LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

VERBA ALPINA

Linked lexical data of different resources and countries in the Alpine region

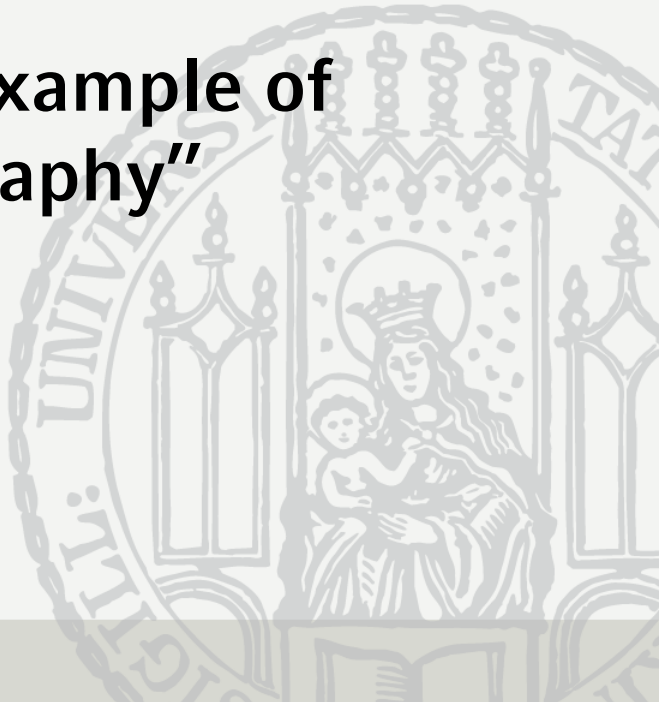
The project *VerbaAlpina* as an example of good practice in "smart lexicography"

Beatrice Colcuc | Christina Mutter

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

eLex Conference – Smart Lexicography

1 – 3 October 2019 – Sintra



Outline

1. Challenges of modern lexicography
   - Changes in research work
2. Project overview
3. Data processing
   - Transcription
   - Tokenization
   - Typification
   - Data access
4. Visualization of data
5. Linking of data

1. CHALLENGES OF MODERN LEXICOGRAPHY

Changes in research work

PAST

- **Analogue World**
  - *Physical obstacles to sharing data*
  - *Publications mainly printed*
  - *Central role of libraries*
- **Individual Data Collection**
  - *Each research project by itself*

PRESENT

- **Way from the analogue to the digital world**
  - *Digitalization → Change of mindset*
  - *Desire and need for sharing data*
- **Data collection with the aim of networking**
  - *Need for sustainability → sharing*
  - *Possibilty to publish and to access all data → more transparency*

1. CHALLENGES OF MODERN LEXICOGRAPHY

Changes in research work

PAST

- **Semasiological vs. Onomasiological**
  - ➢ *Only one perspective possible*
- **Data from one source only**

PRESENT

- **Semasiological and Onomasiological**
  - ➢ *Digitalization allows two perspectives at the same time*
- **Data from many sources**

1. CHALLENGES OF MODERN LEXICOGRAPHY

Changes in research work

FUTURE

- **Only digital data**
  - *Away from paper*
  - *Create more sharing possibilities*
- **Challenge of developing a uniform system for data processing**
- **Sustainability**
  - **F.**_*indable*
  - **A.**_*ccessible*
  - **I.**_*nteroperable*
  - **R.**_*eusable*

2. 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: 11/2017-11/2020 (perspective until 2025)
- Investigation of the multilingual Alpine region
- Combination of (geo-)linguistics and digital humanities

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 dialectical 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^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

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

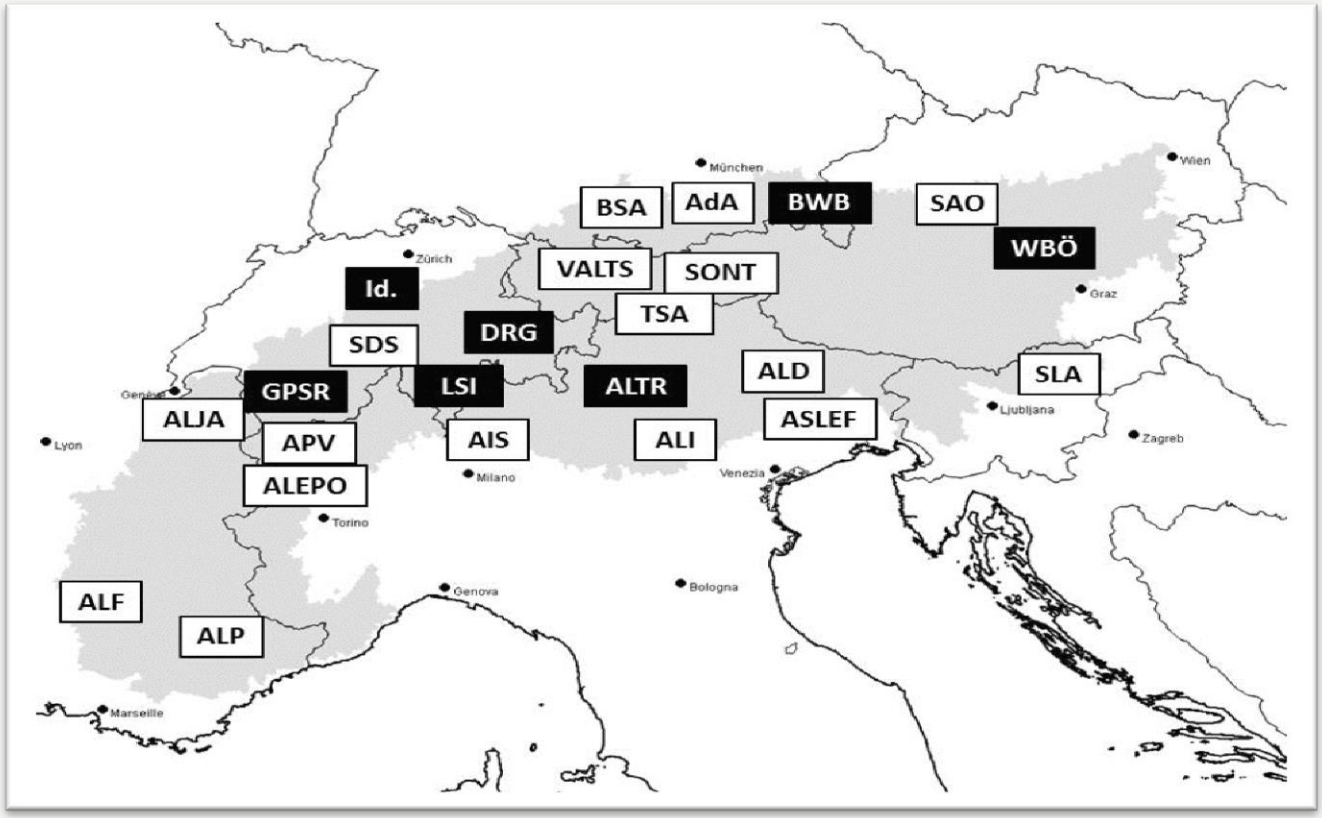
project phase	I	II	III
focus	**culture** • alpine pasture farming • milk processing	**nature** • landscape formations • weather • fauna • flora	**modern life** • ecology • tourism

