LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

VERBAALPINA

The FAIR Principles – an important challenge to variational linguistics

Christina Mutter
https://www.verba-alpina.gwi.uni-muenchen.de/

ICLaVE|11 2022
11-14 April 2022 - online

Outline

1. Project Overview of VerbaAlpina

research aims, area under investigation, conceptual domains, data

2. The FAIR principles

3. FAIR@VerbaAlpina

4. How the FAIR principles are challenging for variational linguistics

Project Overview

- *VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit* (VerbaAlpina. The Alpine cultural region reflected through its multilingualism)
- Funded by the German Research Foundation (DFG)
- 1st term: 10/2014-10/2017, 2nd term: 11/2017-10/2020, 3rd term: 11/2020 – 10/2023 (perspective until 2026)
- Investigation of the multilingual Alpine region
- Combination of (geo-)linguistics and Digital Humanities (DH)

Project Overview

Research Aims

- Selective and analytical investigation of the linguistically and dialectally highly fragmented alpine space in its historico-cultural and historical-linguistic unity
- Overcoming of the traditional limitation of geolinguistic investigation to nation-states
- recognition of connections regarding the etymology of the individual dialectal words
- Setting up a portal by using modern media technology: documentation, data collection, collaborative development
- cooperation with other projects is fundamental for VerbaAlpina

Project Overview

Area under investigation: The Alpine region

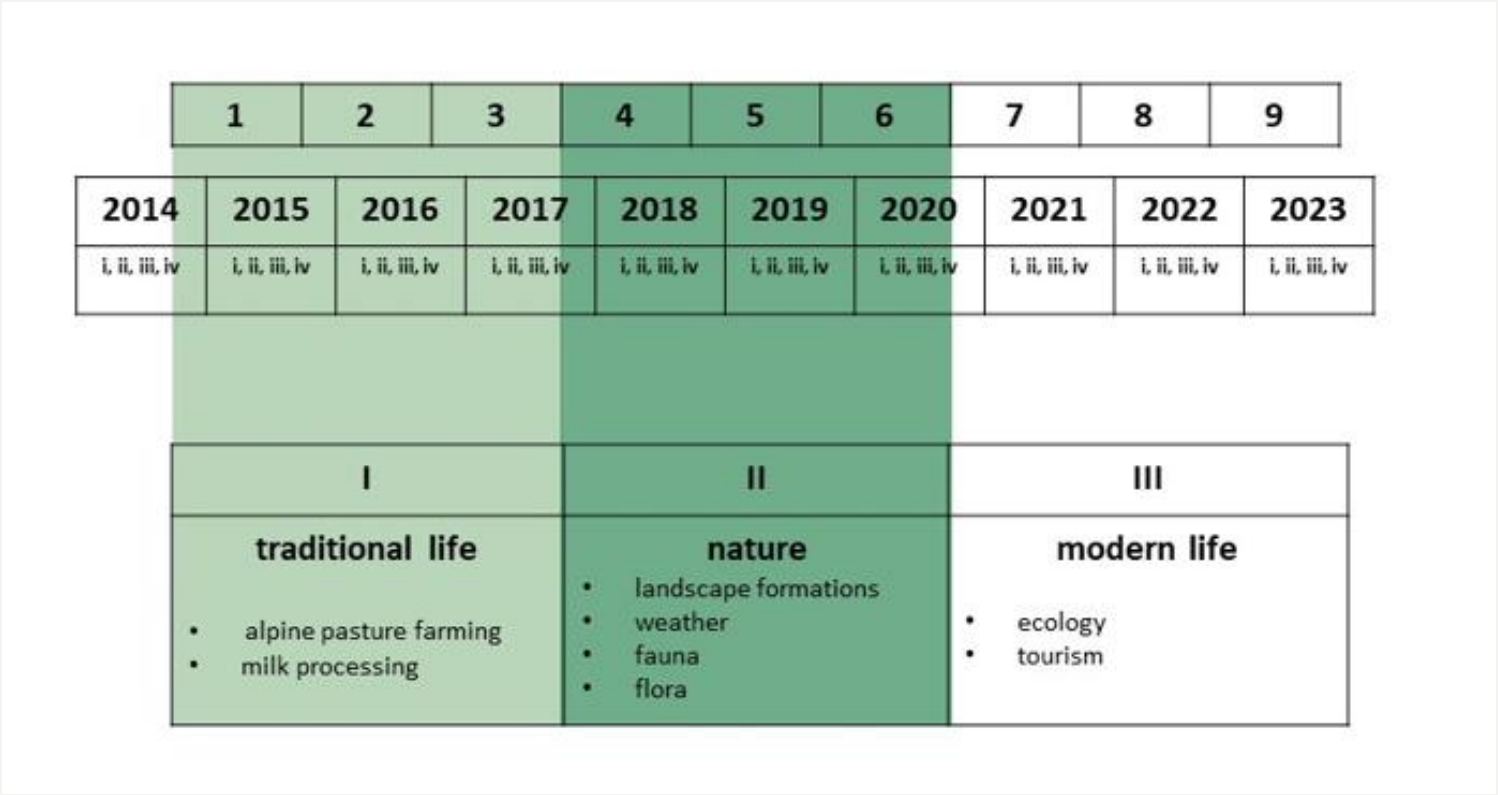
- Area of investigation is limited to the territorial borders defined by the Alpine convention

- surface area of 190,600 km^2, encompasses parts of six different countries (D, A, CH, I, F, SLO) and two entire countries (FL, MC)

- ethnographic and topographic homogeneity and strong linguistic heterogeneity → 3 language families

