Markus Kunzmann, Florian Zacherl

Datengewinnung mittels Crowdsourcing im Dienste der Sprachwissenschaft

Die virtuelle Forschungsumgebung

VerbaAlpina

Digital Humanities Austria 2017
04. – 06. Dezember 2017, Innsbruck

Inhalt

- Projektbeschreibung
- Datenerhebung
- Eingliederung in den Datenbestand
- Visualisierung
- Nachhaltigkeit

Projektbeschreibung

Markus Kunzmann | Florian Zacherl

Projektbeschreibung

- *VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit*
- Förderung durch die Deutsche Forschungsgemeinschaft
- seit 2014, bewilligt bis 2020 (Perspektive bis 2025)
- Untersuchung des sprachenreichen Alpenraums:
  D, A, CH, I, F, FL, SLO, MC
- Umsetzung mittels moderner Konzepte und Methoden in der Verbindung mit den digitalen Möglichkeiten
  (*Digital Humanities*)

Projektbeschreibung

Projektleitung

- Prof. Dr. Thomas Krefeld (Institut für Romanische Philologie)
- Dr. Stephan Lücke (IT-Gruppe Geisteswissenschaften)

MitarbeiterInnen

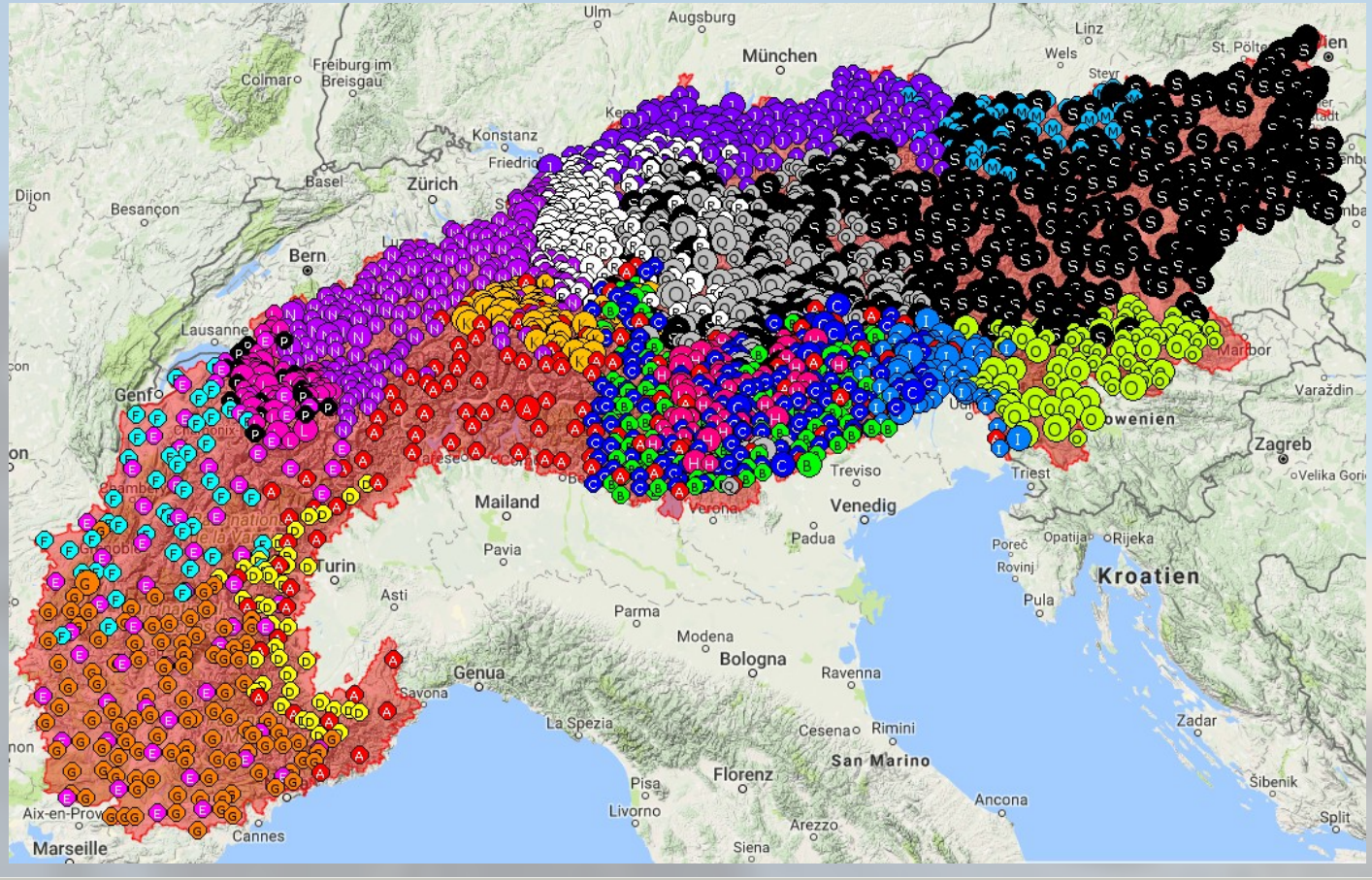
- Christina Mutter (wiss. Koordination)
- Markus Kunzmann (Germanistik)
- Aleksander Wiatr (Romanistik, Slowenisch)
- Florian Zacherl (Informatik)
- David Englmeier (Informatik)
- Monika Hausmann, Filip Hristov, Katharina Knapp, Marina Pantele, Daniela Warras (wissenschaftliche Hilfskräfte)

