




Forschungsdokumentation

Bearbeiten und Verlinken mit OpenRefine und Wikidata

BOKU Wien. 09. Dezember 2019

Christian Erlinger-Schiedlbauer

 @librerli

 0000-0001-7872-9617



#Open Refine

- „stand-alone“-Webapplikation (Open Source)
- Bearbeitung tabellarischer Daten (Bereinigung, Zusammenführung, Verlinkung)
- „Reconciliation“ - Verlinkung der Daten mit externen Datenbanken (Wikidata, GND, LoC)
- Data-Input für Wikidata



#Open Refine

OpenRefine partner_tabelle_stand_06 08_2018_ANNA3110.xlsx [Permalink](#)

Facet / Filter [Undo / Redo 0 / 0](#)

6308 records

Show as: **rows** records Show: 5 10 25 50 records

Refresh [Reset All](#) [Remove All](#)

PARTNER_DE [change](#)
6115 choices Sort by: **name** count [Cluster](#)

Institut für Lebensmittelforschung 1
Österreichische Arbeitsgemeinschaft für Ergonomie 1
Amt für Naturparke, Autonome Provinz Südtirol, Bozen 1
Arba Minch Universität 1
Consorzio per il coordinamento delle ricerche inerenti al sistema lagunare di Venezia 1
Electricite de France 1

All	PARTNER_ID	Stammdatensatz	PARTNER_DE	PARTNER_EN	ACRONYM	STRASSE	POSTLEITZAHL	ORT	LAND_ID
☆	1.	9635	Landwirtschaftliche Universität Tirana	Agricultural University of Tirana	UBT	Koder Kamez	1029	Tirana	8
☆	2.	10368	"Fan S. Noli" University, Korçe, Albania	"Fan S. Noli" University, Korçe, Albania					8
☆	3.	9704	Polytechnische Universität Tirana	Polytechnic University of Tirana	PUoT			Tirana	8
☆	4.	10158	Fan Noli Universität	Fan Noli University	UNKO	Shetitolja, "Rilindasit"	7001	Korça	8
☆	5.	11895	Universität Tirana	University of Tirana	UT	Deshmoret e Kombit	1000	Tirana	8
☆	6.	3276	General Directorate of Forests and Pastures, Forest Protection	General Directorate of Forests and Pastures, Forest Protection		Rruga Sami Frasheri 4		Tirana	8
☆	7.	10862	Menaxhim Agro-mjedisor e Ekonomik	Agro-Environment and Economic Management Center	AEEM	Zef Zubani, Nr5	1016	Tirana	8
☆	8.	6850	Institut Algérien du Pétrole	Algerian Institute for Petrol	IALPET	Avenue du 10 Novembre	35000	Boumerdes	12
☆	9.	8702	Nationale Universität La Plata	National University of La Plata	UNLP		1925	La Plata	32
☆	10.	8700	Nationale Universität Córdoba	National University of Córdoba	UNC		X5000GYA	Córdoba	32



#Open Refine

- Import tabellarischer Daten unterschiedl. Herkunft:
 - Spreadsheets (C/TSV, XLS, ODS, Google Sheet)
 - Datenbank (direkte Anbindung)
 - Copy & Paste
 - XML, JSON
 - Via integriertem Web-Parser



#Open Refine

- Bearbeitung der Daten
 - Erzeugung neuer Spalten, basierend auf den vorgegebenen Inhalten
 - Facettenfilter
 - Cluster-Operationen
- Bearbeitung mittels integrierter Funktionen oder on-Board-Scriptsprachen:
 - GREL (General Refine Expression Language)
 - Jython (Python)
 - Clojure
- Export in diverse tabellarische Formate



#Open Refine

- Bearbeitungsschritte werden einmal ausgeführt (in Zellen werden keine „Formeln“ hinterlegt wie zB in XLS)
 - Bearbeitungshistorie ermöglicht das Wiederholen von einzelnen Schritten.
- Kein Hinzufügen neuer Datensätze in einem bestehenden Refine-Projekt möglich.
- Keine Server-Funktionalität (d.h. Arbeiten in Teams muss „organisiert“ werden.)