Project Overview

Three conceptual domains



Project Overview

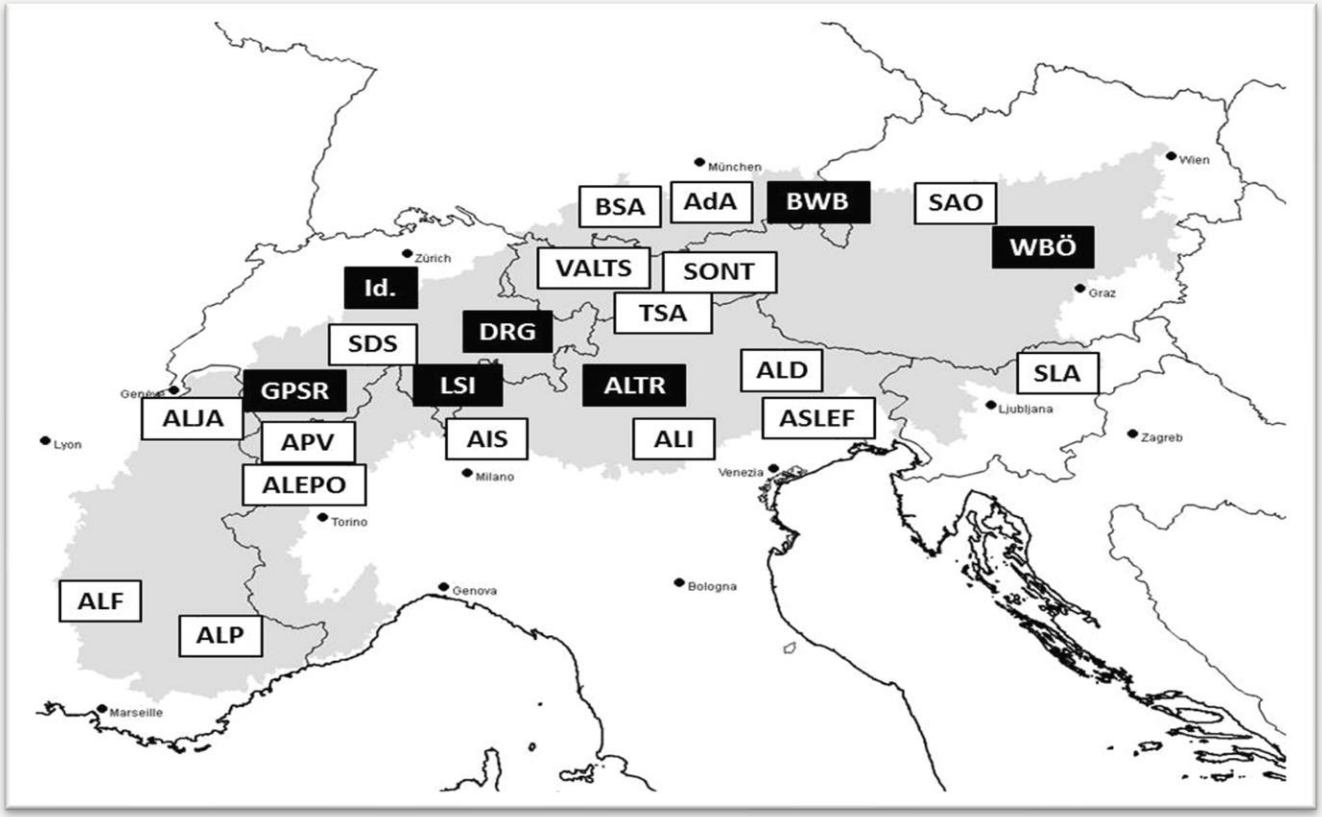
Data

Multiple different sources

- printed atlases/dictionaries (georeferenced)
- digital material from project partners
- crowdsourcing

Project Overview

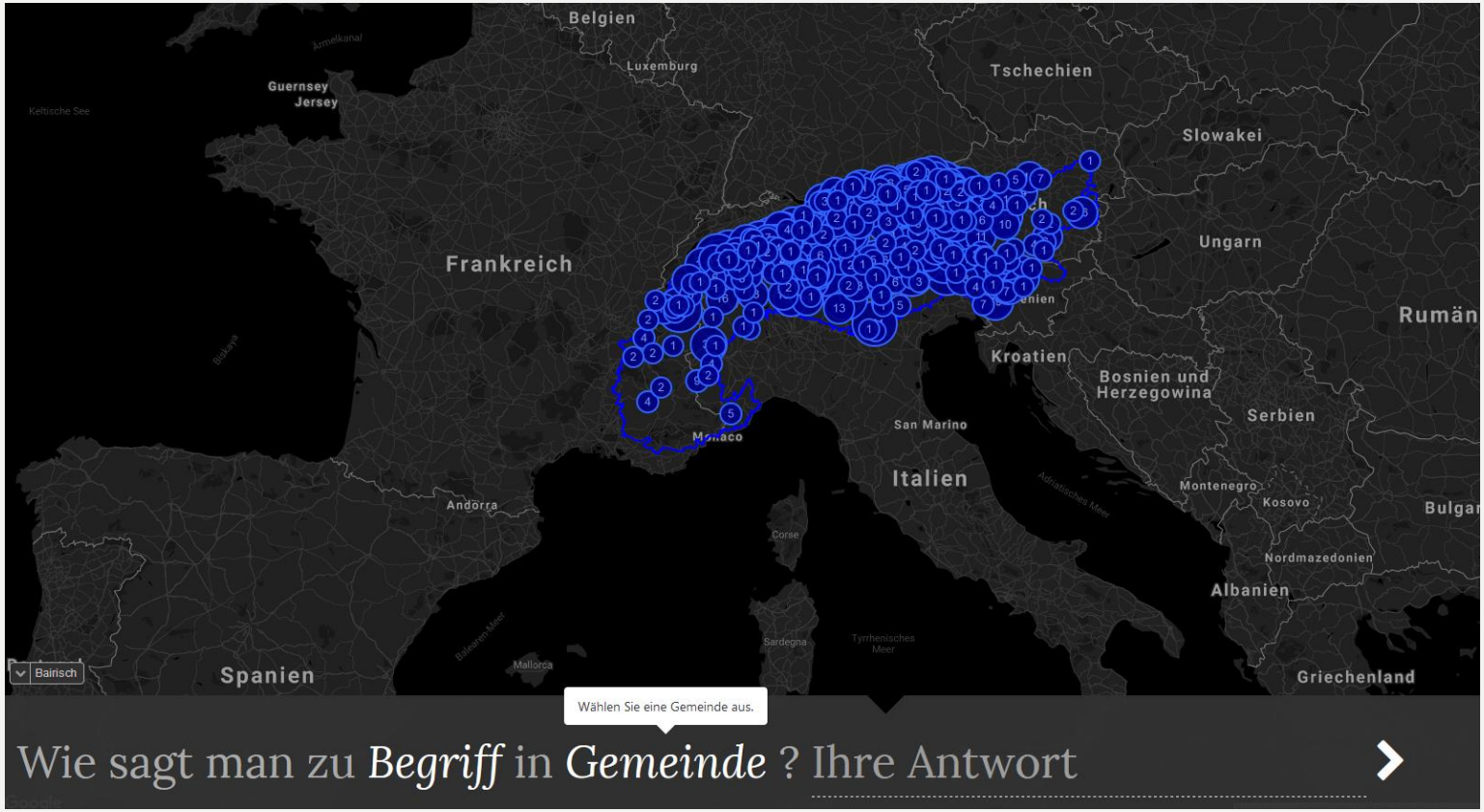
Atlases and Dictionaries in the Alpine region



Project Overview

Crowdsourcing-Tool

www.lmu.de/verbaalpina



The FAIR principles

Research data have to be FAIR:

F_indable

A_ccessible

I_nteroperable

R_eusable

→ principles postulated by Wilkinson et al. 2016 as the guiding principles for scientific data management

The FAIR principles

F_indable → via library catalogues and data aggregators

A_ccessible → via open access licences

I_nteroperable → via compatibility of databases and their interconnection

R_eusable → results from F, A, I

The FAIR principles

What VerbaAlpina does to make its data FAIR

F_indable

- Cooperation with the University Library of Munich University (VerbaAlpina data is available on UB Discover, version 19/1 + 19/2) and the two RDM projects "e-humanities-interdisziplinär" (until 2021) and "GeRDI" (Generic Research Data Infrastructure) (until 2019)

