

# 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)
- 1<sup>st</sup> term: 10/2014-10/2017, 2<sup>nd</sup> term: 11/2017-10/2020, 3<sup>rd</sup> term: 11/2020 – 10/2023 (perspective until 2026)
- Investigation of the multilingual Alpine region
- Combination of (geo-)linguistics and Digital Humanities (DH)



## 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



## 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<sup>2</sup>, 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





## Three conceptual domains

project years	1	2	3	4	5	6	7	8	9	
calendar year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
quarter	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV	I, II, III, IV
project phase	I				II				III	
focus	<b>traditional life</b> <ul style="list-style-type: none"><li>alpine pasture farming</li><li>milk processing</li></ul>				<b>nature</b> <ul style="list-style-type: none"><li>landscape formations</li><li>weather</li><li>fauna</li><li>flora</li></ul>				<b>modern life</b> <ul style="list-style-type: none"><li>ecology</li><li>tourism</li></ul>	



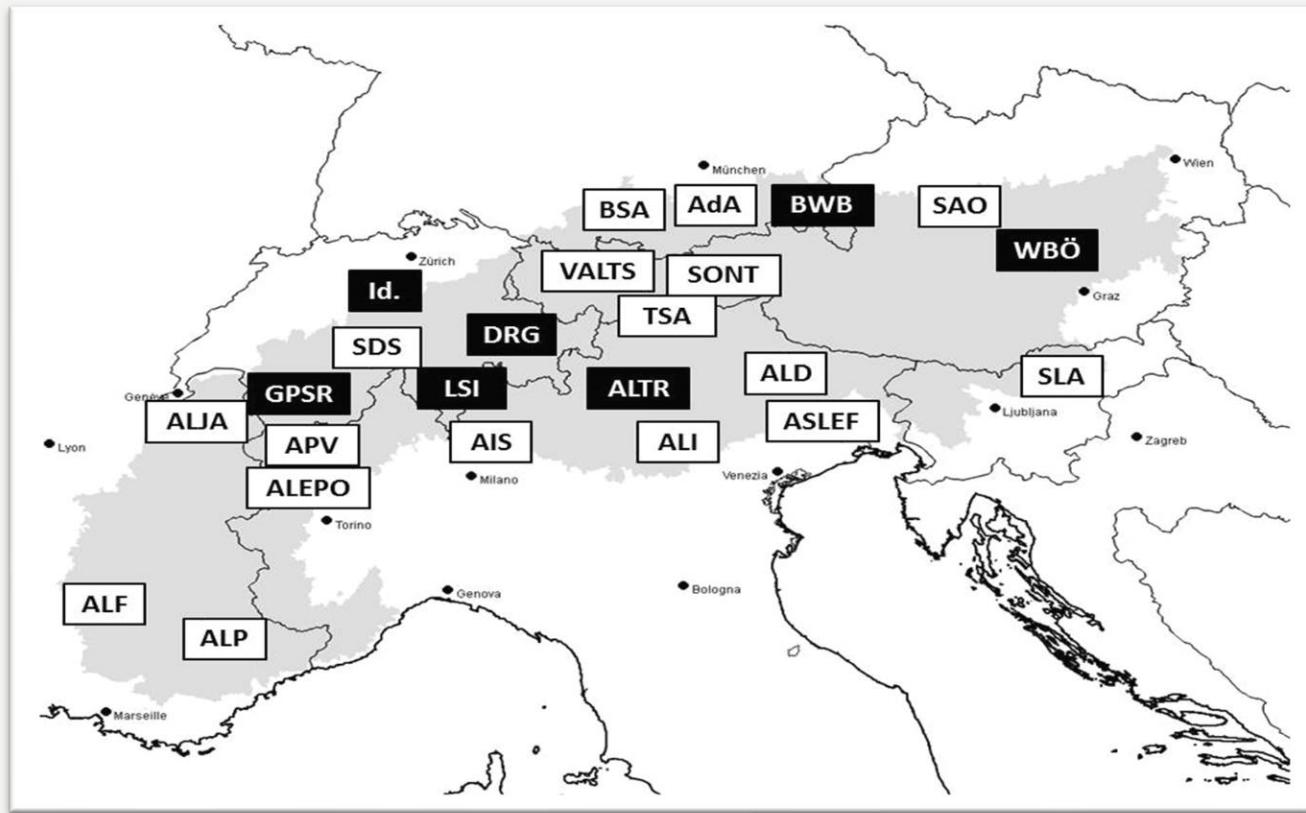
## Data

### Multiple different sources

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



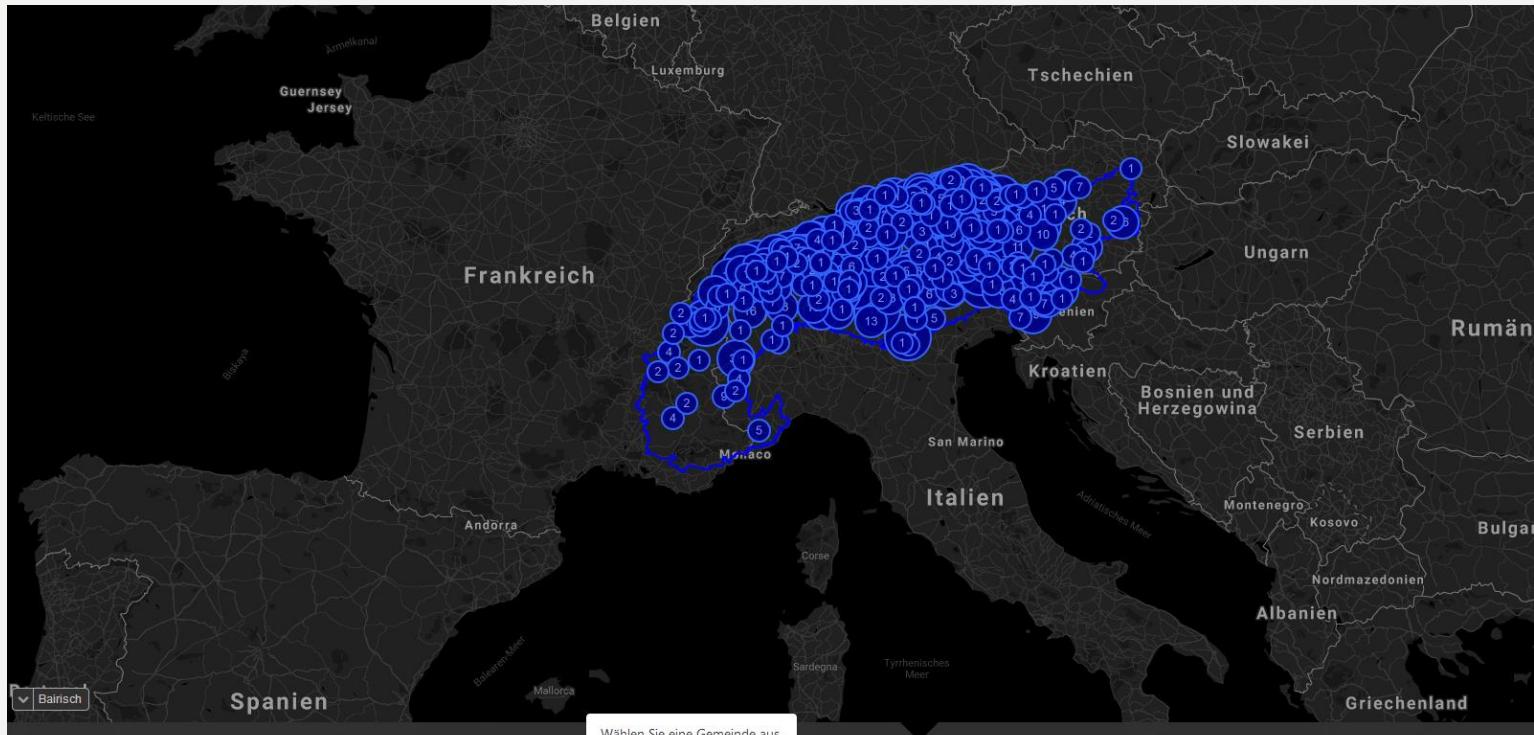
## Atlases and Dictionaries in the Alpine region





## Crowdsourcing-Tool

[www.lmu.de/verbaalpina](http://www.lmu.de/verbaalpina)



Wie sagt man zu Begriff in Gemeinde ? Ihre Antwort





# The FAIR principles

Research data have to be FAIR:

F<sub>indable</sub>

A<sub>ccessible</sub>

I<sub>nteroperable</sub>

R<sub>Reusable</sub>

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

# The FAIR principles



- |                |  |
|----------------|--|
| F_inable       | → 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                                     |