Data

Multiple different sources

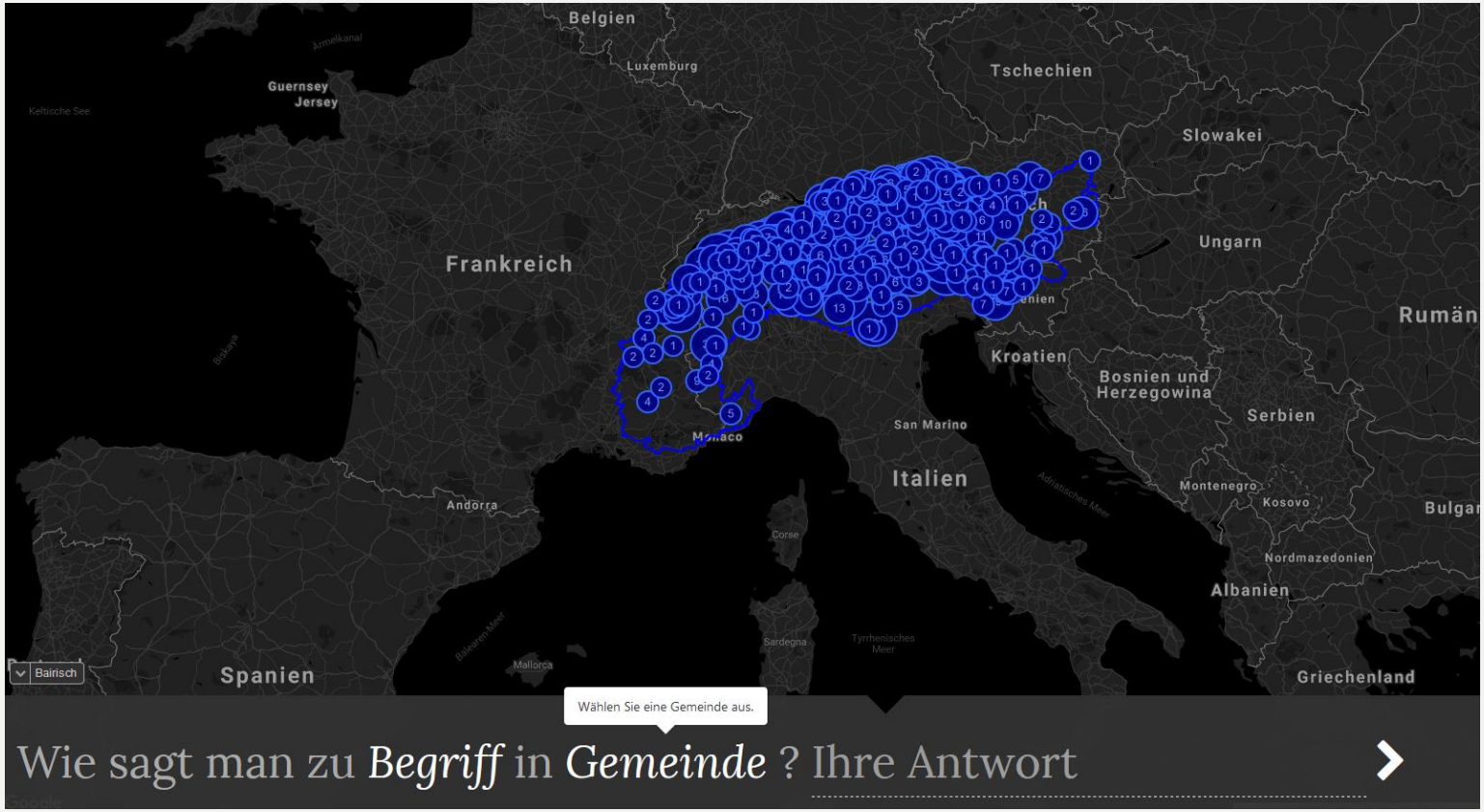
- printed atlases/dictionaries (georeferenced)
- Digital material from project partners
- Crowdsourcing

Atlases and Dictionaries in the Alpine region



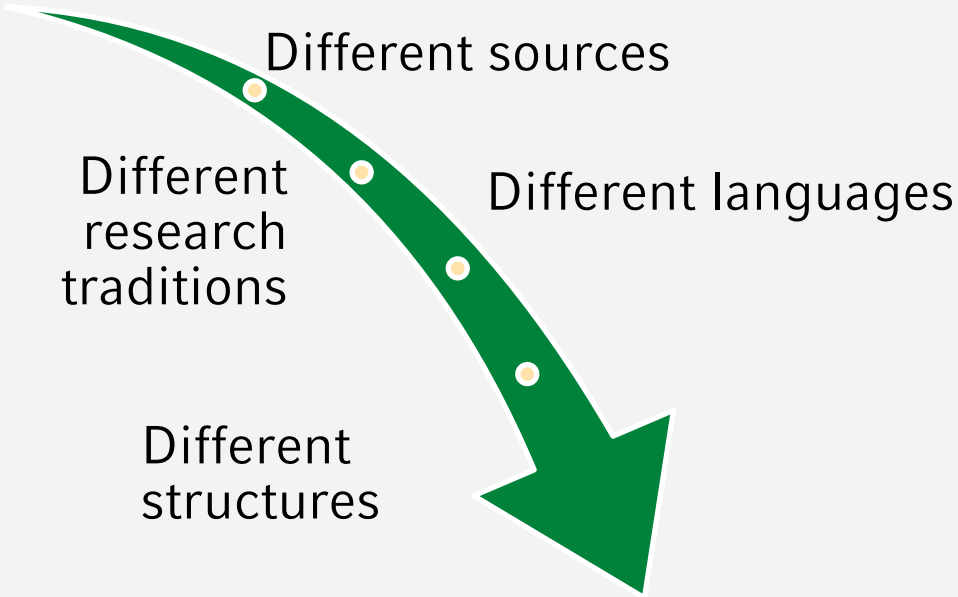
Crowdsourcing-Tool

www.lmu.de/verbaalpina



3. DATA PROCESSING

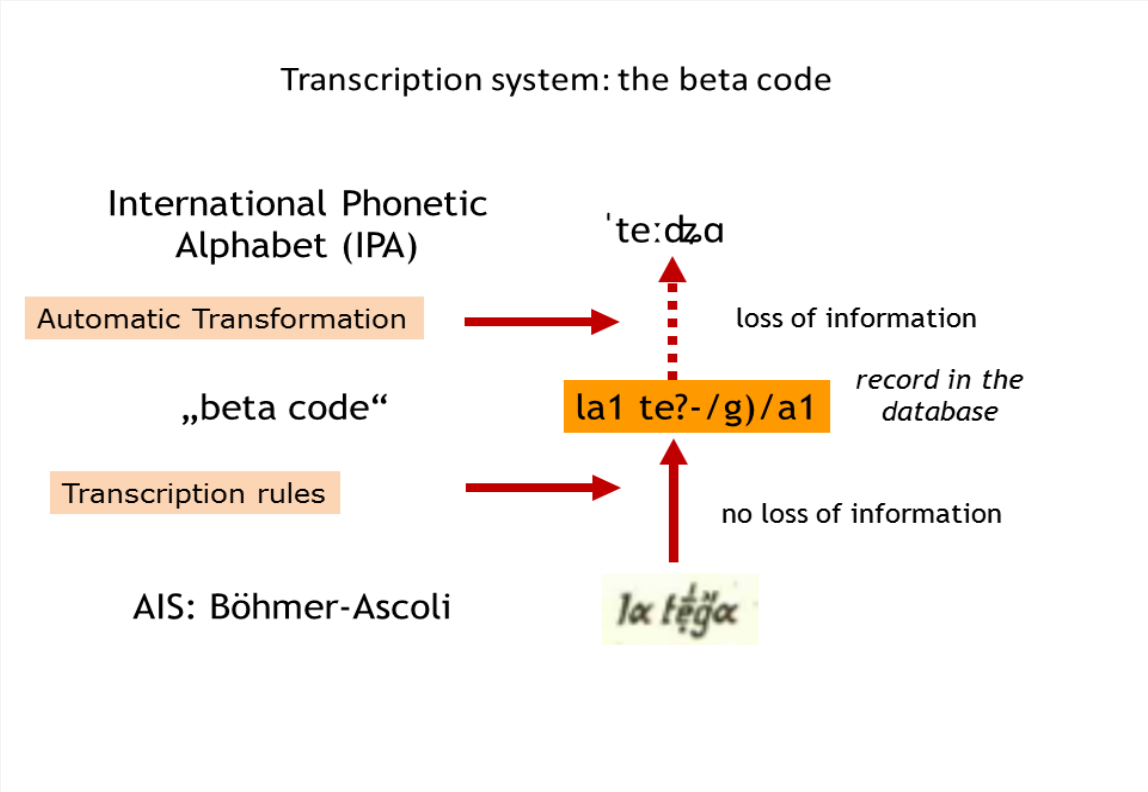
Transcription



→ Relational Database: MySQL

3. DATA PROCESSING

Transcription



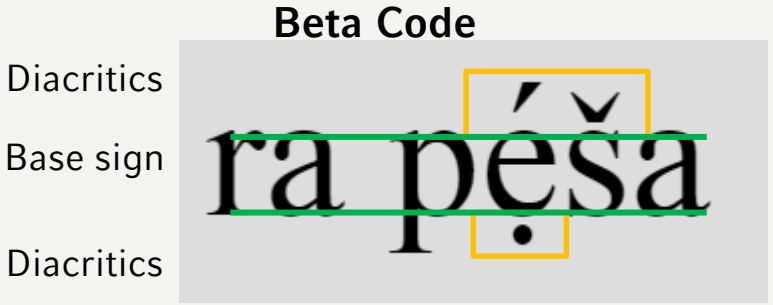
Principles of the transcription:

1. **Reliability of the source**

2. **Easy comparability**

3. DATA PROCESSING

Betacode



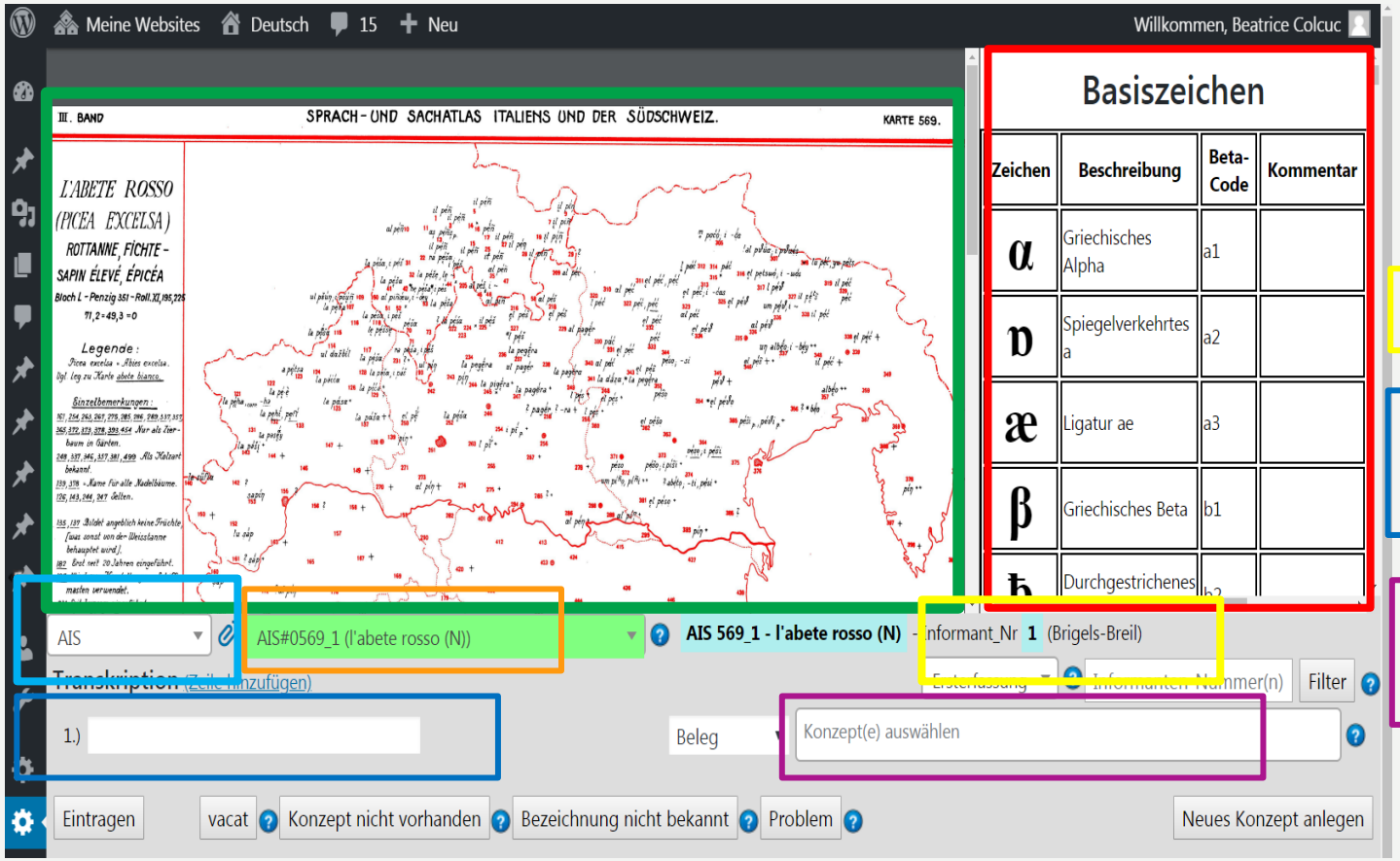
- The system uses only ASCII signs: the data can be transcribed with any computer and by any user
- For every source an unique codepage is created allowing the conversion into IPA

3. DATA PROCESSING

Transcription Tool

Original map

Transcription rules



Source

Stimulus

Informant number

Field for entering the transcription

Concept assignment

3. DATA PROCESSING

Tokenization

Attestation in BetaCode	Attestation in IPA	Concept
una1 mu:g/a1 da1 va/c)/	unɑ mydʒɑ dɑ vˈatɕ	HERD OF COWS
TOKENIZATION		
una1	unɑ	ARTICLE
mu:g/a1	mydʒɑ	HERD
da1	dɑ	PREPOSITION
va/c)/	v ˈatɕ	COW

3. DATA PROCESSING

Typification

Morpho-lexical type: Orthography, language family, POS, gender, affix + base type (lexical base)

	barga	barg	margun	bargun
Language family	roa	roa	roa	roa
Part of speech	noun	noun	noun	noun
Affix	-	-	+	+
Gender	f	m	m	m
Morpho-lexical type	1	2	3	3