Projektbeschreibung

Projektjahre	1	2	3	4	5	6	7	8	9
Kalenderjahre	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Quartale	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

Projektphase	I	II	III
Schwerpunkt	**Kultur** • Almwirtschaft • ~~Volkstümliche Medizin~~ • ~~Traditionelle Küche~~	**Natur** • Landschaftsformen • Wetter • Fauna • Flora • traditionelle Küche	**moderne Lebenswelt** • Ökologie • Tourismus

Auflösung der Grenzen in Bezug auf sprachwissenschaftliche Teilprojekte



Auflösung der Grenzen hinsichtlich wissenschaftlicher Disziplinen

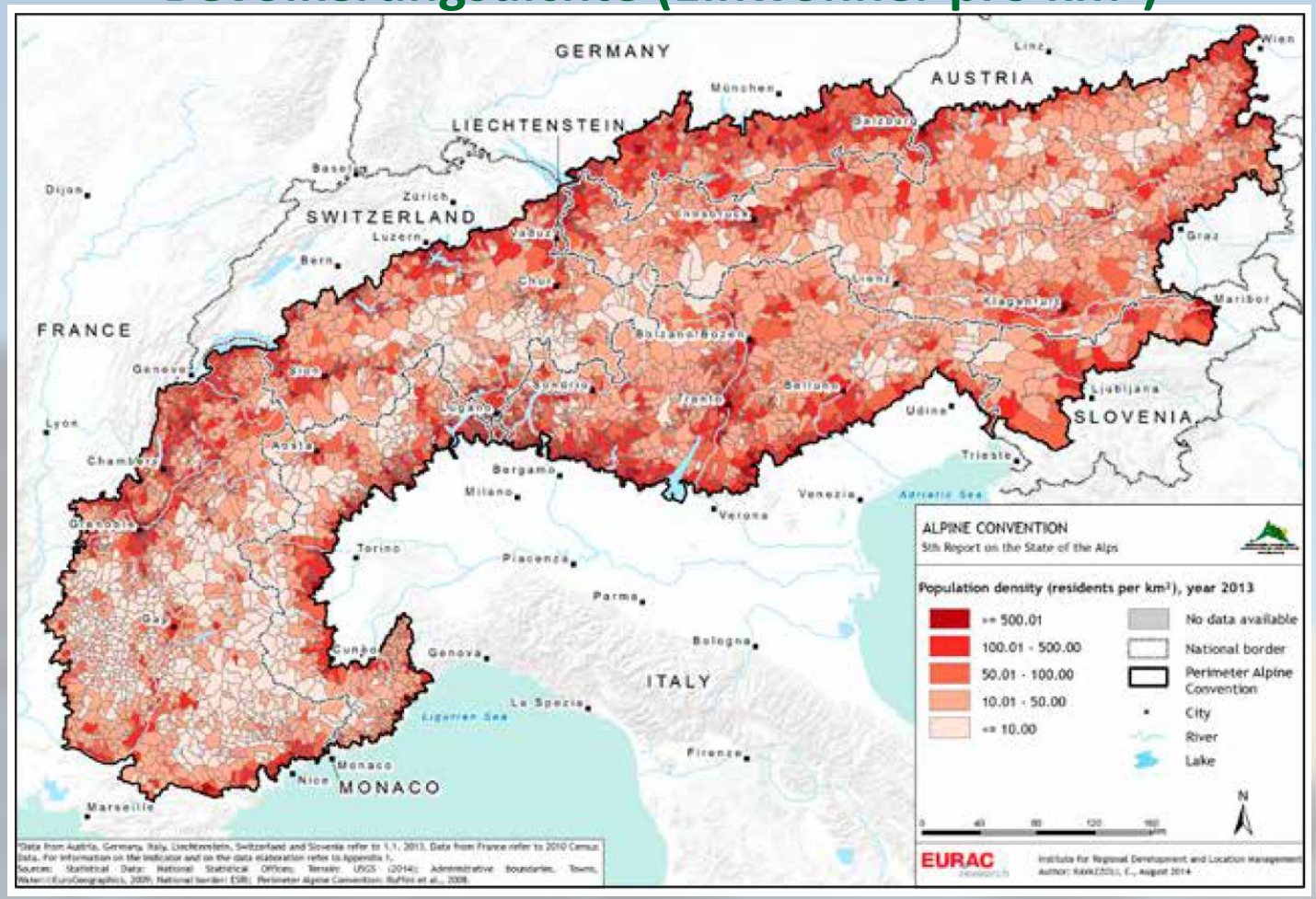
- Sprachwissenschaftler
- Historiker
- Ethnographen und Volkskundler
- Sprach- und Dialektinteressierte