## 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)



## 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



- 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



## 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)



## Persistent identifiers of external institutions

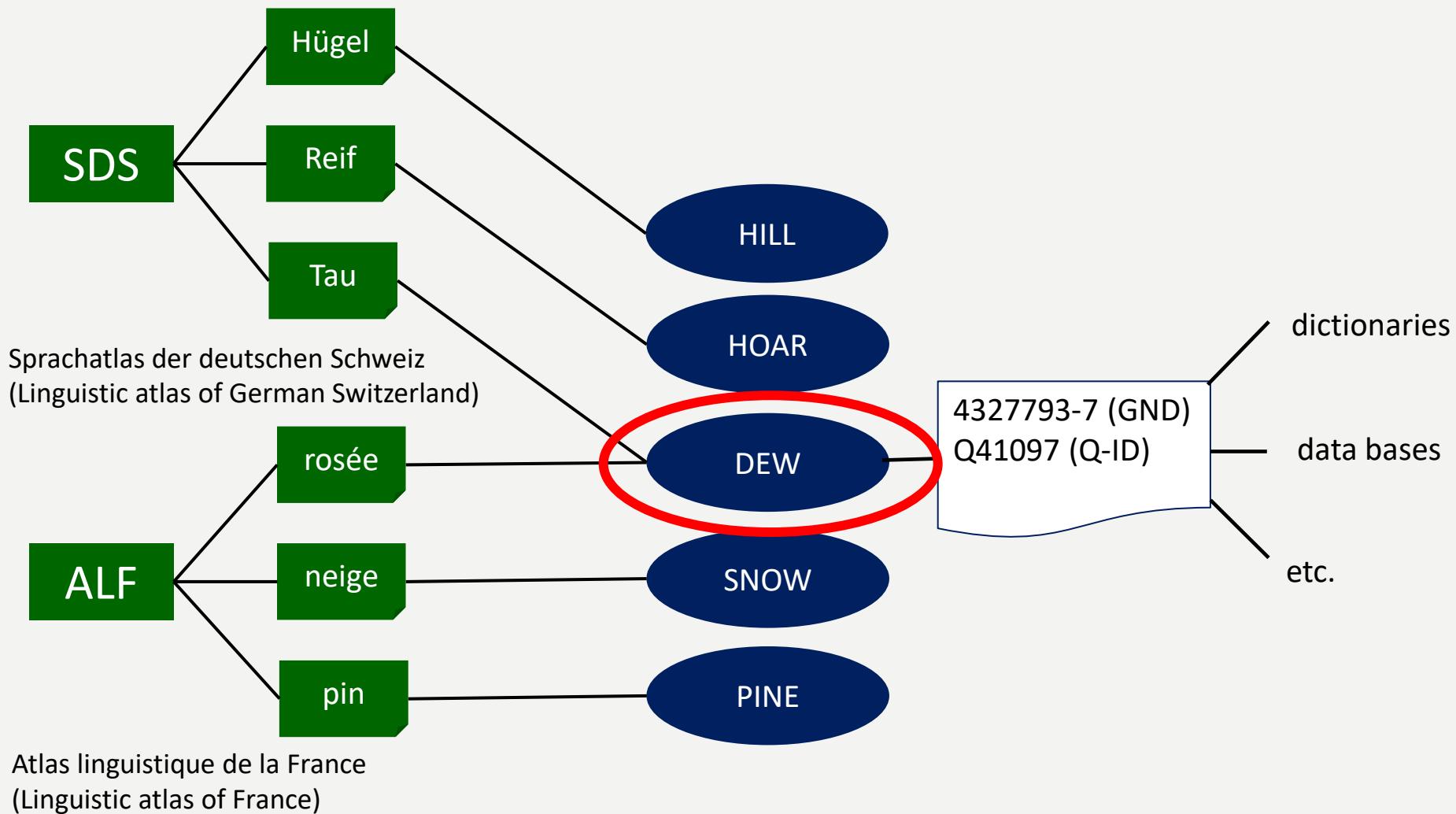
- **Q-IDs** of Wikidata (for concepts), partly also **L-IDs** of Wikidata (for morpho-lexical types)
- partly **GNDs** of the German National Library (for concepts) (*Gemeinsame Normdatei*, „Integrated Authority File“)
- **GeoNames** of www.geonames.org (for municipalities)
- **ISO Codes 639-3** (for 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)





Main page  
Community portal  
Project chat  
Create a new item  
Create a new Lexeme  
Recent changes  
Random item  
Query Service  
Nearby  
Help  
Donate  
  
Tools  
What links here  
Related changes  
Special pages  
Permanent link  
Page information  
Concept URI

Item Discussion

Read View history

Search Wikidata



English Not logged in Talk Contributions Create account



Wiki Loves Earth 2020 photo competition: take photos in nature and support Wikipedia.

## milk (Q8495)

white liquid produced by the mammary glands of mammals

[edit](#)

▼ In more languages

Configure

Language	Label	Description	Also known as
English	milk	white liquid produced by the mammary glands of mammals	
German	Milch	weißliche, undurchsichtige, als Milchfett-in-Wasser-Emulsion vorliegende, von Säugern produzierte Flüssigkeit	
French	lait	liquide biologique comestible produit par les mammifères femelles	
Bavarian	Muich	No description defined	