A_ccessible

- Creative Commons licence (compatible with open access and open source) for all data managed by VerbaAlpina

  (up to version 18/1: CC BY SA 3.0, from 18/2: CC BY SA 4.0)

The FAIR principles

I_nteroperable

- through a fine granulation of the data stock via
  - structured data processing (transcription, tokenization, typification)
  - assignment of norm data (Q-ID, L-ID, GND, GeoNames etc.)
  - enrichment with metadata in DataCite and CIDOC CRM format
  - assignment of persistent identifiers (e.g. DOIs, Digital Object Identifiers)

- access to primary data and metadata (via interactive map, Lexicon Alpinum, API)

R_eusable

- results from F, A, I

FAIR@VerbaAlpina

- **requirements of F, A, R aim to be both human readable and machine readable**

→ apply to human-machine-human communication and to machine-machine communication

- **I → only applies to machine-machine communication**
  → BUT: is crucial for the progress of research

FAIR@VerbaAlpina

I_nteroperable

- through a fine granulation of the data stock
  - structured data processing (transcription, tokenization, typification)
  - assignment of norm data (Q-ID, L-ID, GND, GeoNames etc.)
  - enrichment with metadata in DataCite and CIDOC CRM format
  - assignment of persistent identifiers (e.g. DOIs, Digital Object Identifiers)
- access to primary data and metadata (via interactive map, Lexicon Alpinum, API)

R_eusable

- results from F, A, I

Assignment of norm data

- **Norm data created by VerbaAlpina**

  For the 3 core entities

  - morpho-lexical types → L
  - concepts → C
  - municipalities → A

  e.g.: L1435, „babeurre (m.) (roa.)“
  C612, „ALMHÜTTE“ (*chalet*)
  A60171, „Sils in Engadin/Segl“

- **Persistent identifiers of external institutions**
  (knowledge data bases/norm data bases/reference dictionaries)

FAIR@VerbaAlpina

Persistent identifiers of external institutions

- **Q-IDs** of Wikidata (for concepts), partly also **L-IDs** of Wikidata (for morpho-lexical types)
- partly **GNDs** of the German National Library (for concepts) (*Gemeinsame Normdatei*, „Integrated Authority File“)
- **GeoNames** of www.geonames.org (for municipalities)
- **ISO Codes 639-3** (for languages)
- Identifiers of **reference dictionaries** (for morpho-lexical types + base types)
- **DOIs** (Digital Object Identifiers, assigned to every single data)

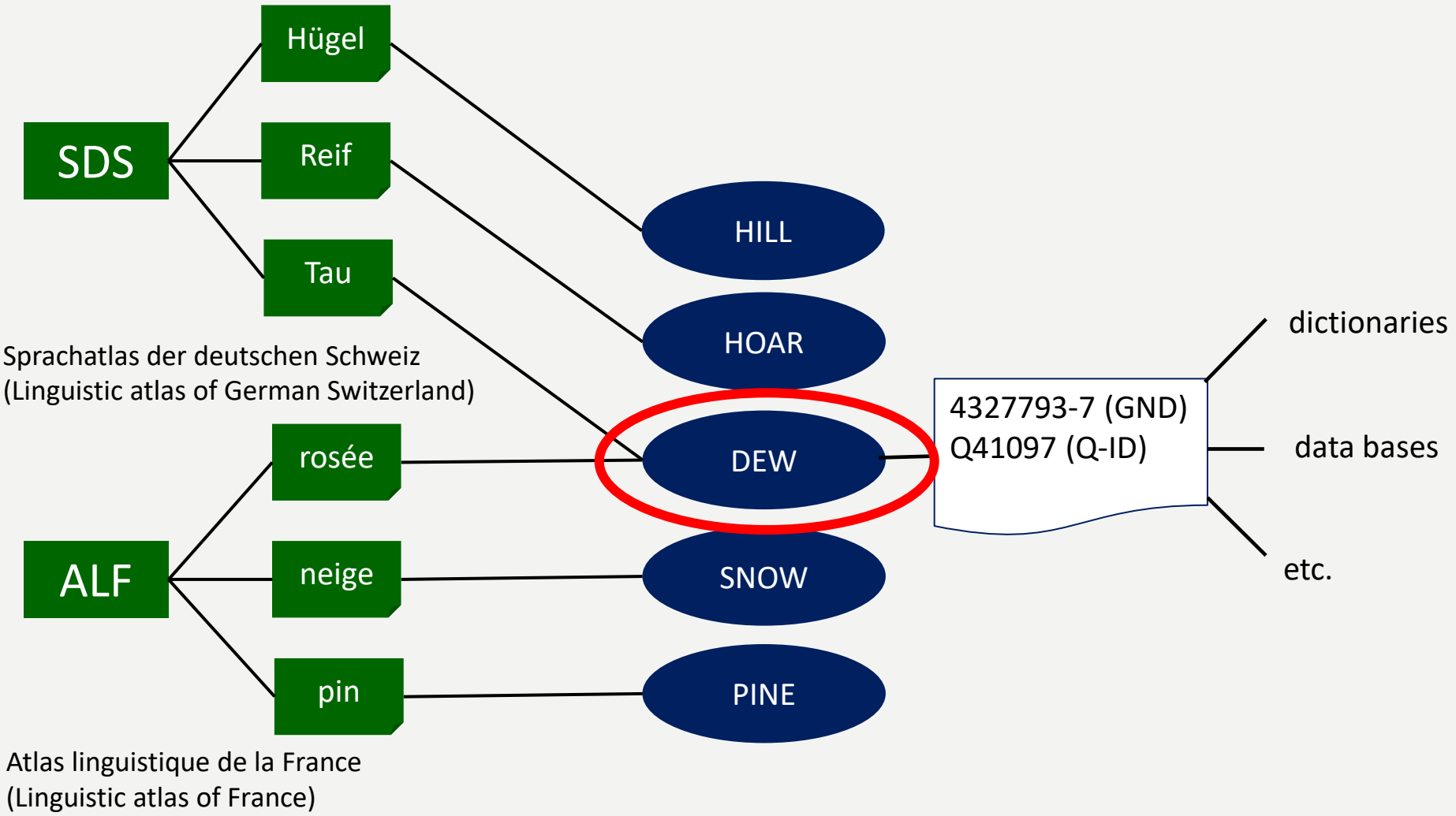
Limerick
2962943

Q54050
(Q-ID)