Datenerhebung

Markus Kunzmann | Florian Zacherl

Datenerhebung

Bevölkerungsdichte (Einwohner pro km²)



Datenerhebung

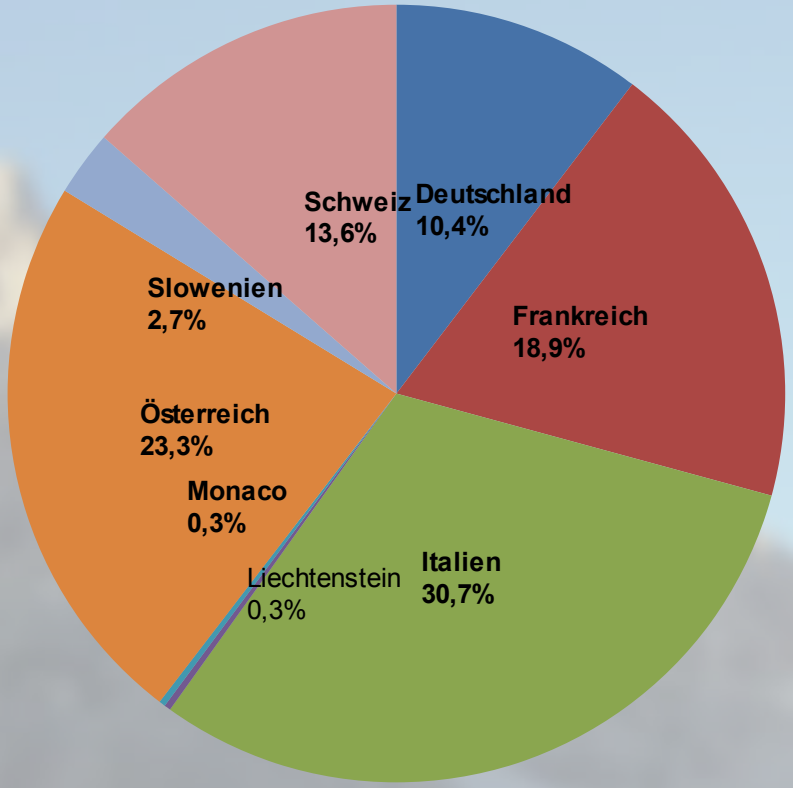
Bevölkerung, Fläche und Bevölkerungsdichte im Alpenraum

	Bewohner des alpinen Raumes	Fläche (km²) des alpinen Raumes	Bevölkerungsdichte im alpinen Raum	Bevölkerungsdichte national
Deutschland	1.476.519	11.160	132,3	225,3
Frankreich	2.683.801	40.801	65,8	103,4
Italien	4.364.538	51.995	83,9	201,8
Liechtenstein	36.838	160	230,2	230,2
Monaco	36.950	2	18.475	18.475
Österreich	3.318.045	54.592	60,8	100,8
Slowenien	385.973	6.796	56,8	101,6
Schweiz	1.929.424	25.211	76,5	201,0
Alpen	14.232.088	190.717	74,6	-

ALPENKONVENTION (2015), S. 17

Datenerhebung

Bewohner des alpinen Raumes nach Staaten



EW Alpenraum:
14.232.088

Crowdsourcing Anforderungen

Welche Daten werden erhoben?