DEUTSCHE NATIONALBIBLIOTHEK

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

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

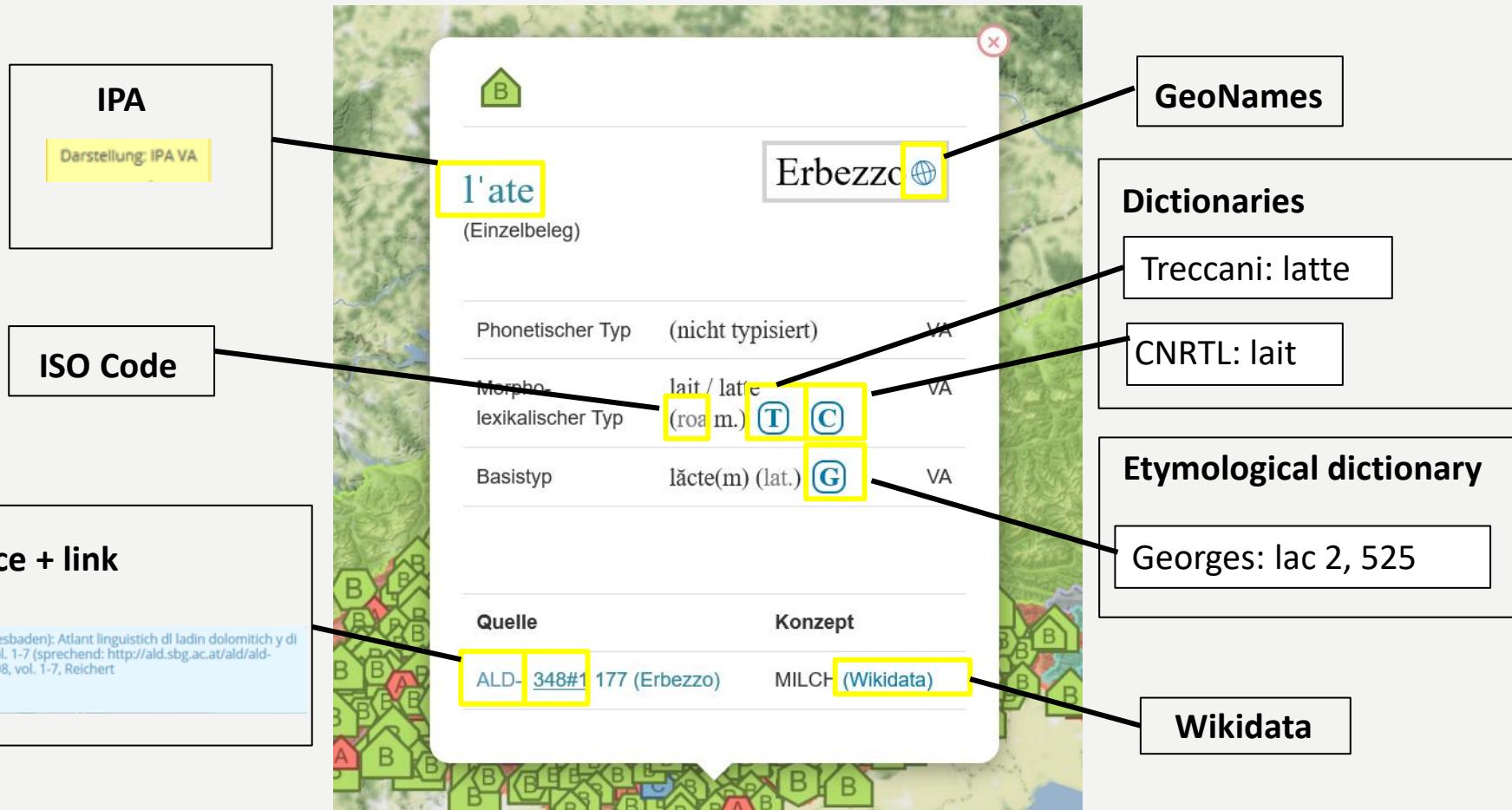
Login →

Über die Deutsche Nationalbibliothek

GND	
Link zu diesem Datensatz	<a href="http://d-nb.info/gnd/4009236-7">http://d-nb.info/gnd/4009236-7</a>
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)



