Bericht der ThULB als für AP3 verantwortlich zeichnender Projektpartner

Rückblick

Nach einem sehr aktiven Auftakt mit einem Workshop zur gewünschten Toolbox-Funktionalität im März 2022 und mehreren Treffen zu den Entwicklungsfortschritten und den neu implementierten Features der Toolbox im Jahresverlauf blieb insbesondere ab Jahreswechsel 2022/2023 Feedback aus dem Projektteam etwa zum User Interface der Toolbox, zur Qualität der Matchingergebnisse, der Umsetzung gewünschter Features oder erforderlichen Anpassung an bei den Partnerinstitutionen bestehende (Interims-)Workflows weitestgehend aus. Die digiCULT-Verbund eG – in Phase II nunmehr externer Impulsgeber – beteiligte sich vor allem in 2022 an der Weiterentwicklung des Personenmatchings zu einem funktionalen Werkzeug. Parallel wurden in AP3 verschiedene Strategien einer nachhaltigen Datenbankanbindung der Toolbox erprobt. Während ein gemeinsam mit der DNB angedachtes neo4j-Datenbackend wegen fehlender Personalressourcen an der DNB nicht aufgesetzt wurde, konnte an der ThULB die auf kulturhistorische Forschungsprojekte fokussierte und inzwischen als Infrastruktur in NFDI4Memory integrierte Wikibase-Instanz FactGrid getestet werden. Dieses Teilprojekt wurde vom FactGrid-Initiator und Community-Manager Olaf Simons betreut und umfassend dokumentiert. Die probeweise Etablierung von GND-Redaktionsworkflows für FactGrid-Daten an der ThULB hat deutlich gemacht, dass die vorhandene Datenqualität und das differenzierte Datenmodell sehr gute Voraussetzungen für GND-Arbeit schaffen, zumal Letzteres ein problemloses Mapping auf die GNDO erlaubte (s. https://tinyurl.com/2o73j3dt). Die dauerhafte Anbindung von FactGrid an die ThULB scheitert derzeit noch an einer fehlenden Strategie für eine technische Betreuung der Wikibase-Instanz, die von der ThULB-IT nicht geleistet werden kann und derzeit über einen jährlich erneuerten Werkvertrag abgedeckt wird (s. auch oben verlinkte Dokumentation).

Obgleich die Personen-Testdatensets aus Phase I in der Toolbox spätestens ab Jahresbeginn 2023 zuverlässig verarbeitet und bewertet werden konnten, wurden damit keine Tests oder Analysen seitens der Partner durchgeführt, mutmaßlich weil die Daten inzwischen veraltet und nicht mehr relevant waren. Zugleich wurden aber auch keine neuen Daten für Tests angeliefert. Für Bauwerke konnte mit Unterstützung des DDK ein erstes Matching implementiert werden, die Feedbackintervalle blieben aber zu lang, um den erwünschten Fortschritt zu ermöglichen, was im Jahresverlauf 2023 mehrfach im Projektteam angesprochen wurde – leider ohne Effekt. Völlig unberücksichtigt blieben die Entitätstypen Geografika und Sachbegriffe, für die es im Projektkontext offenbar keine Anwendungsfälle gab. Ein iteratives Vorgehen, wie es für AP3 lt. Antrag vorgesehen war, erwies sich mit diesen Rahmenbedingungen als nicht umsetzbar. Angelegt entsprechend den Wünschen der Partnerinstitutionen als eine flexible Web-Umgebung zur eigenständigen Installation, Betreuung und Erweiterung durch GND-Agenturen (lokal oder auf einem Server) ist eine weitere Nutzung durch diese Agenturen zumindest im GND4C-Scope vorerst nicht absehbar, da keine Erfahrungen mit der Anwendung gesammelt wurden.