#Wikidata

Content projects



Wikipedia

The free encyclopedia



Wikibooks

Free textbooks and manuals



Wikiversity

Free learning materials and activities



Wikinews

Free-content news



Wiktionary

Free dictionary and thesaurus



Wikisource

Free-content library



Wikiquote

Collection of quotations



Wikivoyage

Free travel guide



Wikimedia Commons

Free media repository



Wikidata

Free knowledge base



Wikispecies

Free directory of species

- Wikidata Teil des Wiki*versums
- Online seit Oktober 2012

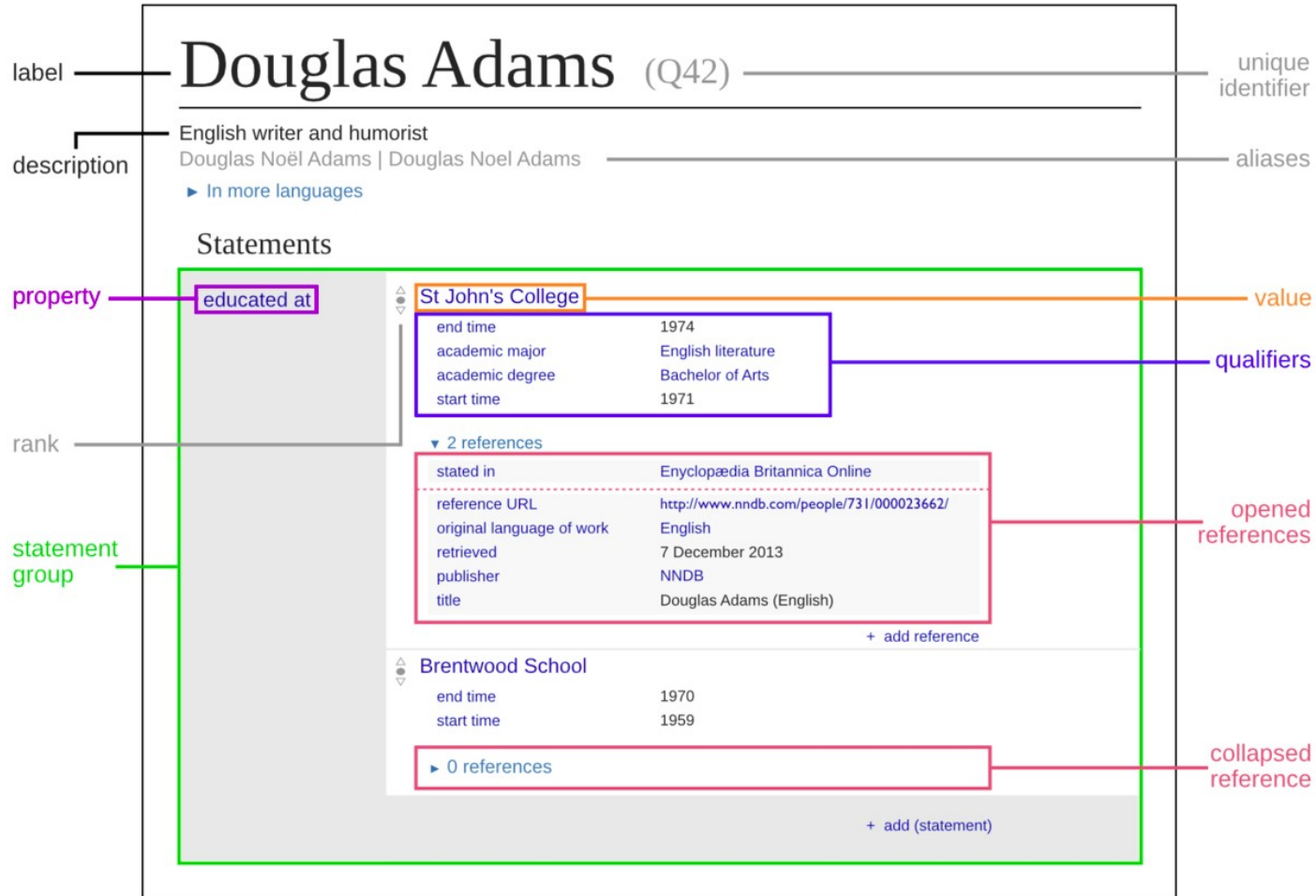