## Detail view of one specific data point





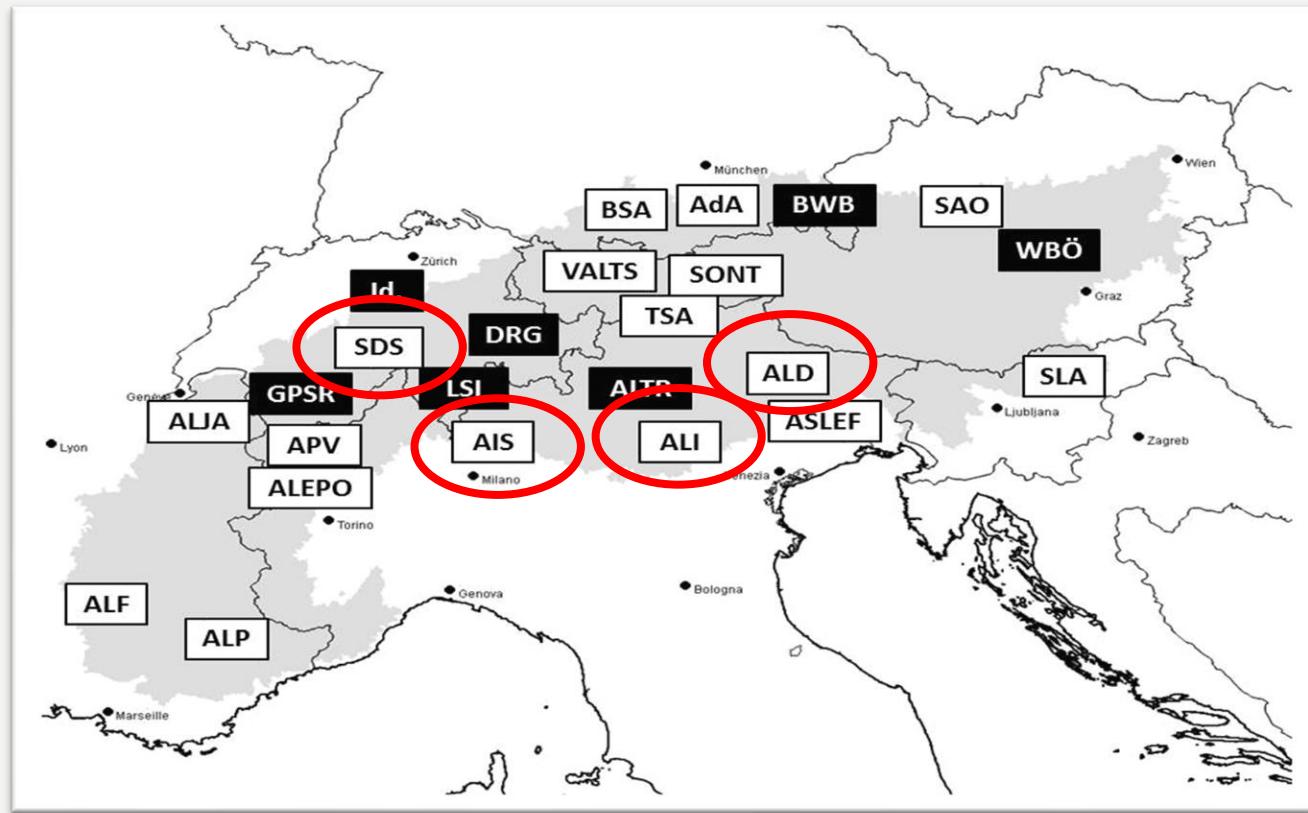
VerbaAlpina tries to create FAIR output

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

**Main reason:** lack of interoperability



## FAIRness of atlases in the Alpine region





## Comparison of the FAIRness of 4 geolinguistic atlases with VerbaAlpina

	Findable		Accessible		Inter-operable	Reusable	
	hum. read.	mach. read.	hum. 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