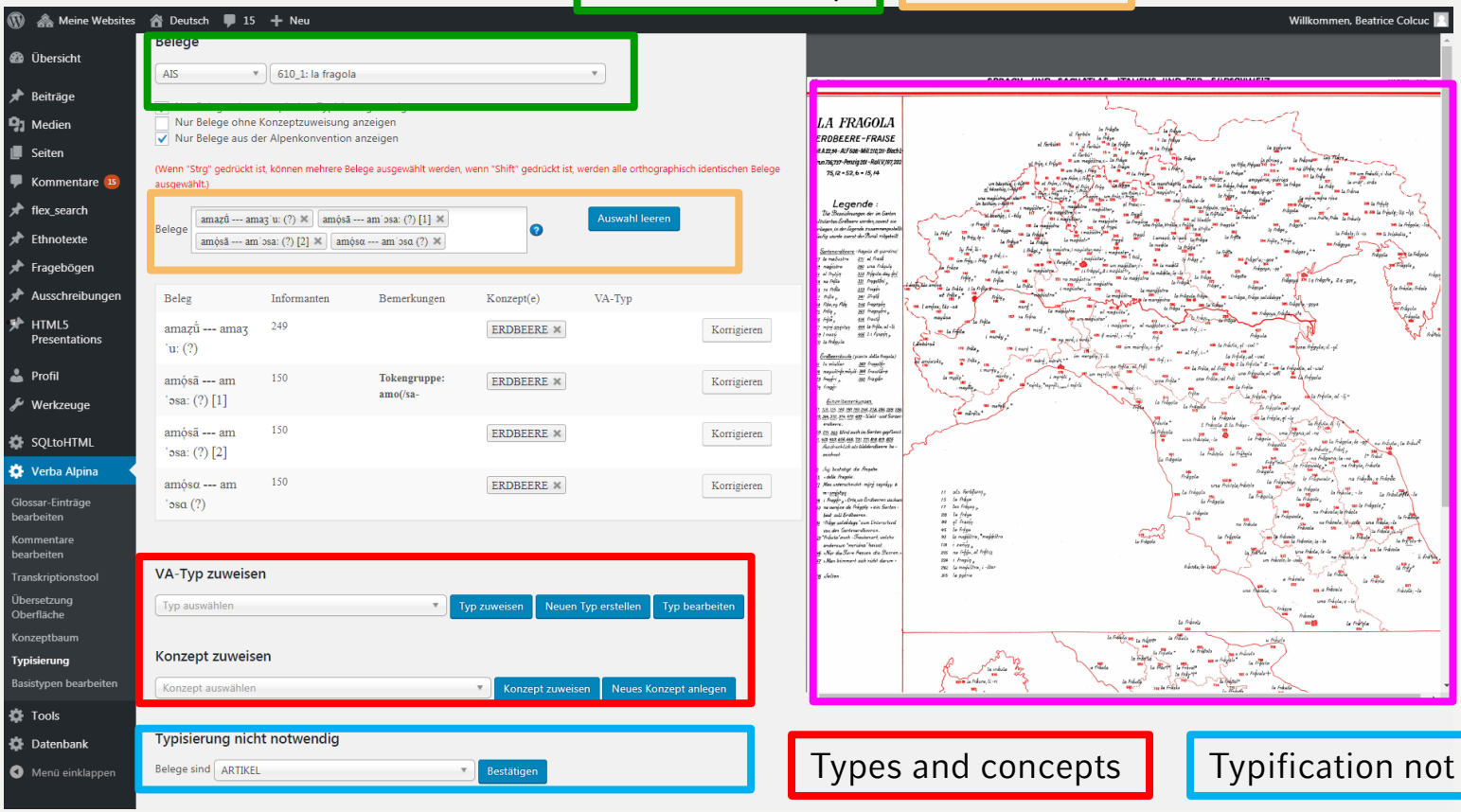
*barga/*barca

3. DATA PROCESSING

Typification tool

Source and map

expressions



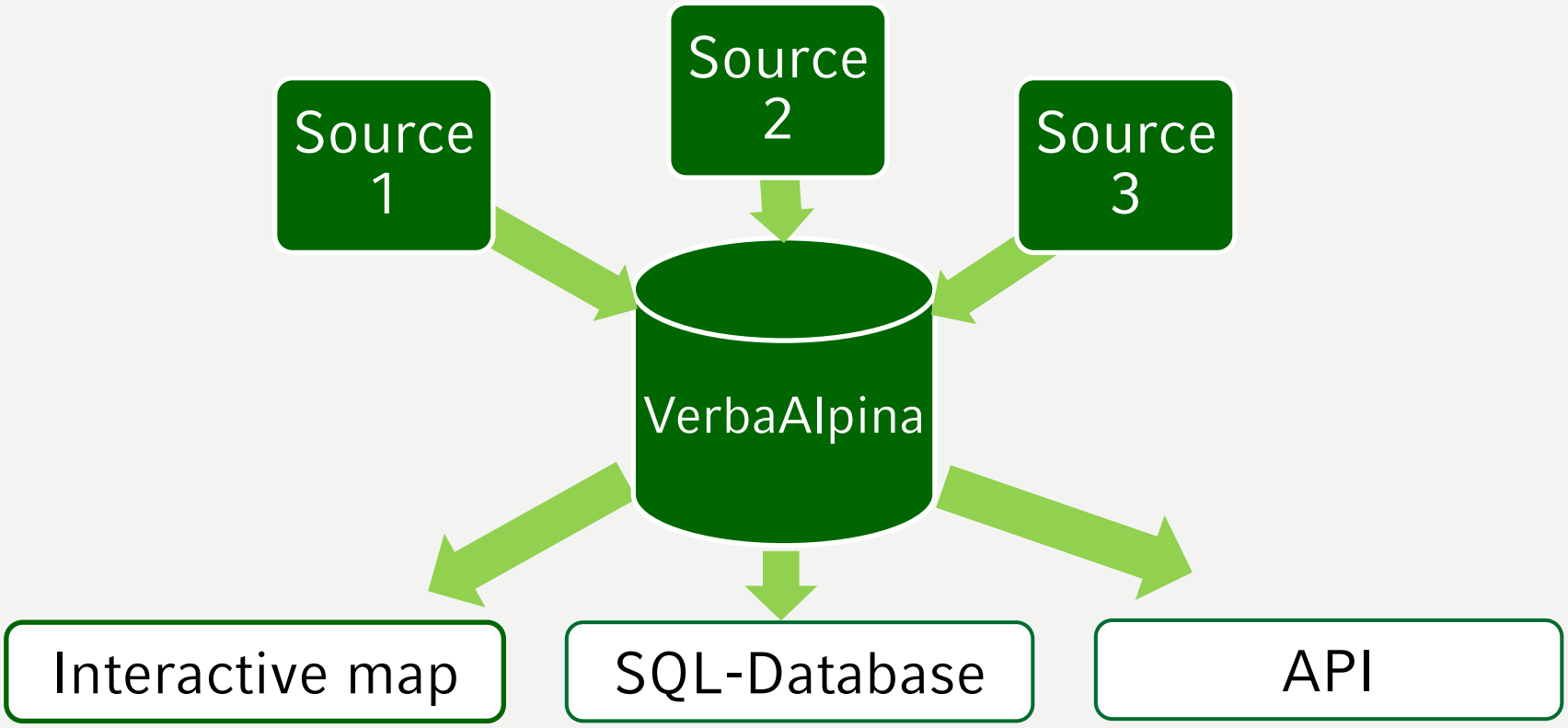
map

Types and concepts

Typification not necessary

3. DATA PROCESSING

Data access



4. VISUALIZATION OF DATA via the interactive map

- onomasiologically and semasiologically
  by using appropriate filters

- **4 different modes of visualization:**

  - geographical/physical
  - abstract
  - qualitative
  - quantitative

4. VISUALIZATION OF DATA via the interactive map

- onomasiologically and semasiologically
  by using appropriate filters

- **4 different modes of visualization:**

  - geographical/physical (NUTS 3 boundaries)
  - abstract
  - qualitative
  - quantitative

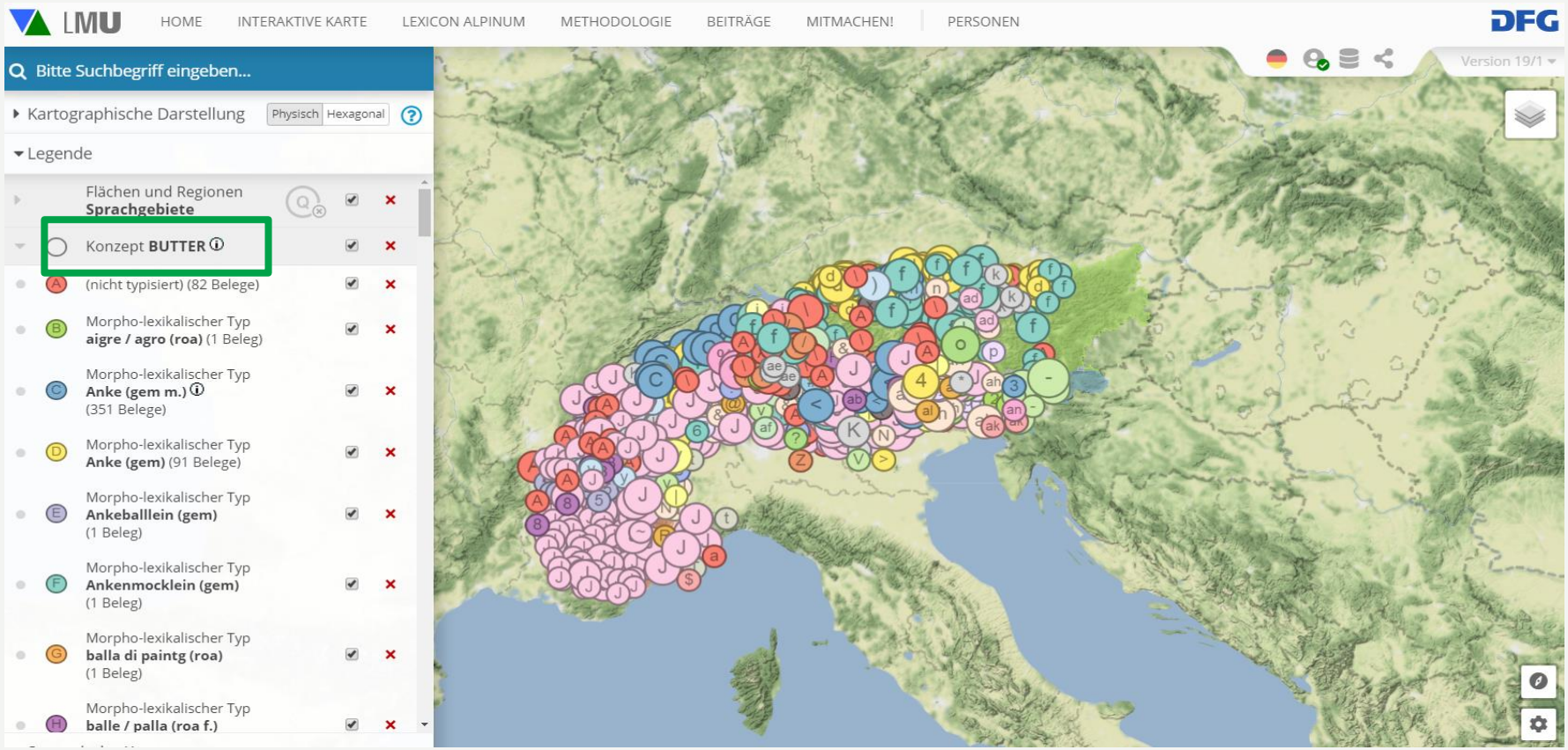
4. VISUALIZATION OF DATA via the interactive map