#Wikidata



Wikidata ist eine zentrale, multilinguale und offene Plattform zur Speicherung von strukturierten und verlinkten Daten

- Freie Software und freie Daten (CC 0)
- Offen zur Bearbeitung (Daten und Struktur) für jeden und jede

#Wikidata - Datenmodell



Wikidata und #openScience

(1) Wikidata als (Norm-)Datenhub

- *Standortregister*
- *Zeitschriften- u. Kongressverzeichnis (für Kongress-Identifikation Projekt [ConfIDent](#) verfolgen)*

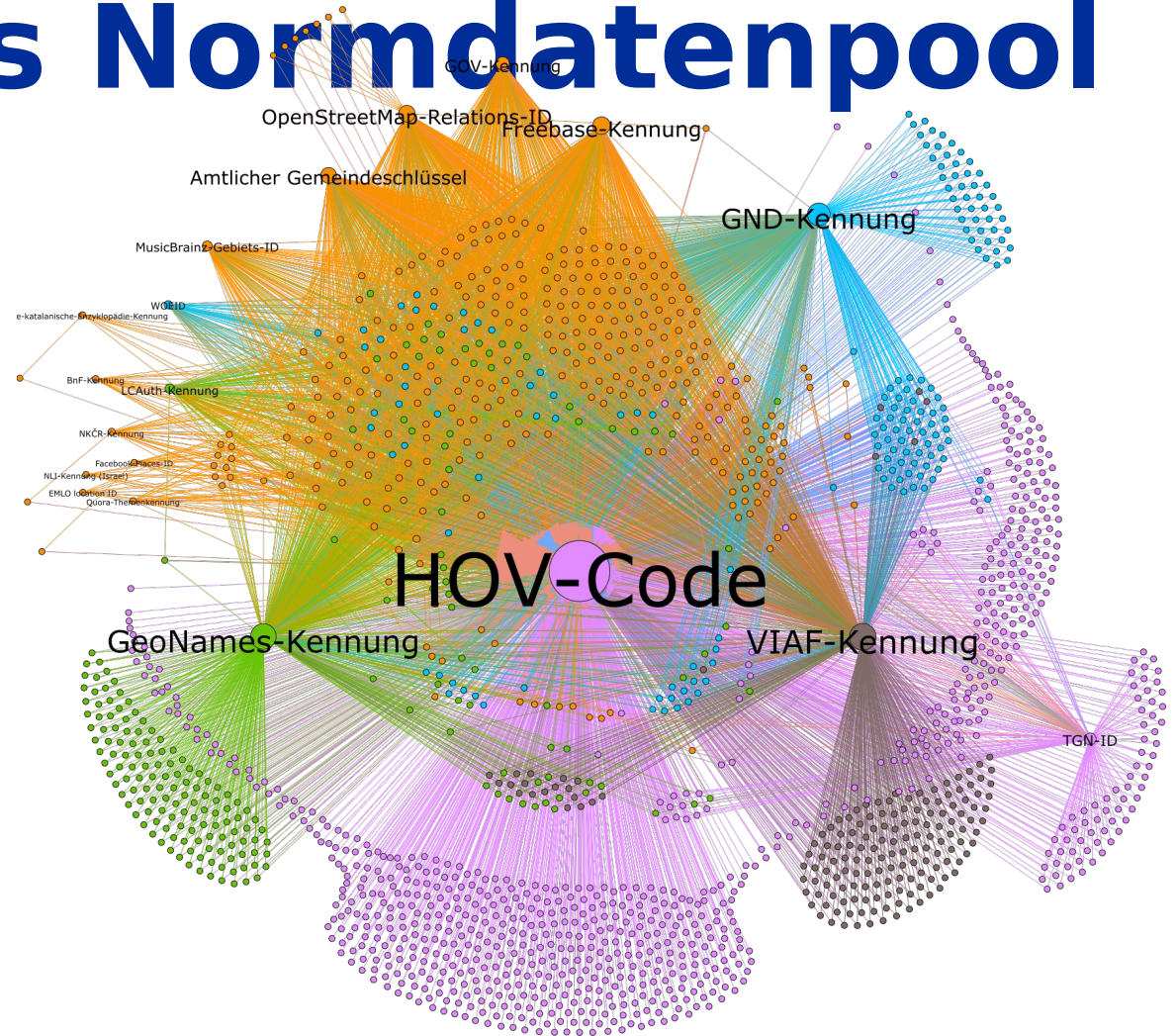
(2) Wikidata als **Bibliographie** (#WikiCite)

