Vernetzen und Verknüpfen
Linked Data in Text+

Text+
Sammlungen - Lexikalische Ressourcen - Editionen
Infrastruktur / Betrieb

Lexikalische Ressourcen

Ausgangslage

Die beteiligten Institutionen haben umfangreiche Datenbestände, die häufig in XML-basierten Formaten vorliegen. Für diese Formate existieren teilweise bereits Konvertierungsroutinen in geeignete LOD-Ontologien (z.B. OntoLex); die öffentliche Bereitstellung von Daten in diesen Formaten ist allerdings selten. Bzgl. konkreter Erfahrungswerte kann u.a. auf existierende Projekte (z.B. ELEXIS) zurückgegriffen werden.

Erste Schritte

→ Erarbeitung eines Working Papers zur Übersicht über relevante Ontologien, Vokabulare und Wissensbasen
→ Bestandsanalyse von Ressourcen, die bereits in LOD-Formaten vorliegen oder in absehbarer Zeit zur Verfügung gestellt werden können.

Perspektive

Viele Ressourcen der beteiligten Institutionen werden im Rahmen der LLOD-Cloud angeboten und mit zentralen Hubs (GND, Wikidata) semantisch verlinkt. Dabei wird unter anderem ein starker Fokus auf dem Bereich der Datenprovenienz, Datenqualität und dem rechtlichen Status von Ressourcen liegen.

Editionen

Ausgangslage

Die Dokumentstrukturierung insbesondere auf Basis von TEI/XML und die Anreicherung von Editionsdaten anhand von Normdaten und kontrollierten Vokabularen ist im Feld digitaler Editionen geübte Praxis: Auszeichnung von Personen und Orten und ihre eindeutige Identifikation, aber auch Organisationen, Institutionen, Ereignisse, Zeitangaben. Zahlreiche digitale Angebote (Geistes- und Kulturwissenschaften, GLAM) exportieren GND-BEACON-Dateien, werten solche aus und integrieren sich so in eine vernetzte Rechercheland-schaft. Participants der Task Area Editionen sind hieran beteiligt, bieten zudem Services an wie STI Linked Data Service, Beaconfinder (MWW) oder correspSearch (BBAW).

Erste Schritte

→ Taskforce zur Bestandsaufnahme und Workshops zur Vermittlung des oft noch "hinter den Kulissen schlummernden" LOD-Potenzials
→ Einbringen der LOD-Perspektive als Querschnittsthema in alle Arbeiten der Datendomäne
→ Mitwirkung in der Text+ LOD–AG sowie am Text+ GND-Forum.

Perspektive

Stärkung der Datenperspektive der vorhandenen Ressourcen, Verfügbarmachung des "Domänenwissens" für Entity Recognition und weitergehende Analyseansätze.

Sammlungen

Sammlungen sind auf der Ebene bibliografischer Daten häufig bereits Teil der LOD-Cloud. Objektdaten liegen aber in sehr unterschiedlichen Datenformaten vor.

Ausgangslage

Erste Schritte

→ Deskriptive Metadaten zu Sammlungen werden semantisch weiter angereichert und über RDF/SPARQL zur Verfügung gestellt.
→ Bisher unannotierte Texte aus dem Text+ Korpus müssen mit maschineller Tiefenerschließung (NERD, Topic Identification, andere disziplinäre Annotationen) semantisch zugänglich gemacht und auf Textebene mit der LOD-Cloud vernetzt werden.

Perspektive

Es sollen Prozesse zur Integration weiterer Sammlungen in die LOD-Cloud etabliert werden.

Linked Open Data (LOD) ist die offene Bereitstellung und Verknüpfung strukturierter Daten und Metadaten, mit dem Ziel, semantische Abfragen zu ermöglichen. Gerade Text+ mit seiner Vielzahl an verteilten, heterogenen Datenbeständen bietet hierfür erhebliches Potenzial, um Datenbestände interoperabel zu gestalten und für eine effiziente Nachnutzung bereitzustellen. Dies erfolgt durch die Nutzung etablierter LOD-Formate in Verbindung mit gemeinsam genutzten Ontologien, kontrollierter Vokabulare sowie durch die Verknüpfung bestehender Daten mit Normdaten. Unter anderem dient die Gemeinsame Normdatei (GND) hierbei als semantisches Werkzeug.