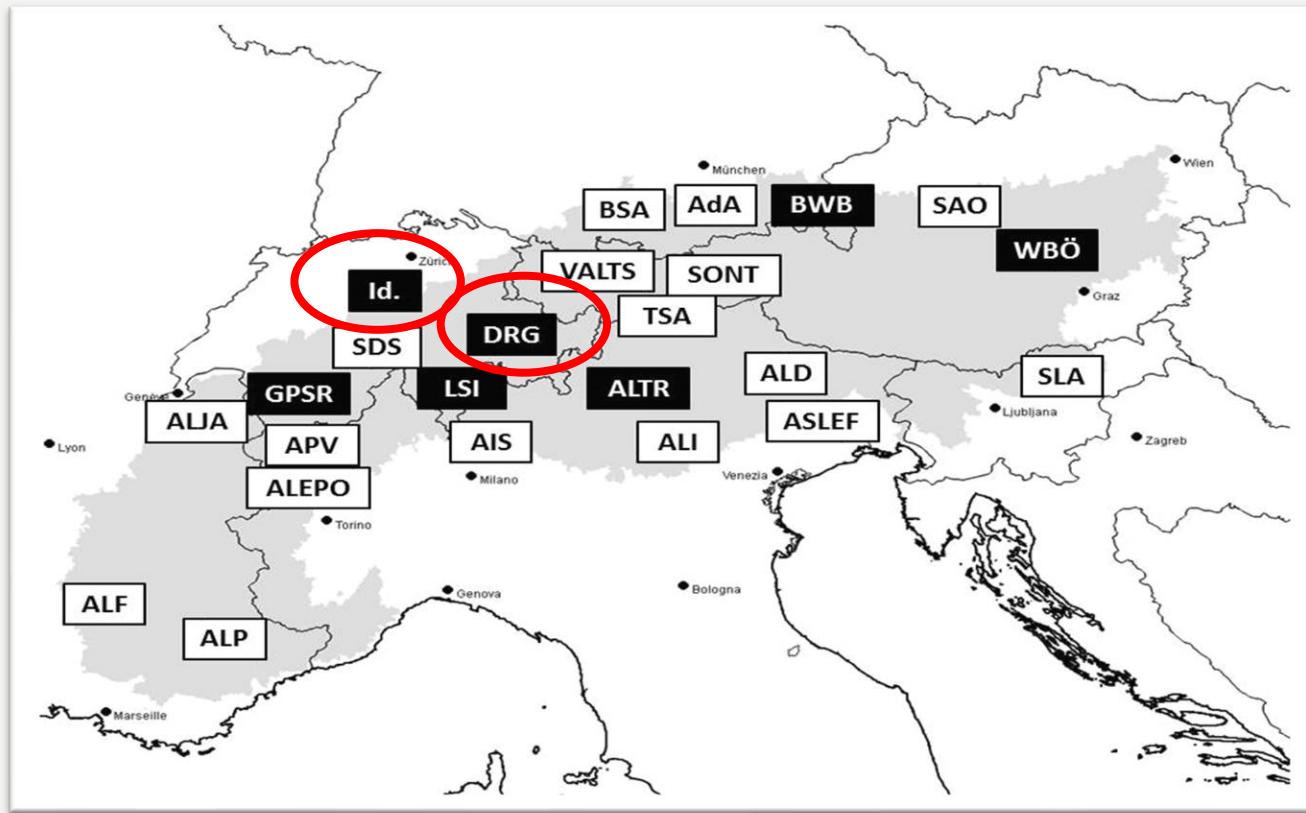


## Lack of FAIRness because...

- 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



**Bd XII 31** [gedruckt 1952]: énerdei ▼ · sélbdei ▼ · Tei N. ▼ · dei II ▼ · Deia N. ▼ · deianen ▼ · deidei I ▼ · deidei II ▼ · deider<sup>ch</sup> deidur<sup>ch</sup>en ▼ · Teien Teilen, Ti(e)jen ▼ · deiënnen ▼ · deihören ▼ · deihin<sup>n</sup> ▼ · deihin(d)e<sup>n</sup> ▼ · deioben ▼ · deiun(d)e<sup>n</sup> ▼

**Bd XII 32:** Suppe(n)teili ▼ · Deuängger ▼ · Theodor N., Tedi ▼ · Theodosius N., Dedli N. ▼ · Theophil N., Fili N., Töfel N. ▼ · di ▼ · dī ▼ · Diabel N. ▼ · didi ▼ · Tita ▼ · titä ▼ · titata ▼ · titi ▼ · Drückteien ▼ · Tuppenteili ▼

verloren gehen. 's ist dei-dei gangen B" (St.²). Aber derchüsing Gotts Willen, soll denn alles deidei gän, was mer so gnueg verdieneu! MWALDEN 1884. Nicht verdoppelt. Bist bald fertig mit dem Chrishacken? Es ruckt em, es ist wie däi ZTöftal. Es dunkt mich, das Fäfli sei bald lär. Ja, 's ist wie däi. ebd.

**Tei, Dei:** Matthäus (vgl. Bd IV 551/2, sowie Tewes), auch Amadeus, Taddäus (vgl. Tedē) W.

**Deia:** Andreas GrVal.

**Dieieldē** s. *Dialdē*.

**Teien** in GRD. (Bed. 1) *Tieien*, *Teiele*<sup>n</sup> (Bed. 2) GRMu. — f.: 1. a) Alp-, Sennhütte GrKl. (veraltet! It Tsch.); s. die Anm. „Welche Liegends kaufft hetten und daruff gebuwen, gezimmeret oder sonst gebesseret... es werend Hüser, Ställ, Tieyen, Wald oder was einer sonst daran gebesseret hette mit Dach oder nüwen Zünen oder Muren ...“ GrD. LB. — b) „Vorraum hinter der Tür mit Küche in den Alphütten“, Wohnraum, an den Heustadel angebaut GrD. (B.). — 2. übertr., gebrechliche, beschränkte, schwerfällige Weibsperson GrChur, D. (B.), He. (Tsch.), L., Mu., Schs, Valz. (Tsch.); Synn. *Noggen* (Bd IV 710); *Tschänggelen*, *Tschappe(le)n*. *Das ist en*

**Ti-ta** n.: Kinderspr., Uhr ZZoll.; vgl. *Tigg-Tagg*.

**ti-ta-ta:** Nachahmung des Taktes des Webstuhles (?). „Ach, der Lohn wird immer kleiner, und die Seide immer reiner, 's ist nüt mē mit däm Titata‘, Vers aus der Zeit, da es mit dem häuslichen Seidenweben zurückging. AfV. 24, 256 (ZMaur).

**ti-tä:** Nachahmung des Geräusches der Dachtraufe.

Rätsel: *Es gät öppis um's Hūs umer und chann nie iner, wānn men's schor wolt inen län, und macht ti-tä Z* (Dän.).

**di-di:** Lockruf für Zicklein, Katzen BsL.; SchR.; ZO. (Stutz); vgl. *de-de I. Halev chumm! Halev didi!* SchR.; s. auch Bd II 1768 o. (ZRafz). *Das ischt doch en Ströf. Die Tüfels Chatz!* [die Würste angefressen hat] Ich schlöhnev si z'Tod uf dem Platz. Wo ischt si echt? Chumm di di di! Stutz, Gem. Auch als Bezeichnung des Tieres BsL.

**ti-ti:** Beteuerungsformel, gewiß BHa.; vgl. *doi, duí*.

**Diabel** m.: 1. Name einer Kuh „mit spitzen Hörnern, die einer Gabel gleich aufgerichtet sind“ WUlr. — 2. Name (einer „Farbe“) im Kartenspiel WVt. — 3. leichter Reisewagen; s. Bd VII 360 M. (E. XVIII., SWbl.). — Frz. *diable*.