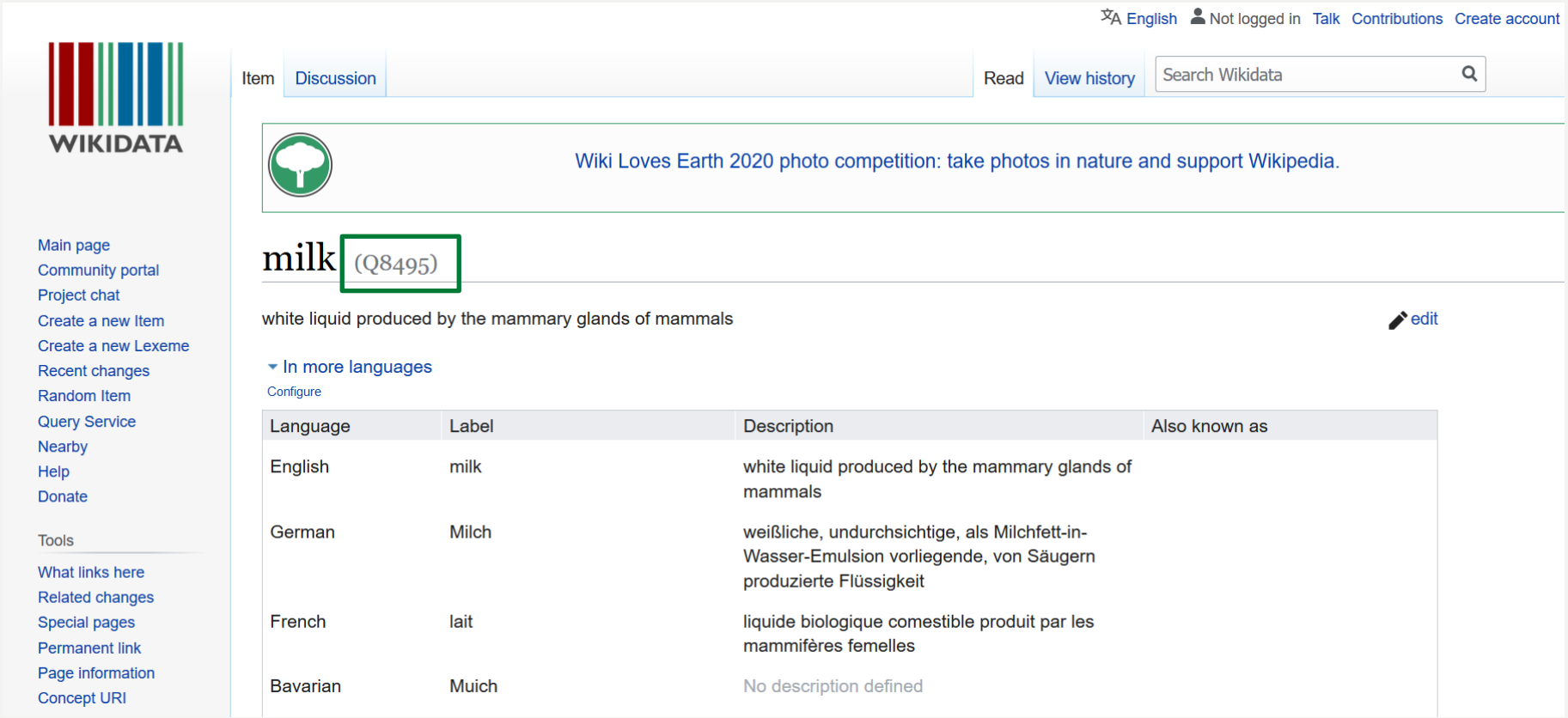
L73260
(L-ID)

4135744-9 (GND)

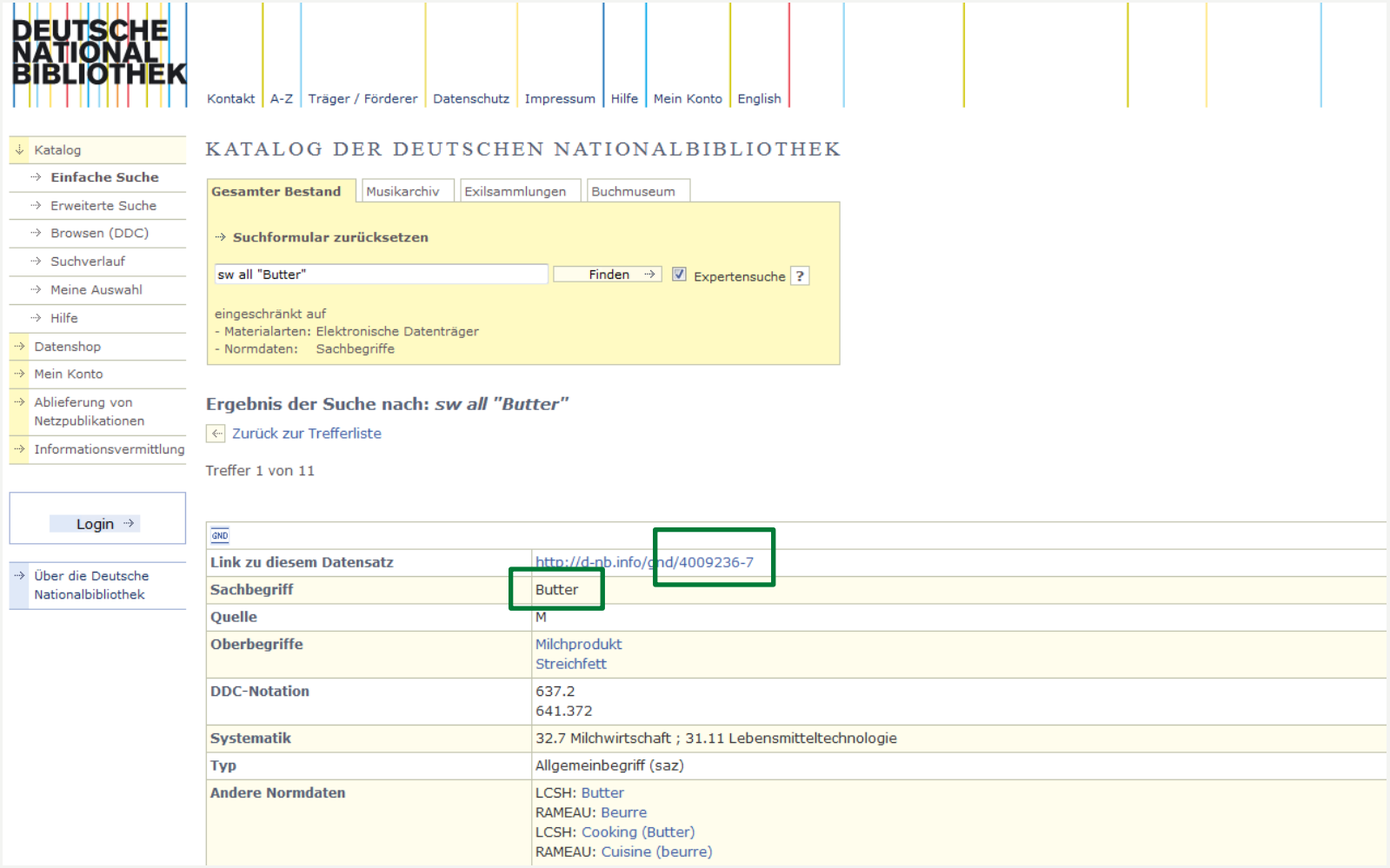
FAIR@VerbaAlpina



FAIR@VerbaAlpina



FAIR@VerbaAlpina



DEUTSCHE NATIONAL BIBLIOTHEK

Kontakt | A-Z | Träger / Förderer | Datenschutz | Impressum | Hilfe | Mein Konto | English