- Konzept
- Gemeinde
- Sprachbeleg (aktuell über Texteingabe)
- Implizit Datum

Zukunft?

- Genus bei Substantiven
- Sprachaufnahme
- phon. Ebene des Materials

Crowdsourcing Anforderungen

Zielgruppe

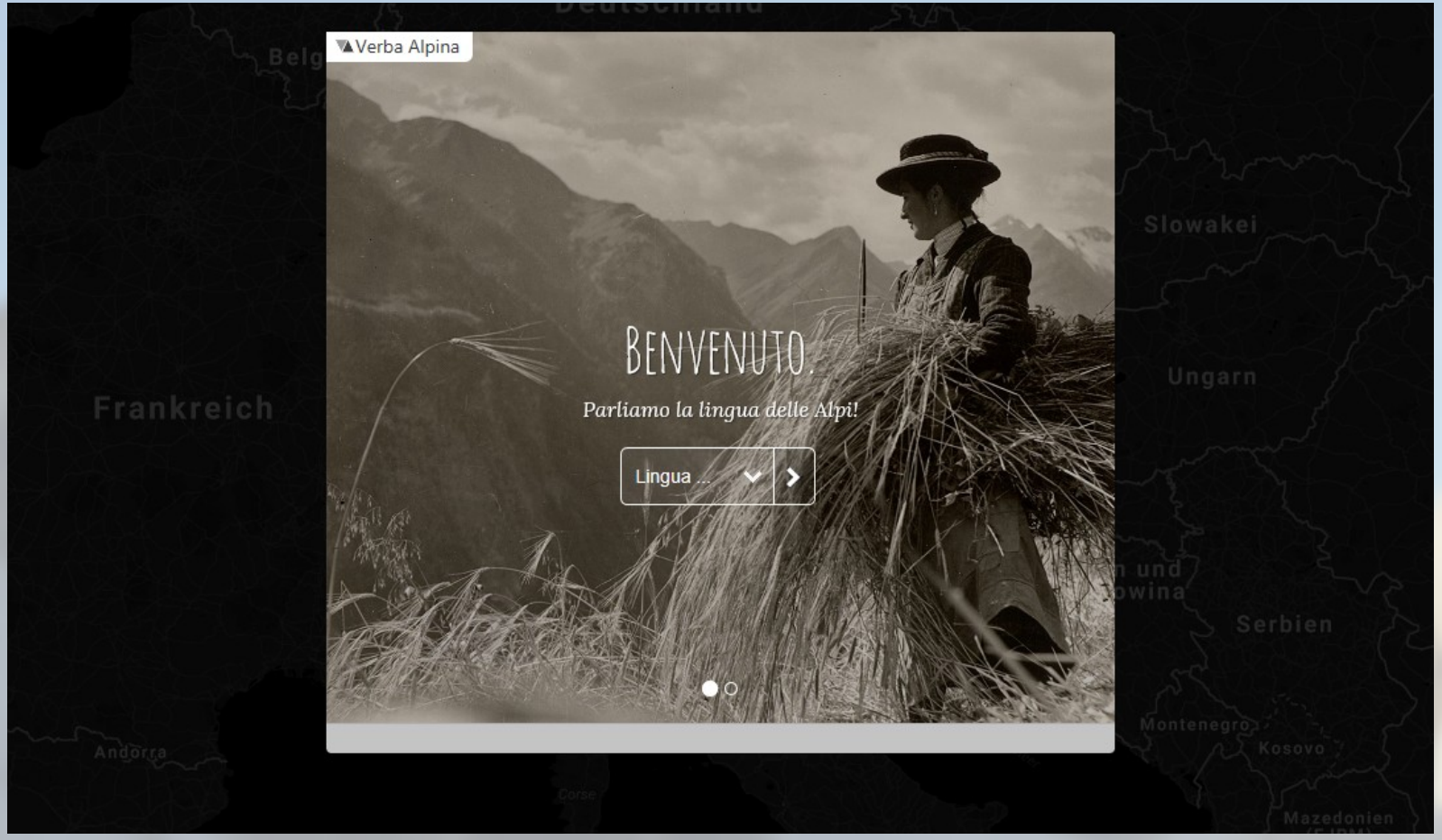
- Wissenschaftler und Laien
- Verzicht auf wissenschaftliche Terminologie