(3) Wikidata als **strukturiert abfragbare Enzyklopädie**

Wikidata als Normdatenpool

- 1.574 Normdateien können verlinkt werden.
(Stand: 2019-12-09)
<https://w.wiki/8CV>
 - GND-ID
 - ORCID
 - ...und eben noch 1.572 weitere!
- Laufend werden neue Verzeichnisse zur Ergänzung vorgeschlagen
<https://www.wikidata.org/wiki/Wikidata:PP/AUTH>

Graphik rechts zeigt die ID des Historischen Ortsverzeichnisses für Sachsen in Konkordanz zu anderen Normdateien auf Wikidata (vgl. Munke, 2019: <https://saxorum.hypotheses.org/2775>)



Wikidata - „the sum of knowledge“

- Ideales Standort- und Personen-Register für Institutionen (Universitäten, Bibliotheken etc.)
 - Verlinkung mit klassischen Registern (zB. GND, Geonames, ISIL)
 - Schnelle Adaptierung an neue Verhältnisse und Verzeichnung neuer Links und Ressourcen (Social-Media-Links, Online-Kataloge und Schnittstellen)
- Beispiele
 - Karte der Universitäten in Österreich <https://w.wiki/DXV>
 - Angehörige der BOKU mit ORCID und ggf. GND-ID <https://w.wiki/DXW>
<https://w.wiki/DXW>

Wikidata #scholia - Visualisierung von Forschungsergebnissen

- Scholia (<https://tools.wmflabs.org/scholia>) dient zur Visualisierung von Forschungsergebnissen in Wikidata verzeichnet
 - Personen
 - Zeitschriften
 - Organisationen
 - Artikel
- BOKU auf Scholia:
<https://tools.wmflabs.org/scholia/organization/Q876520>



Wikidata als Bibliographie

- (1) ca 45 % aller Items in Wikidata sind bibliographische Einträge.
(Wissenschaftliche Artikel zum Aufbau einer offenen Zitationsdatenbank)
- (2) Offen zur Ergänzung weiterer Artikeldaten unterschiedlicher Zeitschriften und zur Anreicherung mit Daten (Autoren-Verlinkung, Inhaltserschließung, Zitationsdaten)
- (3) Prinzipiell auch offen als Katalogplattform für Bücher (da mangelt es aber noch an einem klaren und akzeptierten Datenmodell)
- (4) **Nutzung als offener (Artikel-)Index für kleine wie große Bibliotheken**

Zusammenfassung

- **OpenRefine** ermöglicht die **Bearbeitung** von „messy“ Forschungsdaten
- Erweiterung der Bearbeitungskapazitäten von OpenRefine bspw. mit Jupyter Notebooks
- Im Zusammenspiel mit **Wikidata** können (Forschungs-)daten **disambiguiert** und (nachhaltig) **verlinkt** werden.
- Offene Daten zum **Forschungsouptut** ermöglichen kollaborative Bearbeitung und Visualisierung und sind Bestandteil von **#openScience**