- Katalog
- Einfache Suche
- Erweiterte Suche
- Browsen (DDC)
- Suchverlauf
- Meine Auswahl
- Hilfe
- Datenshop
- Mein Konto
- Ablieferung von Netzpublikationen
- Informationsvermittlung

Login

Über die Deutsche Nationalbibliothek

KATALOG DER DEUTSCHEN NATIONALBIBLIOTHEK

Gesamter Bestand | Musikarchiv | Exilsammlungen | Buchmuseum

Suchformular zurücksetzen

sw all "Butter" Finden Expertensuche ?

eingeschränkt auf
- Materialarten: Elektronische Datenträger
- Normdaten: Sachbegriffe

Ergebnis der Suche nach: *sw all "Butter"*

Zurück zur Trefferliste

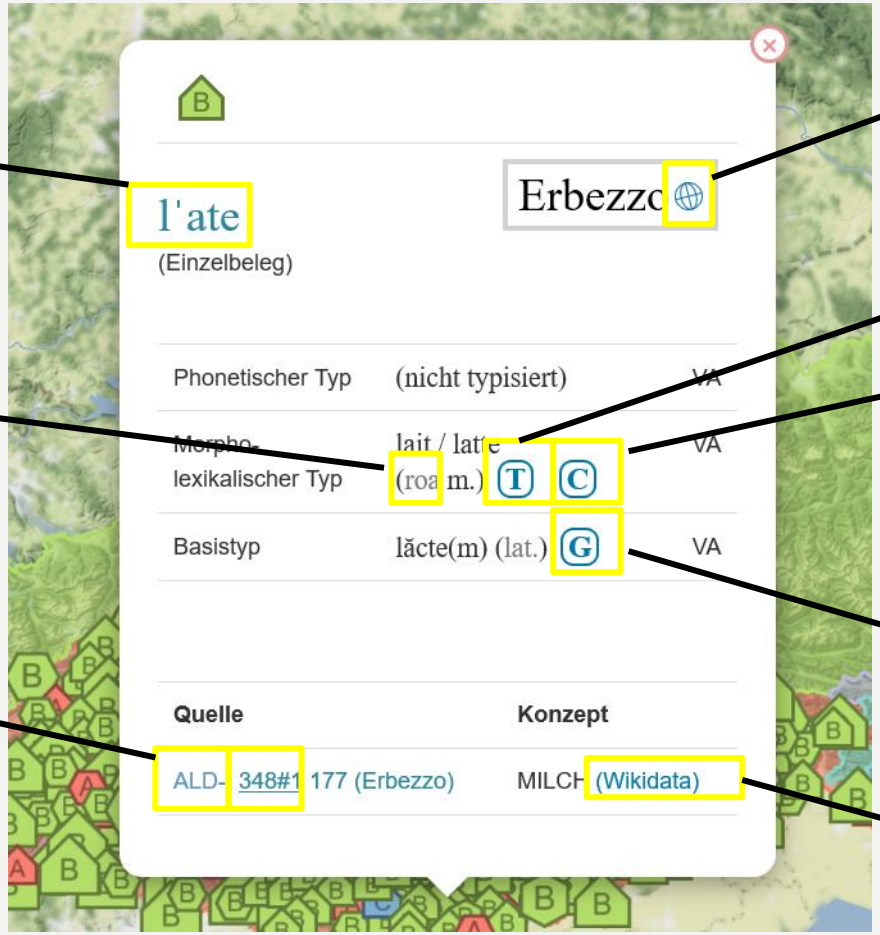
Treffer 1 von 11

GND

Link zu diesem Datensatz	http://d-nb.info/gnd/4009236-7
Sachbegriff	Butter
Quelle	M
Oberbegriffe	Milchprodukt Streichfett
DDC-Notation	637.2 641.372
Systematik	32.7 Milchwirtschaft ; 31.11 Lebensmitteltechnologie
Typ	Allgemeinbegriff (saz)
Andere Normdaten	LCSH: Butter RAMEAU: Beurre LCSH: Cooking (Butter) RAMEAU: Cuisine (beurre)

FAIR@VerbaAlpina

Detail view of one specific data point



IPA

Darstellung: IPA VA

ISO Code

Source + link

Goebl, Hans (Wiesbaden): Atlant linguistich dl ladin dolomitich y di dialec vejins l, vol. 1-7 (sprechend: http://ald.sbg.ac.at/ald/ald-i/index.php), 1998, vol. 1-7, Reichert

Link

GeoNames

Dictionaries

Treccani: latte

CNRTL: lait

Etymological dictionary

Georges: lac 2, 525

Wikidata

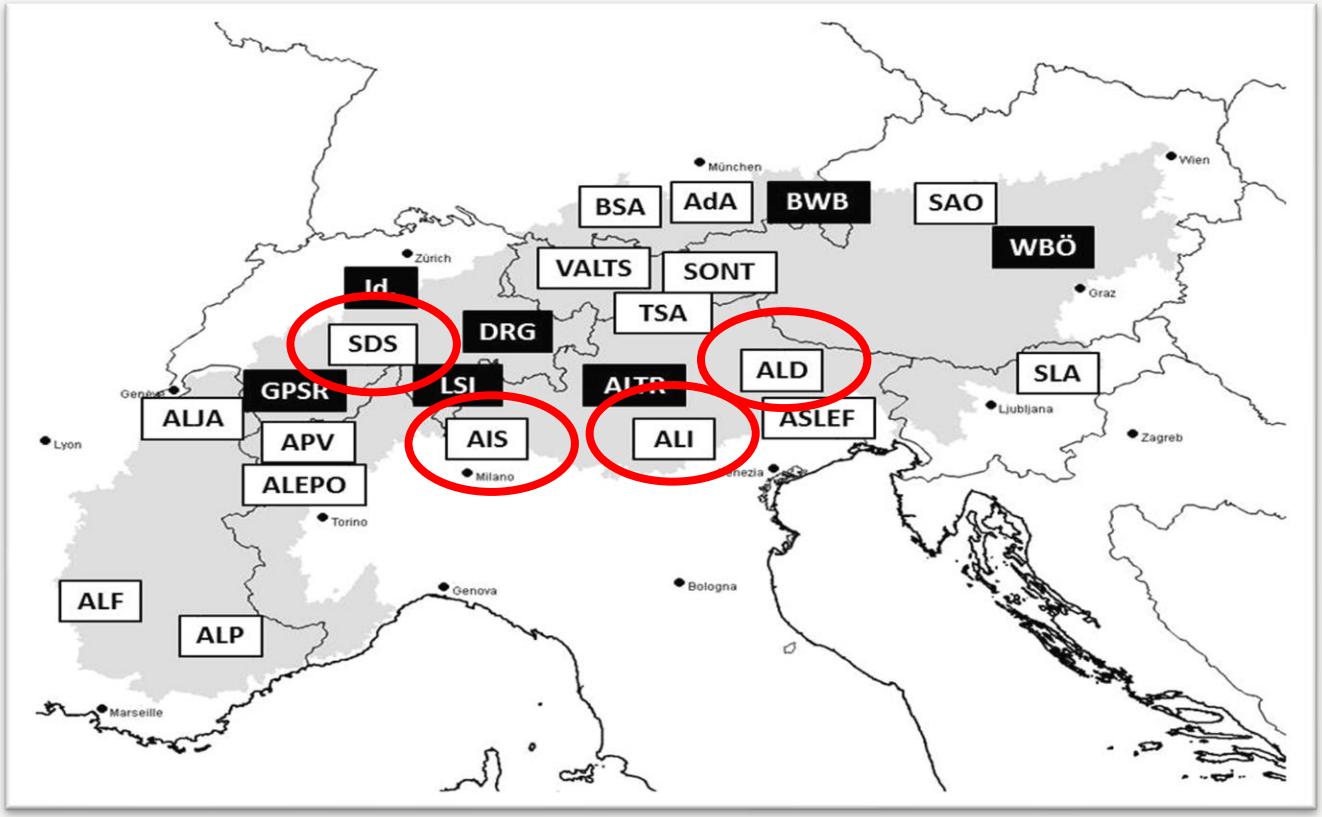
How the FAIR principles are challenging for variational linguistics