- onomasiologically and semasiologically
  by using appropriate filters

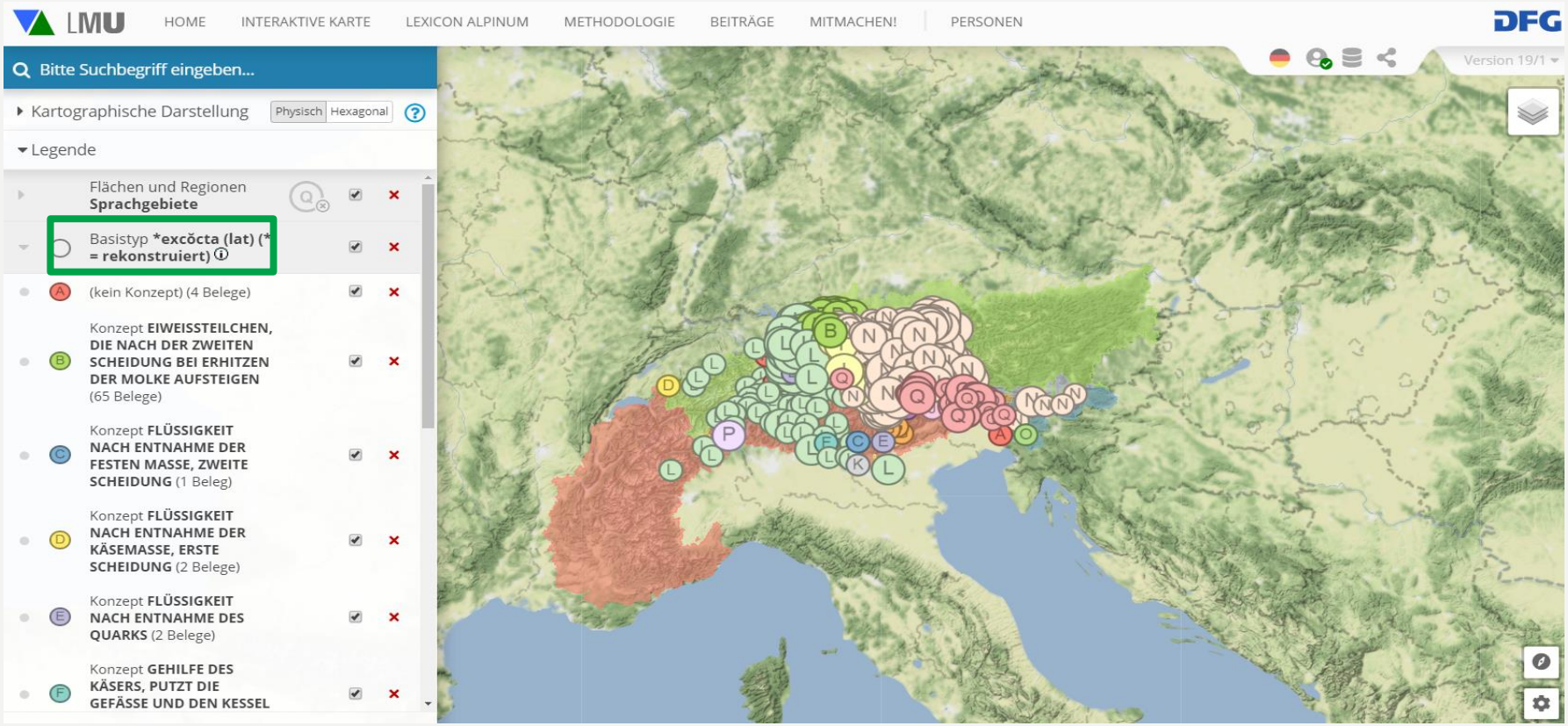
- **4 different modes of visualization:**

  - geographical/physical (NUTS 3 boundaries)
  - abstract (hexagons/honeycomb map)
  - qualitative
  - quantitative

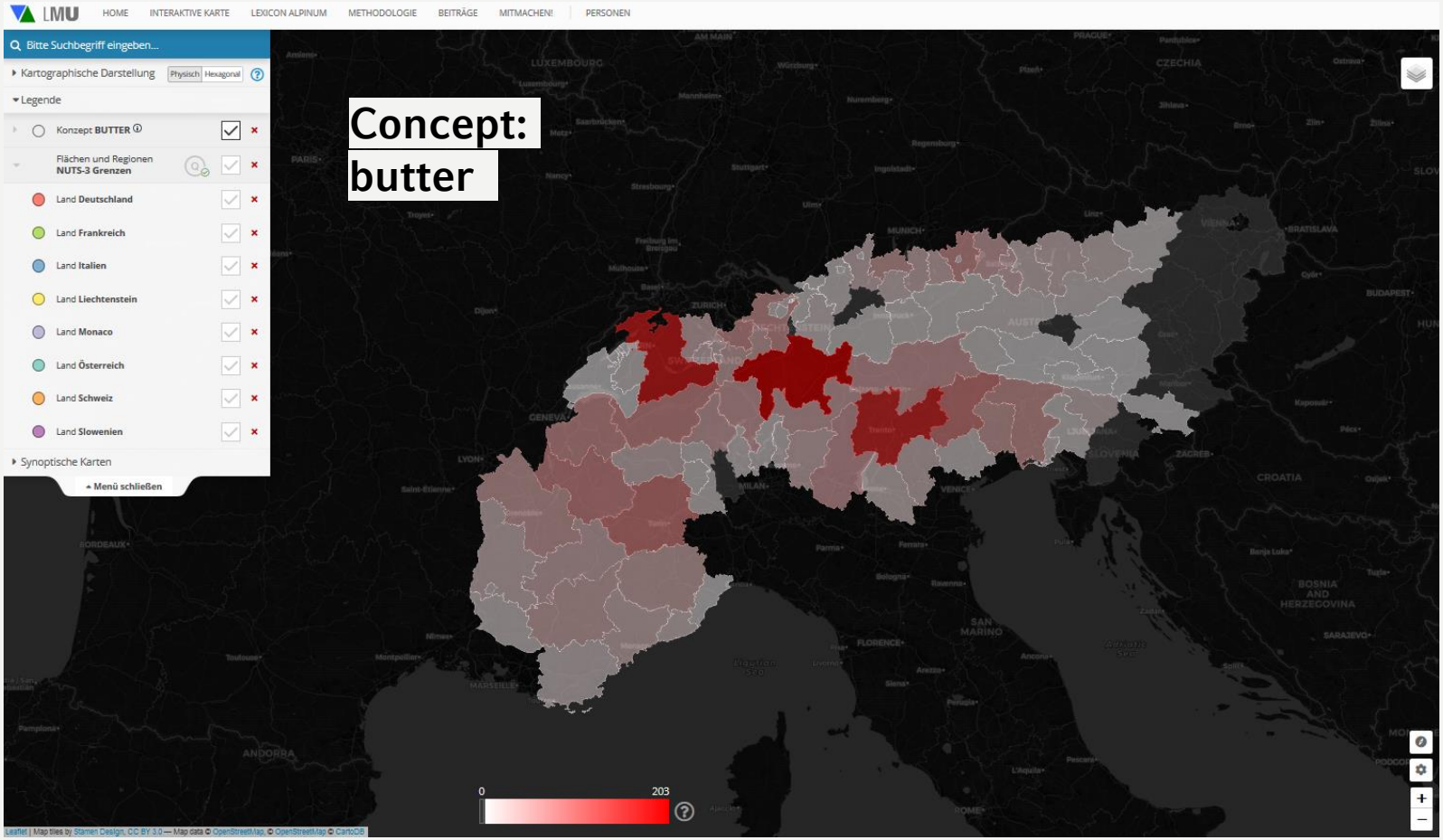
geographical qualitative (onomasiological perspective)



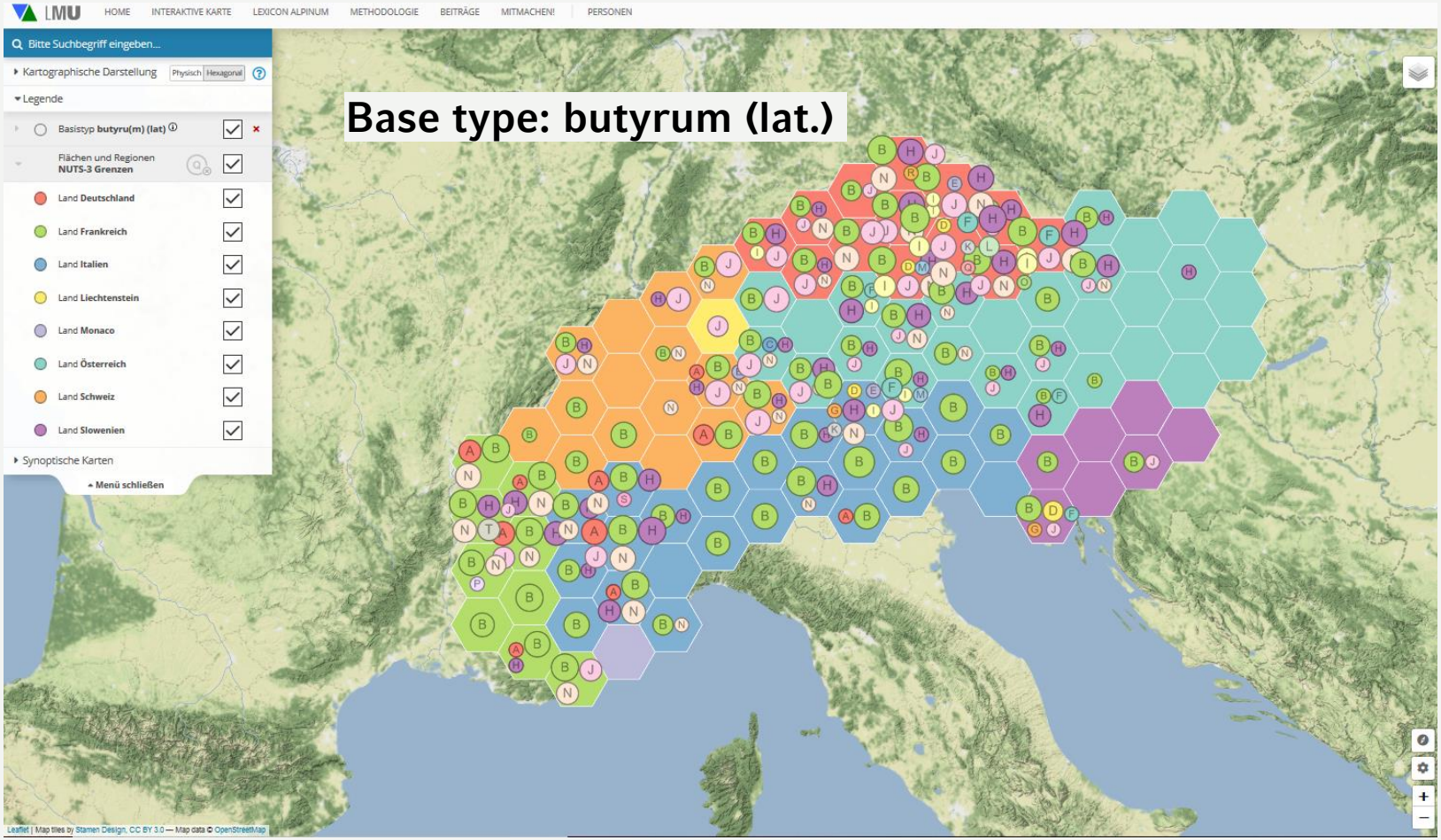
geographical qualitative (semasiological perspective)