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



[online.drg.ch/main.aspx#35ac0b4c3795c91831f3534a40a0c5f0](http://online.drg.ch/main.aspx#35ac0b4c3795c91831f3534a40a0c5f0)

FsK — Milch Home - Milch SQL Fiddle Ilion-Troja Deutsche Dialekte i... Italienisch kostenlo... italiano/LeLingue/L... NavigAIS - AIS Navi... Google Maps lat/lo... TUO TL



Dicziunari  
Rumantsch  
Grischun

deutsch | rumantsch

Institut | Über das Wörterbuch | Angebot | Publikationen | Shop | Fototeca Online | DRG-Online

DRG-ONLINE ↓

«Zurück

## BARGUN

**Neue Suche**

Suchtipps

Abkürzungen

Vorworte

Hinweise

A

B

C

D

E

(DRG 02 / 192) | Artikel als PDF laden

Markieren Sie einfach grosszügig eine Textstelle, um die Band- und Seitenzahl anzuzeigen

Innerhalb des Artikels suchen

Suchen

**BARGUN** surselv., MARGUN oengad., m. 'Heustadel auf Bergwiesen, Alpstall, Alphütte auf der obersten Weidestufe' usw., siehe unten. Vgl. Karte B II, p. 179.

### I. *bargun*

C 1 *bargóij*, C 20 *bargéun*, C 21–23 *bargáun*, C 24, 3–4 *bargún*, C 45 *barvúij*, C 50 *bragúij*, C 60, 61 *bargúij*, C 62 *galbúij*, C 66 *balgúij*, *galbúij*, C 68, 69 *bargúij*, C 70 *bargíij*, C 81 *bargeun*, *barkéun*, C 93 *bargúij*, S 10, 13, 21, 25, 31–42, (64?) *bargún*. Zur Endung cf. Tab. 1 a. – Deutschbünden: Untervaz *pärgü*, Maienfeld *pärgü*, pl. *pärgü* (cf. auch MEINHERZ 116), Valzeina *pärgáu* neben *börgä*, Seewis, Fanas, Schiera, Firma, Tuzein *pärgáu*, Tenaz Kloster *pärgáu*, St. Antönien *pärgáu* neben *härgä*, St. Galler

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



## 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		
"written idioms"	Puter	Q688309	Upper Engadine	"Rhaeto-Romance" (Rätoromanisch)	roh
	Vallader	Q690226	Lower Engadine		
	Surmiran	Q690216	Surmiran-Albula		
	Sursilvan	Q688348	Sursilvan		
	Sutsilvan	Q688272	Sursilvan-Oberland		
spoken dialect	Jauer	Q690181	Sutsilvan		



## ▪ Concept assignment is sometimes unclear

Example: concept GHIACCIAO (GLACIER)