VerbaAlpina tries to create FAIR output

BUT: many data sources (the input) are far away from being FAIR

Main reason: lack of interoperability

FAIRness of atlases in the Alpine region



How the FAIR principles are challenging for variational linguistics

Comparison of the FAIRness of 4 geolinguistic atlases with VerbaAlpina

	Findable		Accessible		Inter-operable	Reusable	
	hum. read.	mach. read.	hum. read.	mach. read.	mach. read.	hum. read.	mach. read.
ALI	–	–	–	–	–	–	–
SDS	+	+	+	–	–	+	–
AIS	+	+	+	–	–	+	–
ALD	+	+	+	–	–	+	+
VA	+	+	+	+	+	+	+

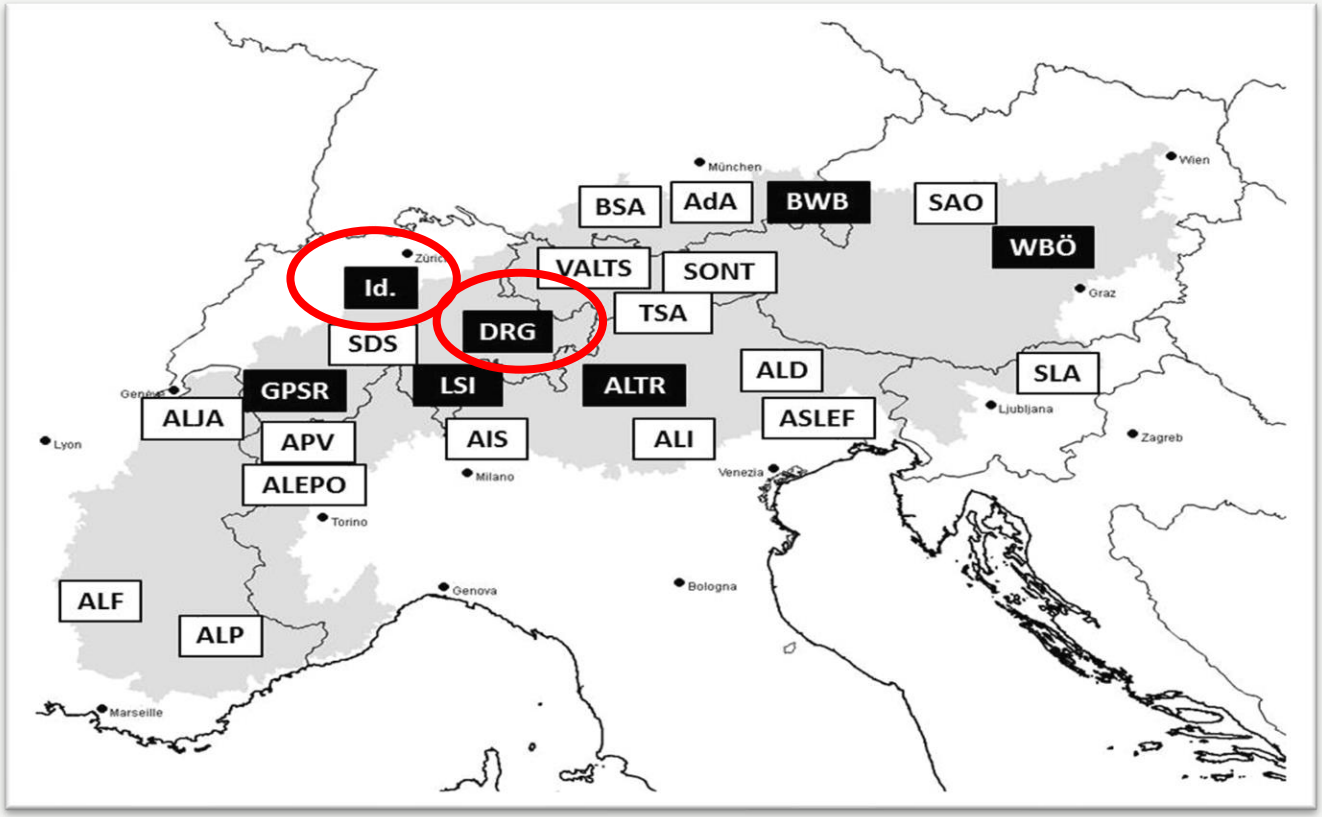
ALI = Atlante linguistico italiano
SDS = Sprachatlas der Deutschen Schweiz
AIS = Sprach- und Sachatlas Italiens und der Südschweiz
ALD = Atlant linguistich dl ladin dolomitich y di dialec vejins
VA = VerbaAlpina

How the FAIR principles are challenging for variational linguistics

Lack of FAIRness because…

- linguistic atlases are usually only accessible as printed works
- only a few offer at least the most basic level of digitisation i. e. digital photos (scans), e.g. AIS (NavigAIS), SDS for original material
- no older atlas has yet been prepared in the form of a structured corpus that also allows export of the data
- only the ALD is based on a non- interoperable digital format that proved to be machine-readable after certain adaptations