Ausblick

- Wikidata erzeugt ein ergänzendes und gemeinschaftliches Verständnis für die Erschließung unserer Daten.
- Wikidata erlaubt die Etablierung einer nachhaltigen Infrastruktur zur Disambigierung und Verlinkung lokaler Datenbestände.
- Unser Auftrag:
 - **Für die freie Datenhoheit solcher Bestände eintreten.**
 - **Freie Datenbestände gemeinsam pflegen und ergänzen.**

Quellen/Links

- Allison-Cassin, Stacy & Scott, Dan 2018. Wikidata: a platform for your library's linked open data. *code4lib Journal* 40, . <https://journal.code4lib.org/articles/13424> [Stand 2019-07-10].
- Ayers, Phoebe u. a. 2019. WikiCite 2018-2019: Citations for the sum of all human knowledge. https://figshare.com/articles/WikiCite_2018-2019_Citations_for_the_sum_of_all_human_knowledge/8947451 [Stand 2019-07-31].
- Bartov, Asaf 2018. Wiki + data = Wikidata (and why you should care). <https://www.youtube.com/watch?v=24DOvuZWaD0&t=239s> [Stand 2019-03-13].
- Delpuech, Antonin 2019. OpenTapioca: Lightweight Entity Linking for Wikidata. *arXiv:1904.09131 [cs]* . <http://arxiv.org/abs/1904.09131> [Stand 2019-04-23].
- Erlinger-Schiedlbauer, Christian u. a. 2018. Semantische Suchabfragen mit der Linked Open Data Cloud generieren SocietyByte. <https://www.societybyte.swiss/2018/10/24/semantische-suchabfragen-mit-der-linked-open-data-cloud-generieren/> [Stand 2019-03-13].
- Fauconnier, Sandra 2019. Structured Data on Commons and GLAM - Wikimania 2019.pdf - Wikimania. https://commons.wikimedia.org/wiki/File:Structured_Data_on_Commons_and_GLAM_-_Wikimania_2019.pdf [Stand 2019-09-11].
- Fontenelle, Giovanna 2019. An International Knowledge Base for all Heritage Institutions (Part 1*). SocietyByte. <https://www.societybyte.swiss/2019/07/04/an-international-knowledge-base-for-all-heritage-institutions-part-1/> [Stand 2019-07-10].
- Lemus-Rojas, Mairelys & Pintscher, Lydia 2017. Wikidata and Libraries: Facilitating Open Knowledge. <http://hdl.handle.net/1805/16690> [Stand 2019-03-13].
- Lih, Andrew 2019. Linking the World's Knowledge through Wikidata : A vision for connected cultural heritage with the crowd. <https://docs.google.com/presentation/d/1LaTHCU7dXh2c06CbzPmHNruMwbayordhd2G860EbGks> [Stand 2019-05-14].
- Neubert, Joachim 2017. Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings. In *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop. NKOS 2017*. <http://ceur-ws.org/Vol-1937/paper2.pdf> [Stand 2019-03-13].
- Stinson, Alex 2017. Wikidata in Collections: Building a Universal Language for Connecting GLAM Catalogs Medium. <https://medium.com/freely-sharing-the-sum-of-all-knowledge/wikidata-in-collections-building-a-universal-language-for-connecting-glam-catalogs-59b14aa3214c> [Stand 2019-03-13].
- van Veen, Theo 2017. Wikidata as a universal (library) thesaurus. https://commons.wikimedia.org/wiki/File:Wikidata_as_universal_library_thesaurus_-_tvv.pdf [Stand 2019-03-13].
- Vrandečić, Denny 2018. Toward an Abstract Wikipedia. In *Proceedings of the 31st International Workshop on Description Logics*. <http://ceur-ws.org/Vol-2211/paper-03.pdf> [Stand 2019-03-13].
- Vrandečić, Denny & Krötzsch, Markus 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57, 10, 78–85.