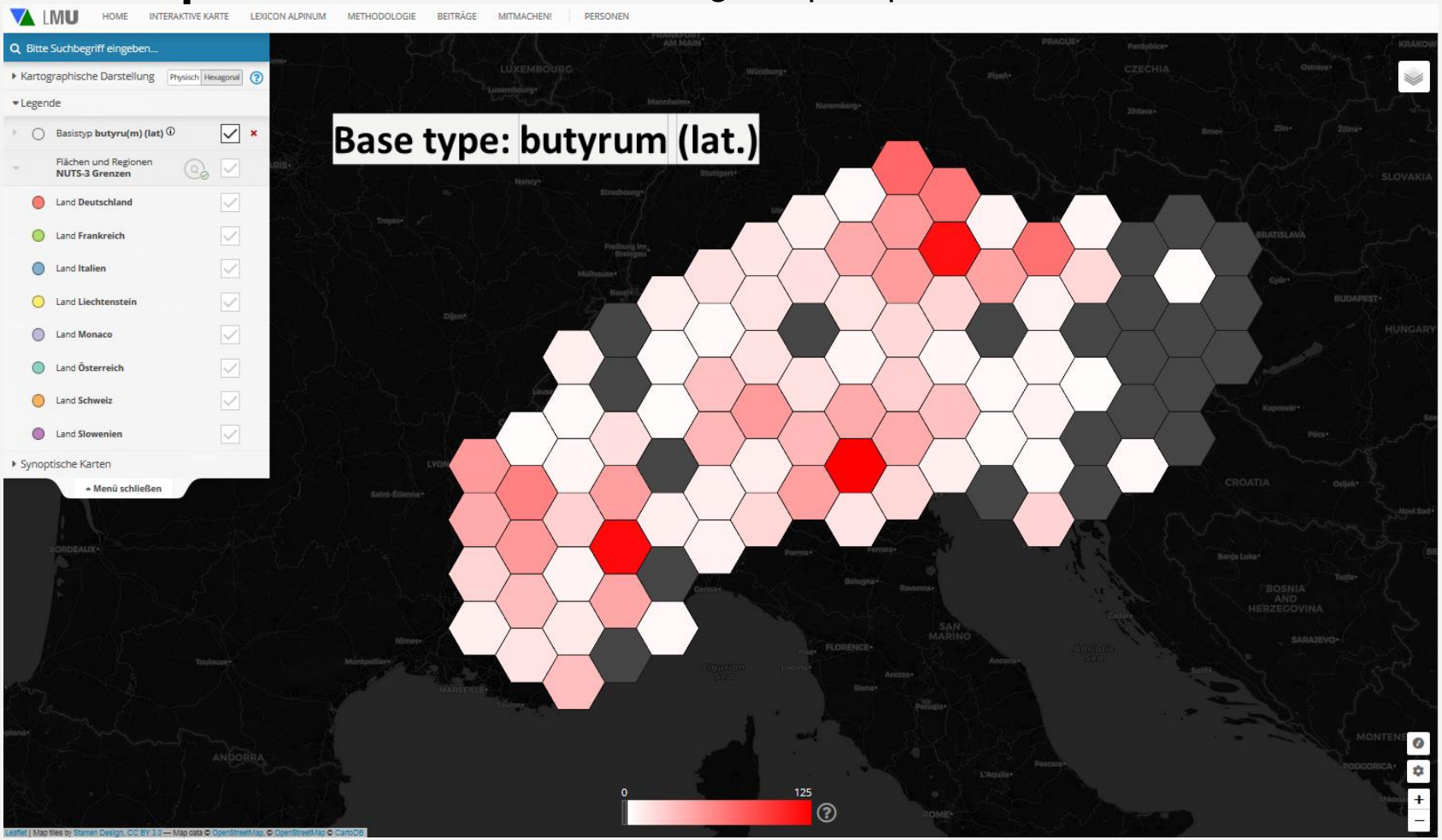
geographical quantitative (onomasiological perspective)



abstract qualitative (onomasiological perspective)



abstract quantitative (onomasiological perspective)



5. LINKING OF DATA

Guaranteed by:

1) Enrichment with **metadata**

2) Assignment of persistent **identifiers** of different knowledge data bases/norm data bases/reference dictionaries (GeoNames, Q-ID/L-ID, GND, DOI)

1) Enrichment with metadata

a) via the process of typification

→ assignment of basic types, morpho-lexical types, phonetic types

b) via the enrichment with sociolinguistic and sociodemographic data

→ language, municipality

(if available: age and sex of the informant)