Interoperability of dictionaries in the Alpine region



- **Schweizerisches Idiotikon**

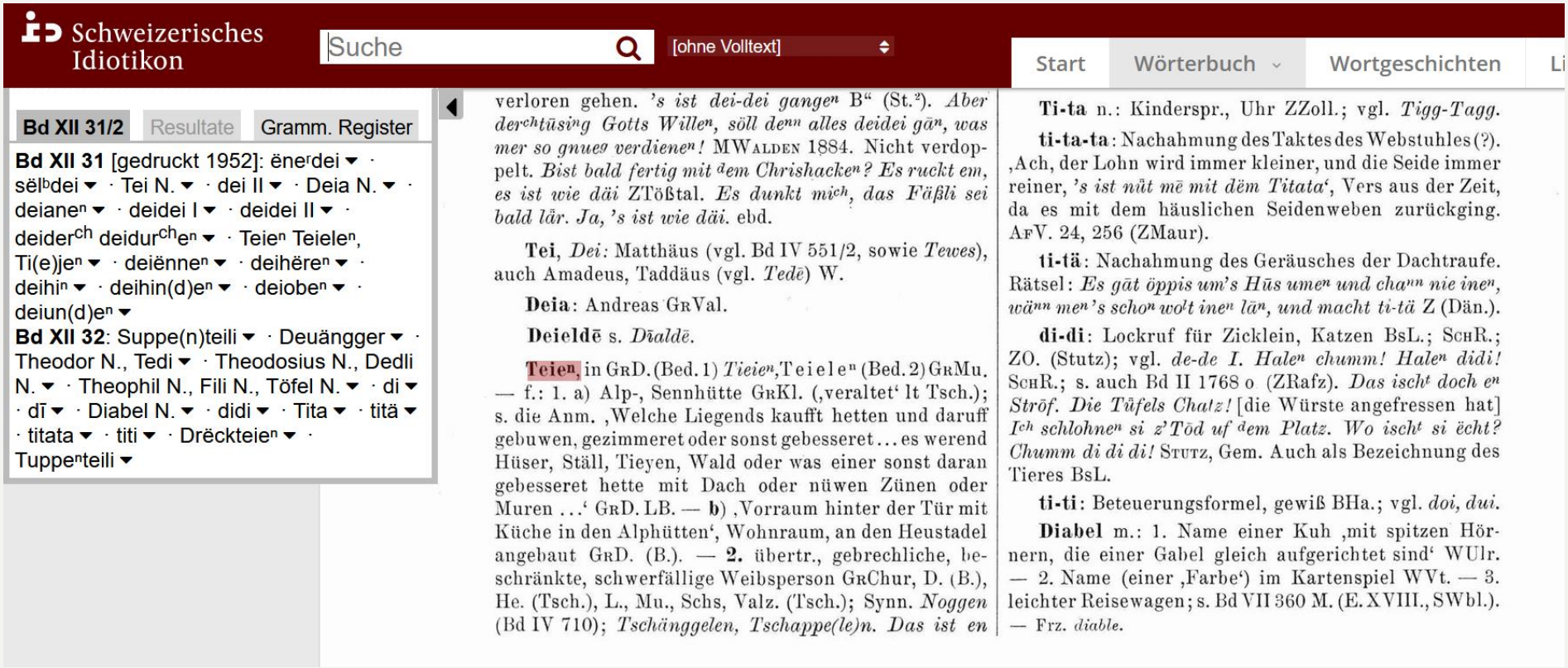
Many Cons

- referencing is possible, but not precise enough
- structure of URLs is cryptic/semantically intransparent

  Example: *Teie(n)* (Alpine chalet)
  https://digital.idiotikon.ch/idtkn/faksimile.php?band=12&spalte=31&hl=229458

- referencing is based on the pages of the printed edition, not on the specific entries (naming of volume and column in the URL)
- all entries on one page are addressed via the same URL (no individual URLs)
- URLs refer to a scan of the printed volume and retain its page/column logic as reference system (see URL)
- entries cannot be copied/pasted (image and no HTML code)

How the FAIR principles are challenging for variational linguistics



Source: Schweizerisches Idiotikon https://digital.idiotikon.ch/idtkn/faksimile.php?band=12&spalte=31&hl=229458

- **Dicziunari Rumantsch Grischun (DRG)**

Pros

- **Links do not refer to the book page as in the Idiotikon, but to the corresponding entry**

  Example: BARGUN (Hay barn in mountain pastures)
  http://online.drg.ch/main.aspx#35ac0b4c3795c91831f3534a40a0c5f0

- **Contents are available in HTML and can be copied/pasted and cited**

Cons

- **Identifiers are too long and make intuitive and automatic generation of specific URLs impossible**

How the FAIR principles are challenging for variational linguistics

Source: DRG-Online http://online.drg.ch/#35ac0b4c3795c91831f3534a40a0c5f0

How the FAIR principles are challenging for variational linguistics

Difficulties in the assignment of norm data

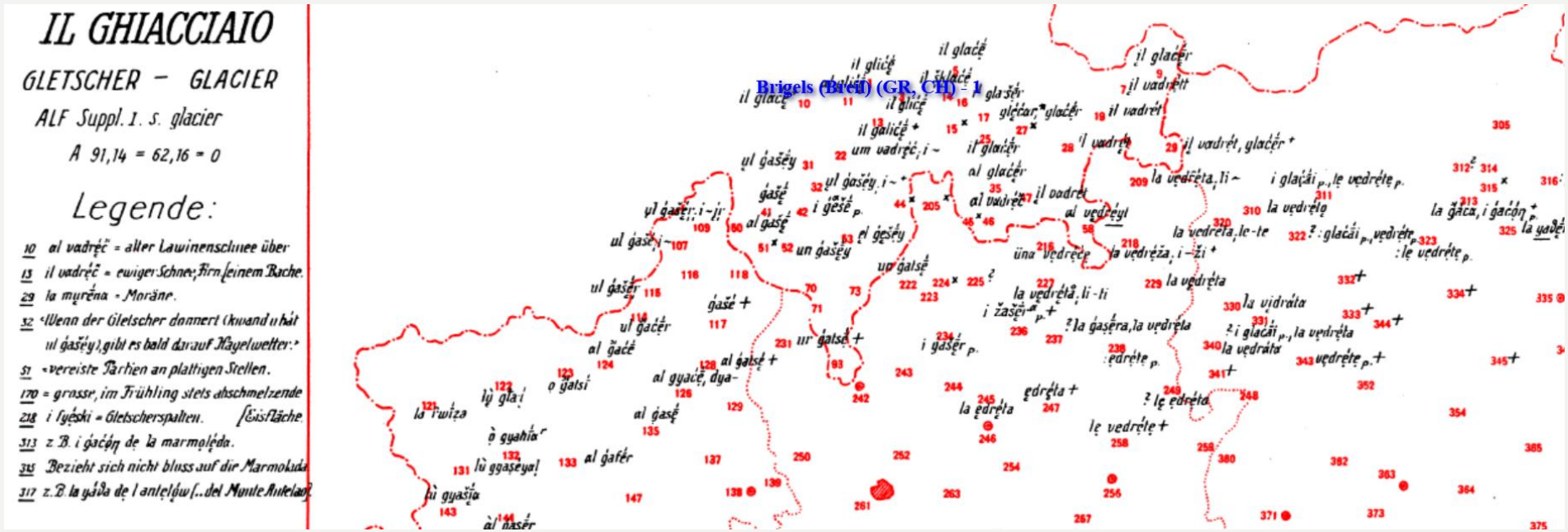
- ISO-Codes 639-3 are not detailed enough