### *IL GHIACCIAIO*

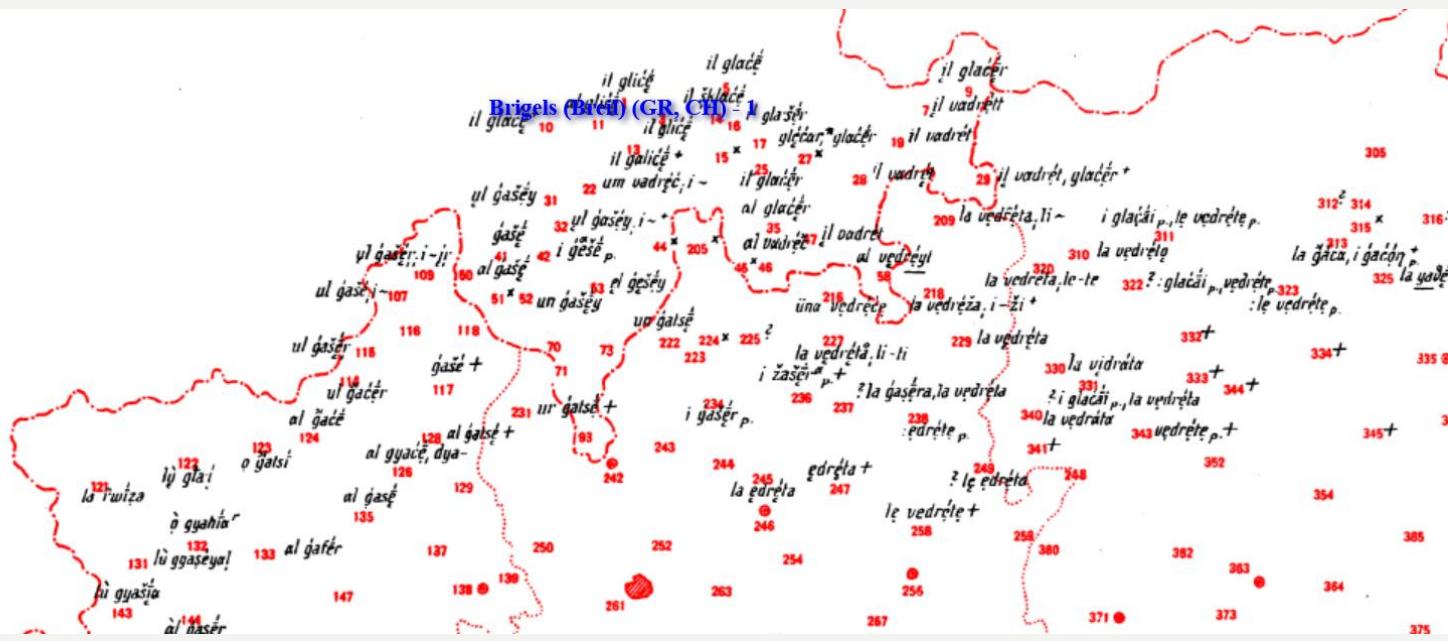
*GLETSCHER – GLACIER*

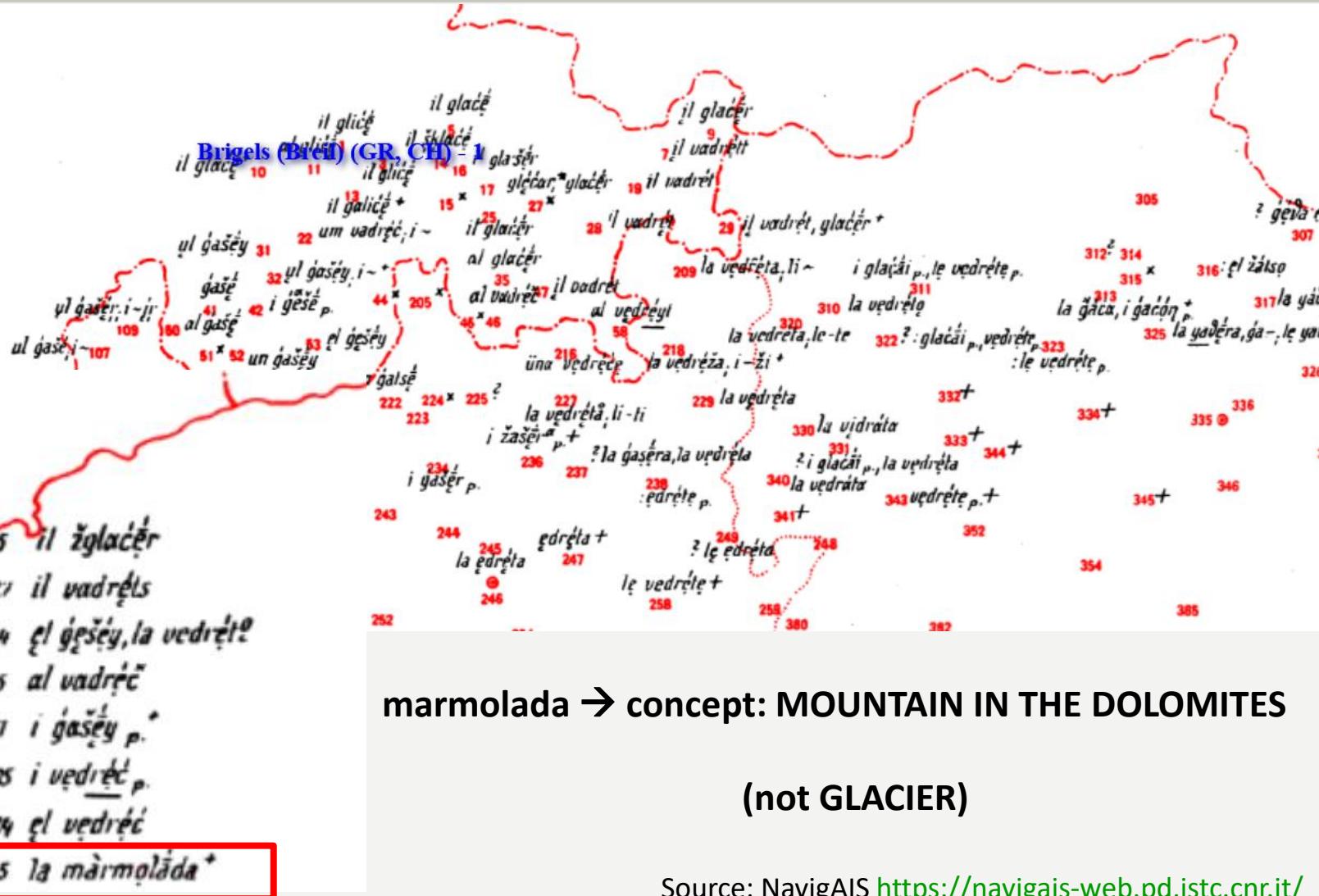
ALF Suppl. I. s. *glacier*

A 91,14 = 62,16 = 0

#### Legende:

- 10 *al vadréč* = alter Lawinenschluee über
- 13 *il vadréč* = ewiger Schnee; fñrleinem Bach.
- 29 *la myréna* - Moräne.
- 32 «Wenn der Gleitscher donnert (kwand u hät ul gáščy) gibt es bald darauf Hägelwetter.»
- 31 »vereiste Tächen an plattigen Stellen.
- 170 = grasse, im Frühling stets abschmelzende
- 218 *i lyéški* = Gleitscherspalten. /Eisfläche.
- 313 z.B. *i gáčq̊ de la marmoléda*.
- 315 Bezieht sich nicht bluss auf die Marmokida
- 317 z.B. *la yáda de l antélyw*.. del Monte Antelao.





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](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](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](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](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](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](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/>