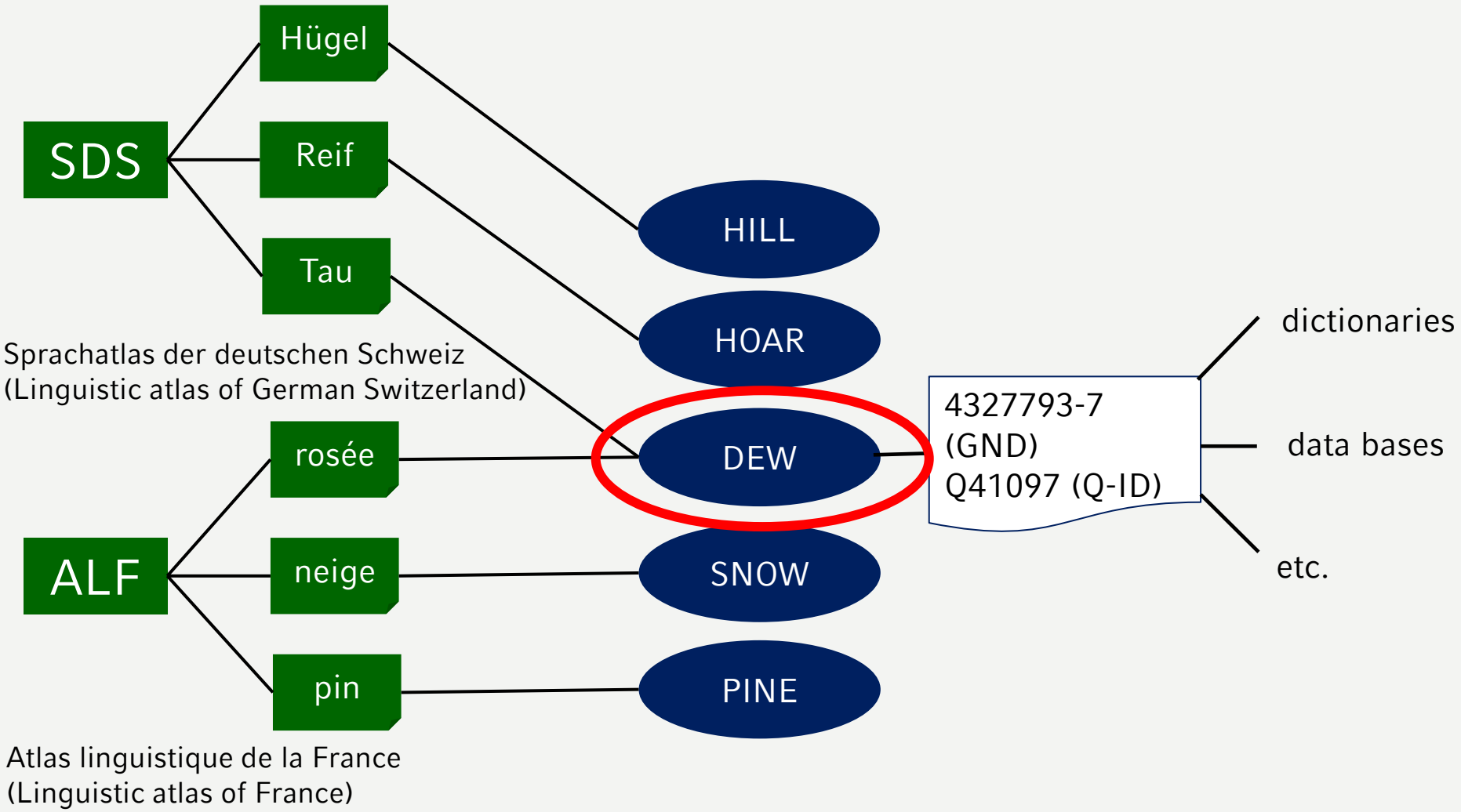
2) Assignment of different identifiers

OBJECTIVE: **Interconnection of VA data with data of other projects**

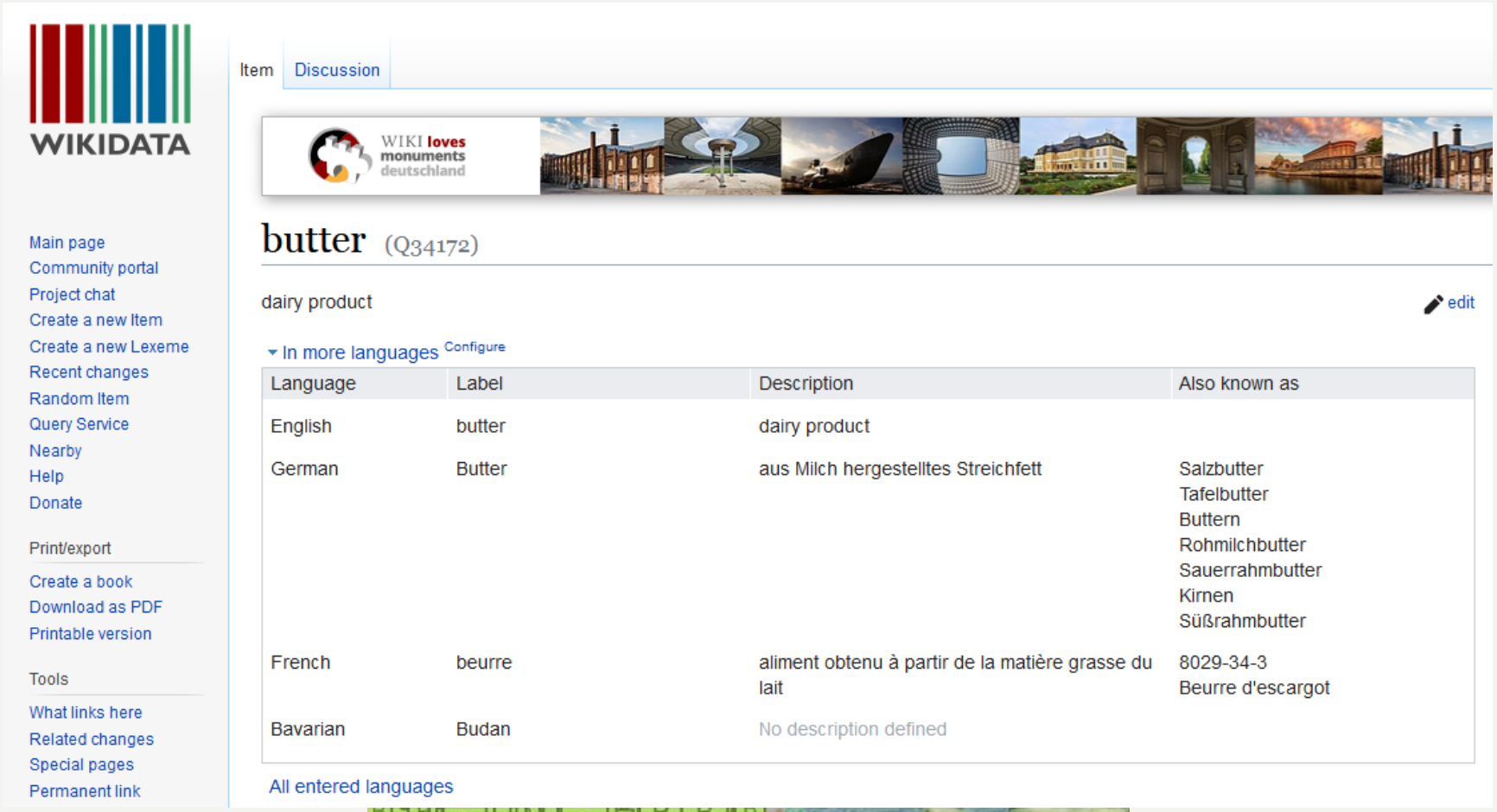
- **GeoNames** of www.geonames.org (for municipalities)
- **Q-IDs** of Wikidata (for concepts)
- partly also: **L-IDs** of Wikidata (for morpho-lexical types) + **GNDs** of the German National Library (for concepts) (*Gemeinsame Normdatei*, „Integrated Authority File“)
- **ISO-Codes 639-3** (for languages)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

Sintra 2262912	Q54050 (Q-ID)	L73260 (L-ID)	4135744-9 (GND)

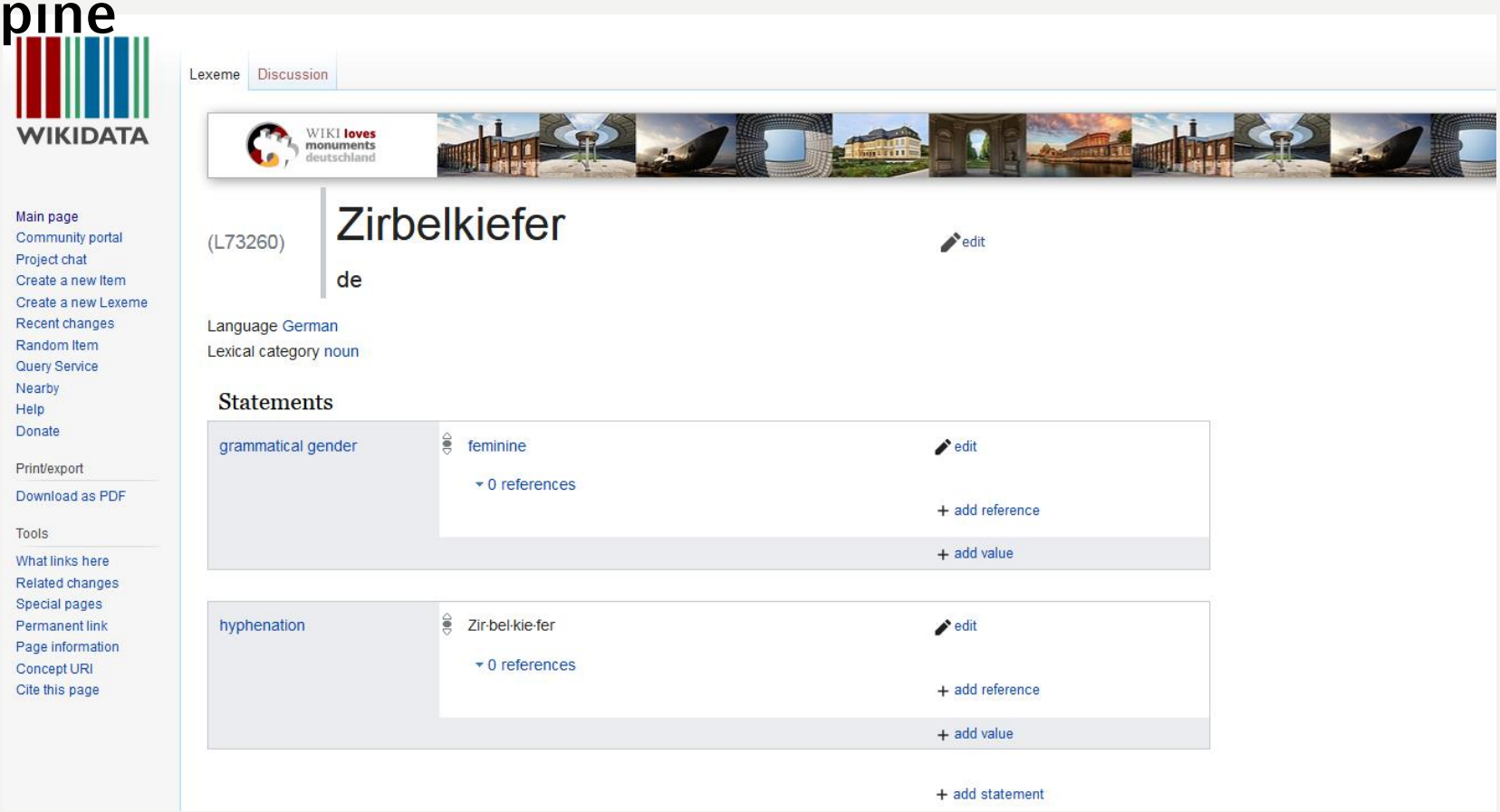
Linked lexical data of different resources and countries in the Alpine region