Die universelle Toolbox als uneinlösbares Versprechen

Davon unabhängig bleibt das Kernproblem der Entwicklung generischer Werkzeuge wie der Toolbox für die GND-Arbeit bestehen: Die Prozesse sind in den einzelnen Institutionen zu individuell und zu komplex, die Daten zu divers, um mit einem einheitlichen Softwarewerkzeug in einer Oberfläche "as-is" bearbeitet werden zu

können. In praktisch allen Bearbeitungsschritten sind in den in GND4C praktizierten Workflows Anpassungen und Eingriffe je Datenlieferung nötig, um zu einer fehlerfreien Verarbeitung und aussagekräftigen Resultaten und validen Transformationen zu gelangen. So wünschenswert eine integrierte GND4C-Toolbox für alle redaktionsrelevanten Vorgänge auch ist, so wenig realistisch ist es, die in den Quell- und Zieldaten verankerte Komplexität und Vielgestaltigkeit in fest verdrahtete, einheitliche Prozesse zu gießen, ohne das für jeden Entitätstyp hochspezifische GND-Regelwerk zu ignorieren und in der Vereinheitlichung die Daten soweit zu reduzieren, dass das Resultat kaum noch (etwa für ein zuverlässiges Matching oder gar eine Einspielung) nutzbar ist.

Die überzogenen Erwartungen an eine alle Herausforderungen der GND-Redaktion technisch adressierende GND-Toolbox sind ein Erbe aus Phase I von GND4C, in dem sich die Projektpartner auch in Phase II nicht völlig frei machen konnten – was auch für die Arbeit an der Toolbox in AP3 gilt. Das für die Entwicklung eines generischen Werkzeugs nötige, umfassende Verständnis für die zu partikularisierenden Prozesse der GND-(Redaktions-)Arbeit ist nun bei den Agenturpartnern nach zwei Phasen GND4C in einem Maße vorhanden, dass ein solches Vorhaben mit den entsprechenden Ressourcen und in enger Zusammenarbeit mit aktiven, Daten regelmäßig prozessierenden GND-Agenturen umsetzbar wäre. Eine solche Toolbox würde verstanden als eine Umgebung für individuell anpassbare Skripte, die beliebig verschaltet und verkettet werden können, um die unterschiedlichen Quelldaten, Zielsysteme (Abgleiche mit der GND müssen über andere Systeme wie Wikidata, Geonames, OSM angereichert und validiert werden), Entitätstypen und Zielformate (neben GND-MARCXML z. B. die Ausgangsformate der Lieferinstitutionen oder LIDO für einen Rückimport) zu bedienen.

Vision

Eine Python-basierte Umgebung, wie sie als Toolbox in Projektphase II entwickelt wurde – zur Verwaltung der Datensets in einem generalisierten, leistungsfähigen Interims- und Transferformat wie EntityXML (das im Übrigen mit Text+ von einem externen Partner und erst während Phase 2 für einen anderen Anwendungsfall entwickelt wurde) zum Management der Skripte, der Verschaltung und Kontrolle der Abläufe und der Generierung und Validierung von Exporten – ist dafür der ideale Ausgangspunkt wegen der guten Lesbarkeit der Scripte und der leichten Zugänglichkeit auch für Programmier-Anfänger*innen.

Kombiniert werden sollte ein solches Abgleichs- und Data-Wrangling-Werkzeug mit einer offenen Datenbank zur Verwaltung und Publikation von GND-Kandidaten und nicht-GND-geeigneten Datensätzen, die oft den Großteil einer Datenlieferung ausmachen und für die von den Lieferinstitutionen Identifikatoren benötigt werden. Eine etablierte Wikibase-Instanz wie FactGrid wäre wegen der hohen Transparenz, intensiven Nutzung und unkomplizierten Datenpflege an der ThULB dafür das Werkzeug der Wahl. Zudem ist die GND im Datenmodell von Anfang an mitgedacht, was die Instanz zu einem hervorragenden Instrument zur Vorbereitung und Zugänglichmachung von GND-Kandidaten (zur entsprechenden Markierung existiert ein eigenes Property) bzw. (noch) nicht für die GND geeigneten Datensätzen macht.