Infrastruktur / Betrieb

Ausgangslage

Ein zentrales, datendomänenübergreifendes Aufgabengebiet sind die Interoperabilität und Wiederverwertung von Ressourcen und deren Verwendung in LOD-Kontexten. Dafür werden geeignete Infrastrukturen geschaffen bzw. erweitert, und eine dedizierte GND-Agentur eingerichtet. Die Community wird u. a. durch Workshops und die Veröffentlichungen von Leitfäden unterstützt.

Erste Schritte

→ Die Koordination einer datendomänenübergreifenden Zusammenstellung vorhand-ener LOD-Ressourcen wurde initiiert, um eine bessere Übersicht auf das heterogene, breite Feld zu ermöglichen.
→ Das erste GND-Forum Text+ wurde erfolgreich durchgeführt, als erstes Angebot zum Austausch über und der Mitgestaltung an der GND.

Perspektive

Zentral ist die Etablierung gemeinsamer Modelle mit den drei Datendomänen. Auf technischer Seite wird eine Verknüpfung mit existierenden Infrastrukturkomponenten benötigt. Weiterführende Foren und Workshops sollen Forschenden auch außerhalb von Text+ leichtere Anknüpfungspunkte ermöglichen.

Facetten und Herausforderungen von Daten und Metadaten als LOD

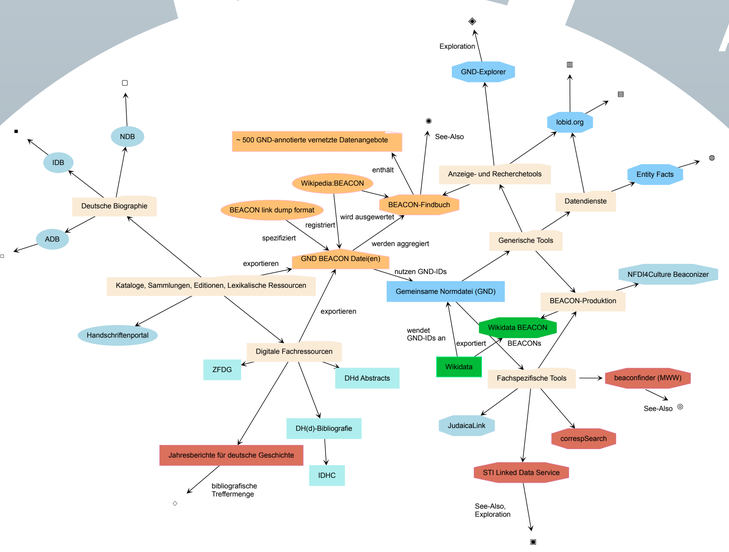
Repräsentation in RDF
- RDF-Konverter
- JSON-LD in statischem HTML
- Harvesting über SPARQL
- Standardisierung / Harmonisierung / Mapping von Ontologien und Vokabularen

Normdatenreferenzen
- Normdatenprovider
- Entitäten-Erkennung & -Disambiguierung
- Sichtbarkeit & Auffindbarkeit semantischer Verknüpfungen
- Qualitätsmanagement

Ressourcenverknüpfung
- Registraturen
- Bestands- und Kompatibilitätssichtungen
- Mapping von Ontologien und Vokabularen

- Wie bettet sich Linked Data in die Text+ Suchinfrastruktur ein
- Was soll die AG LOD konkret leisten
- Welche Schulungsmaßnahmen zu Linked Data werden benötigt
- Wie sichern wir durchgehende Qualitätsstandards
- Welche rechtlichen Einschränkungen sind zu beachten
- Welchen Mehrwert erwartet Text+ von Linked Data
- Wie vernetzt sich Text+ mit anderen NFDI-Konsortien
- Welche LD-Ressourcen werden in Text+ angeboten

Aktuelle Maßnahmen



Wie lässt sich die bottom-up gewachsene "BEACON-Landschaft", die auf Basis von GND-Annotationen domänenübergreifend bibliografische Daten, Nachschlagewerke, Digitalisate, Editionen und auch Sekundärliteratur online vernetzt bzw. vernetzbar macht, aufgreifen und in eine stabile und qualitätsgesicherte LOD-Infrastruktur integrieren?

