268.html

Heath, Tom / Bizer, Christian (2011): "Linked Data: Evolving the Web into a Global Data Space", 1st edition, in: *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 1, No. 1 [San Rafael, Calif.]: Morgan & Claypool, S. 1-136, doi: https://doi.org/10.2200/S00334ED1V01Y201102WBE001

IFLA (1998): "Functional Requirements for Bibliographic Records: Final Report", in: *IFLA Series on Bibliographic Control* 19 (former UBCIM). München: K.G. Saur Verlag.

ISO standard 15836 (2017): "The Dublin Core Metadata Element Set"

Odebrecht, Carolin (2018): "MKM – ein Metamodell für Korpusmetadaten. Dokumentation und Wiederverwendung historischer Korpora", Dissertation. Humboldt-Universität zu Berlin, Sprach- und literaturwissenschaftliche Fakultät, Berlin. doi: https://doi.org/10.18452/19407

Sahle, Patrick (2013): "Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels", 3 Bände, Norderstedt: Books on Demand, in: *Schriften des Instituts für Dokumentologie und Editorik*, Bände 7-9.

Trilcke, Peer (2013): "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft", in: Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds): *Empirie in der Literaturwissenschaft*. Paderborn: Mentis 201–247.

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand J. / et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship". Sci Data 3, 160018. doi: https://doi.org/10.1038/sdata.2016.18

W3C OWL Working Group (2012): Web Ontology Language (OWL), https://www.w3.org/OWL/