Varieties according to Lia Rumantscha		Wikidata Q-ID	Glottolog	Constitutions CH, GR	ISO-693.3
"standard language"	Rumantsch Grischun	Q688873	Rumantsch Grischun	"Rhaeto-Romance" (Räto-romanisch)	roh
"written idioms"	Puter	Q688309	Upper Engadine		
	Vallader	Q690226	Lower Engadine		
	Surmiran	Q690216	Surmiran-Albula		
	Sursilvan	Q688348	Sursilvan		
			Sursilvan-Oberland		
	Sutsilvan	Q688272	Sutsilvan		
spoken dialect	Jauer	Q690181			

How the FAIR principles are challenging for variational linguistics

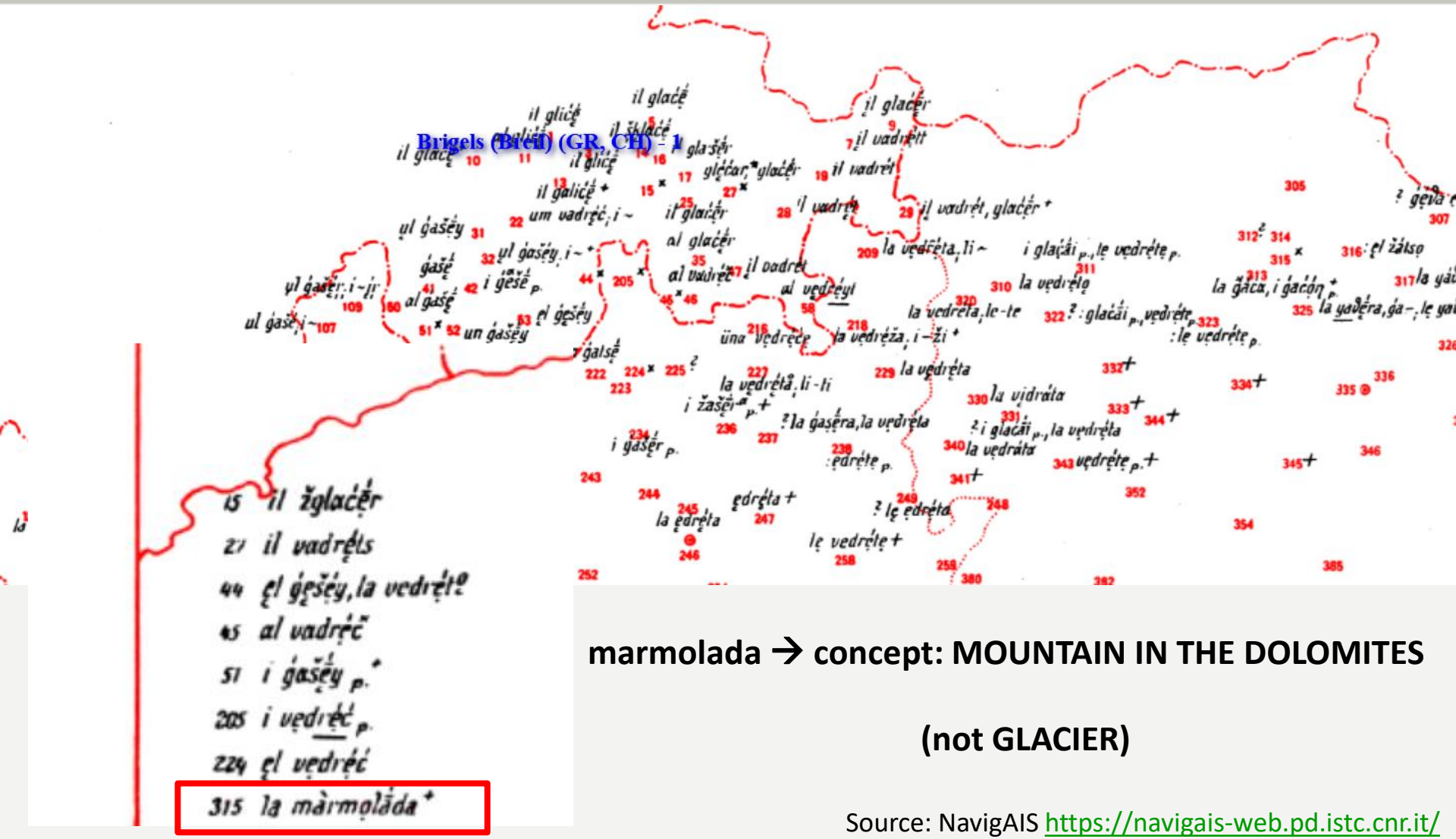
- **Concept assignment is sometimes unclear**

Example: concept GHIACCIAO (GLACIER)



Source: NavigAIS https://navigais-web.pd.istc.cnr.it/

How the FAIR principles are challenging for variational linguistics



marmolada → concept: MOUNTAIN IN THE DOLOMITES

(not GLACIER)

Source: NavigAIS https://navigais-web.pd.istc.cnr.it/

References

- Alpine Convention. Contracting parties. https://www.alpconv.org/en/home/ [accessed 26 June 2020].
- Force11 (eds.), "The Fair Data Principles", https://www.force11.org/group/fairgroup/fairprinciples [accessed 26 June 2020]
- Kümmet, S. / Lücke, S. / Schulz, J. / Zacherl, F.: s.v. "Forschungsdatenmanagement", in: VerbaAlpina-de 19/1 (Erstellt: 18/2, letzte Änderung: 18/2), Methodologie, https://doi.org/10.5282/verba alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DF%23112
- Thomas Krefeld & Stephan Lücke (2019): Kleinsprachen, Dialekte und die FAIR-Prinzipien (am Beispiel von VerbaAlpina), Version 1 (02.10.2019, 15:57). In: Korpus im Text, Serie A, 48389. url: http://www.kit.gwi.uni-muenchen.de/?p=48389&v=1
- Krefeld, Thomas / Lücke, Stephan (2020): 54 Monate VerbaAlpina – auf dem Weg zur FAIRness, in: Ladinia, vol. XLIII, 139-156
- Lücke, S.: s.v. "FAIR-Prinzipien", in: VerbaAlpina-de 19/1 (Erstellt: 18/2, letzte Änderung: 18/2), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DF%23128
- Lücke, S.: s.v. "Metadaten", in: VerbaAlpina-de 19/1, Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DM%23140
- Lücke, S.: s.v. "Normdaten", in: VerbaAlpina-de 19/1 (Erstellt: 18/2, letzte Änderung: 18/2), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DN%23114
- Lücke, S. / Schulz, J.: s.v. "Digital Object Identifier (DOI)", in: VerbaAlpina-de 19/1 (Erstellt: 16/1), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DD%2373
- Mutter, C.: s.v. "Wikidata", in: VerbaAlpina-de 19/1 (Erstellt: 18/1), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DW%23105
- M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 03 2016. doi: 10.1038/sdata.2016.18

Thank you for your attention!

https://www.verba-alpina.gwi.uni-muenchen.de/