Zur Sichtung der vielseitigen LOD-Ressourcen und Formate beteiligter Einrichtungen wird fortlaufend eine tabellarische Bestandsdokumentation erarbeitet, um die Grundlagen für Standardisierungs-, Harmonisierungs- und Mappingmaßnahmen zu legen, Handlungsempfehlungen zu erarbeiten und Datenbestände interoperabel in die LOD-Cloud zu integrieren.

	A	B	C	D	E	F	G	H	I	J	
2	**Partner**	**Kontaktdaten (E-Mail)**	**Name der Ressource**	**Derzeitiges Dateiformat**	**(P)ID für die Ressource**	**PID Schema (Handle, DOI, URN, eigenes System, keins)**	**Anzahl der Entitäten**	**Sprache(n)**	**Zugriffsrechte (Vollzugriff, Lizenz/Beschränkt, akademische Nutzung, kein Zugriff, unbekannt)**	**Bis wann als LOD?**	**Geplante LOD-Ontologie**
	DNB	r.pollak@dnb.de	DNB-Titeldaten	Erfassungsformat: Pica	d-nb.info/id	eigenes System (Permalinks DNB)	ca. 24 Millionen	Katalogisierungssprache: deutsch	Vollzugriff	produktiv	DINI-AG-KIM Empfehlungen (produktiv) BIBframe (Prototyp)
1	DNB	siehe oben	GND (gesamt und einzelne Entitätentypen wie Personen, Körperschaften etc.)	Erfassungsformat: Pica	d-nb.info/GND/id	eigenes System (Permalinks DNB)	ca. 9 Millionen	Katalogisierungssprache: deutsch	Vollzugriff	produktiv	GND-Ontologie (produktiv Entity Facts (produktiv)
	NB	siehe oben	DBSM (Metadaten)	Erfassungsformat: Pica	d-nb.info/id	eigenes System (Permalinks DNB)	?	Katalogisierungssprache: deutsch	Vollzugriff	ab 2024	DINI-AG-KIM Empfehl- DNB Metadata Term
		siehe oben	DEA (Metadaten)	Erfassungsformat: Pica	d-nb.info/id	eigenes System (Permalinks DNB)	?	Katalogisierungssprache: deutsch	Vollzugriff	ab 2025	DINI-AG-KIM E/ DNB Metadat
		siehe oben	Cosokinkordanz GND-STW	Erfassungsformat: Pica	d-nb.info/GND/id zbw.eu/stw/descriptor/id	eigenes System (Permalinks DNB, ZBW)	ca. 10.000	?		ab Juni 2022	SKOS Lab Jur
		-iben	Cosokinkordanz GND-Mesh	Erfassungsformat: Pica	d-nb.info/GND/id id.nlm.nih.gov/mesh/vocab #id	eigenes System (Permalinks DNB, Thesaurus National Library of Medicine)	ca. 10.000	Katalogisierungssprachen: deutsch/ englisch		ab Juni 2022	SK
			weitere Cosokinkordanzen zur GND	Erfassungsformat: Pica						Herbst 2022	
				-tes TüISO 24622-1 (XML in CMDI) –		Handle	670 digitale Objekte	Deutsch und Englisch	Vollzugriff		

Titel: Vernetzen und Verknüpfen – Linked Data in Text+
Autoren: Arbeitsgruppe LOD
Text+ Plenary 2022, Mannheim

Text+ ist Teil der Nationale Forschungsdaten Infrastruktur

Gefördert durch Deutsche Forschungsgemeinschaft

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) unter der Projektnummer 460033370.