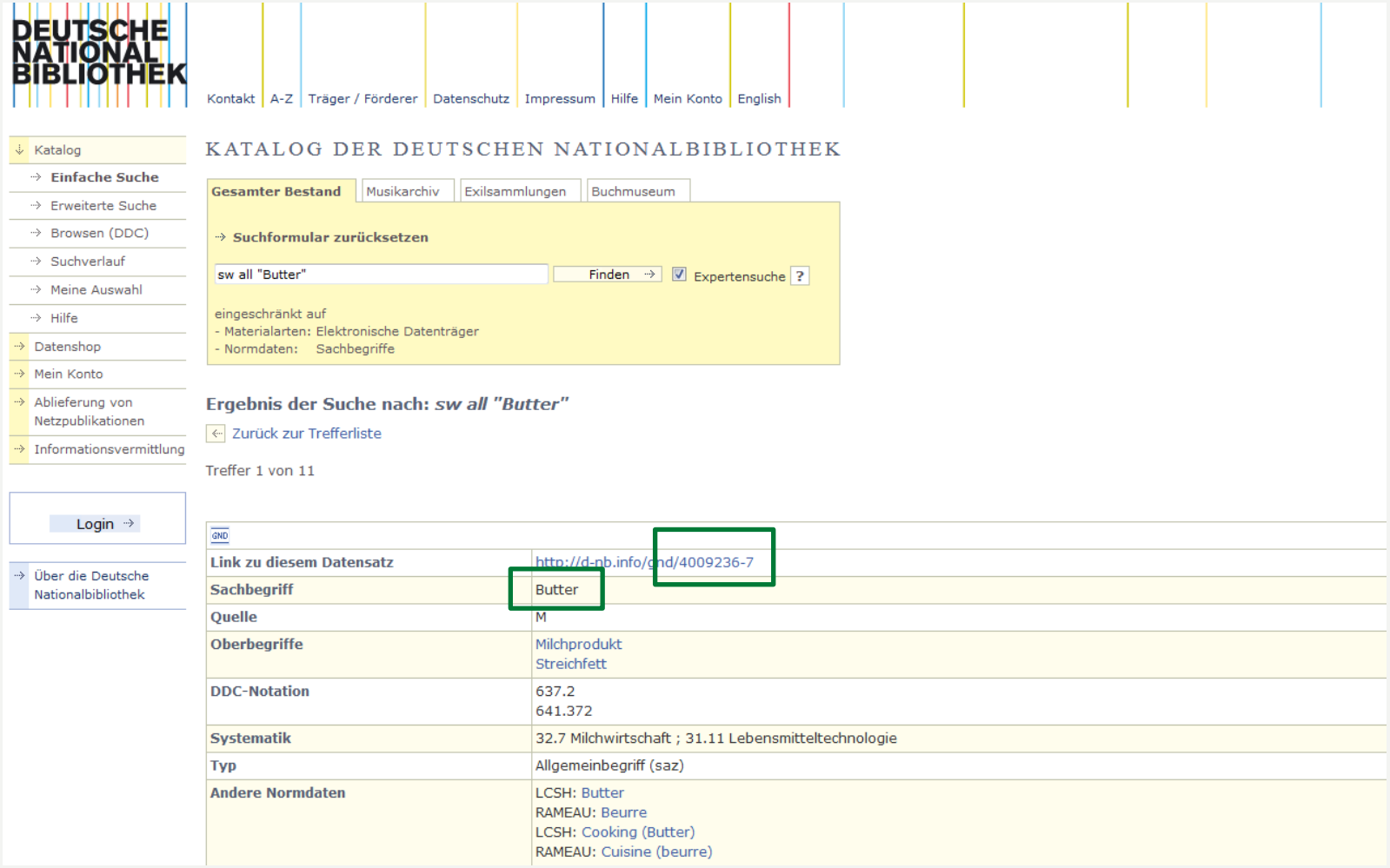
Linked lexical data of different resources and countries in the Alpine region



„Zirbelkiefer“ – Swiss stone pine



Linked lexical data of different resources and countries in the Alpine region



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

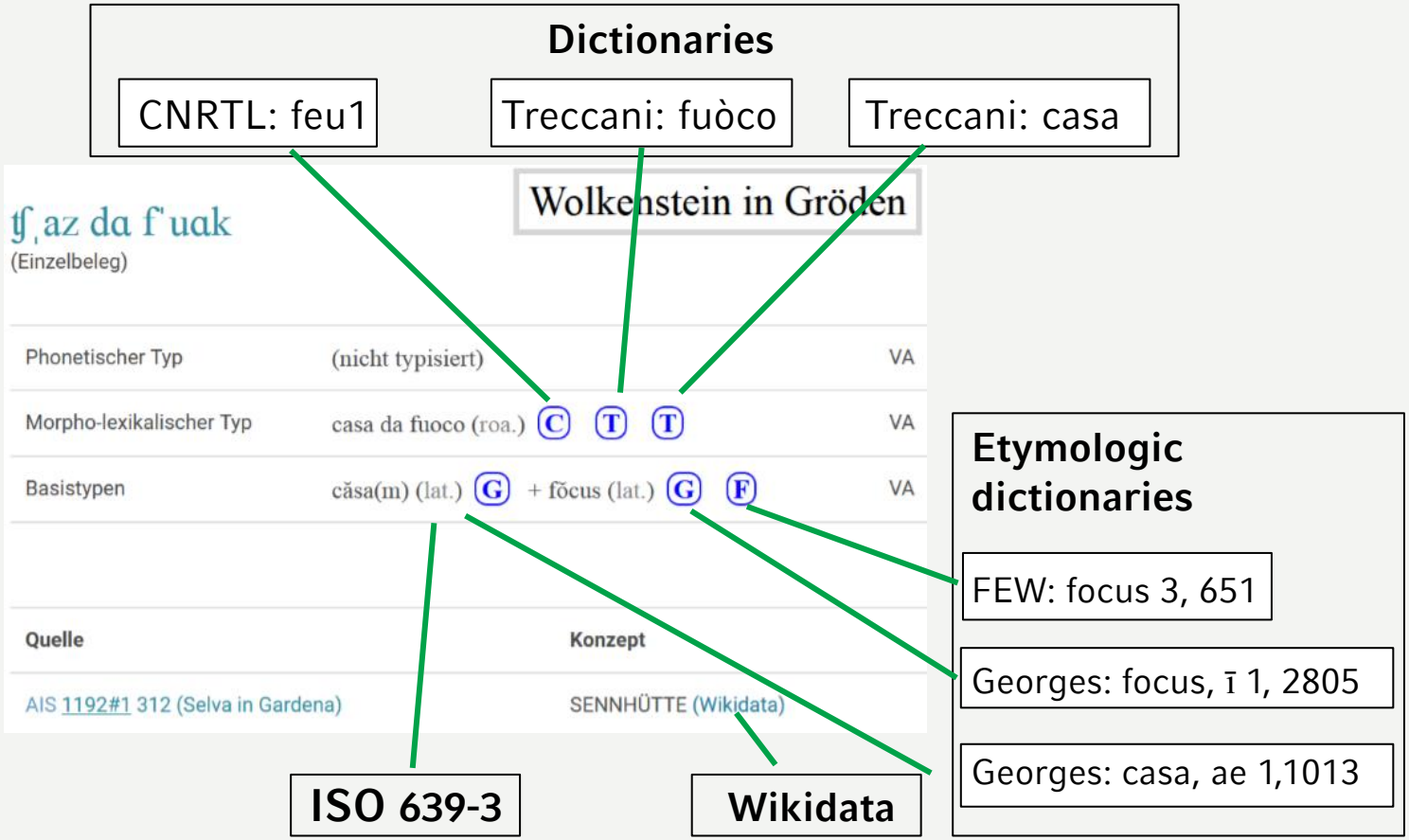
Treffer 1 von 11

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)

- **Identifiers of reference dictionaries**

→ at present 43 dictionaries (e.g. Treccani, FEW online, DRG, Duden, etc.)



- **DOIs** (Digital Object Identifier)

→ assigned to every single data
(attestations, morpho-lexical types, concepts, municipalities)

→ accessibility also if URL (Uniform Resource Locator) changes

Advantage: sustainable citability of electronic resources

Bibliography

Krefeld, T.: s.v. "Transkription", in: VerbaAlpina-de 19/1 (Erstellt: 16/1, letzte Änderung: 16/2), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DT%2357

Krefeld, T. / Lücke, S.: s.v. "Betacode", in: VerbaAlpina-de 19/1 (Erstellt: 16/1, letzte Änderung: 16/2), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DB%237

Krefeld, T. / Lücke, S.: s.v. "Typisierung", in: VerbaAlpina-de 19/1 (Erstellt: 16/1, letzte Änderung: 18/2), Methodologie, https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DT%2358

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

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

Thank you for your attention!

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