Bedienbarkeit

- Möglichst einfache Bedienbarkeit für mobile und Desktop-Geräte
- Simples responsives Design
- Registrierung nicht notwendig, Nutzer wird aber auf erweiterte Möglichkeiten hingewiesen

Datenerhebung

Crowdsourcing-Portal



Crowdsourcing Mailingaktion

Verlage
- Zeitschriften Almwirtschaft und Milchverarbeitung

Ausbildung
- Berufsschulen und sonstige Ausbildungsstätten

Gewerbe
- Käsereien
- Molkerein

Almen, Kultur
- Vereine
- Museen

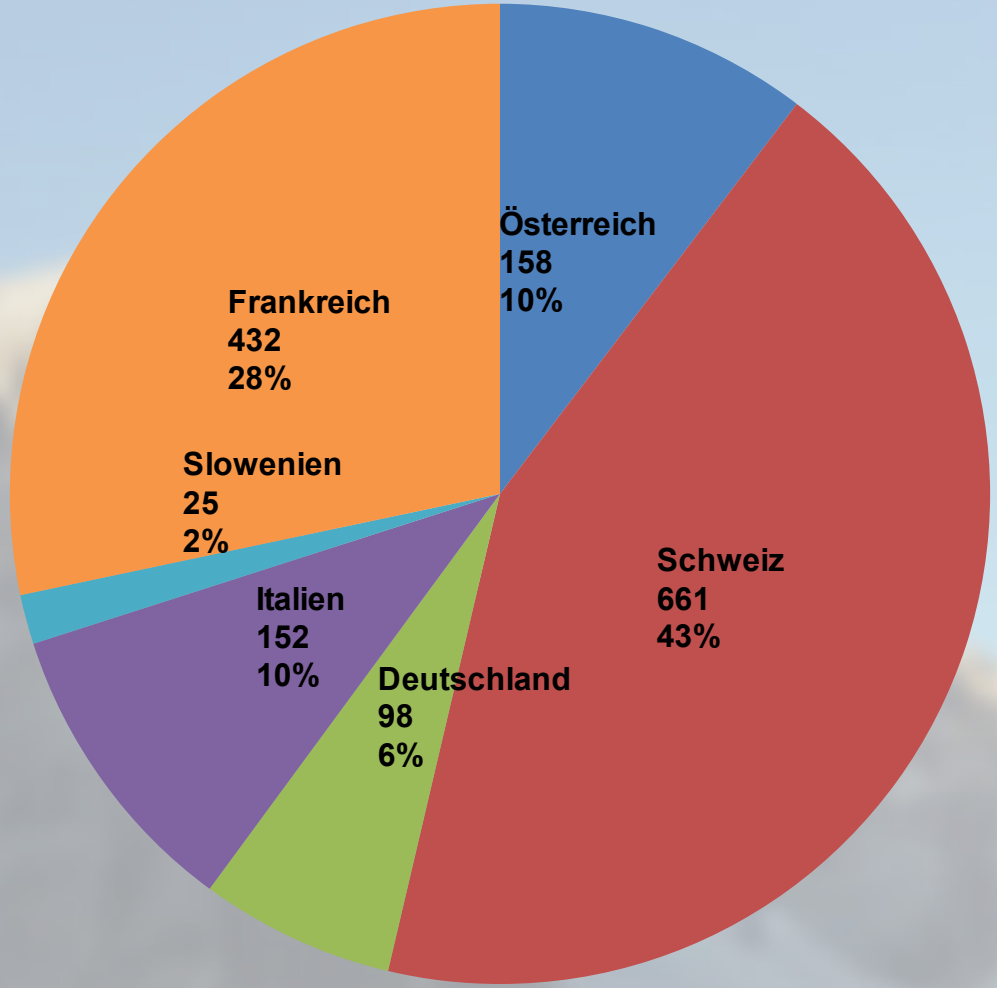
Portale
- Fachportale der Milchbranche (z.B. z‘alp.ch)

Medien
- Zeitungen
- Fernsehen
- Radio
- Verteiler

Private Kontakte

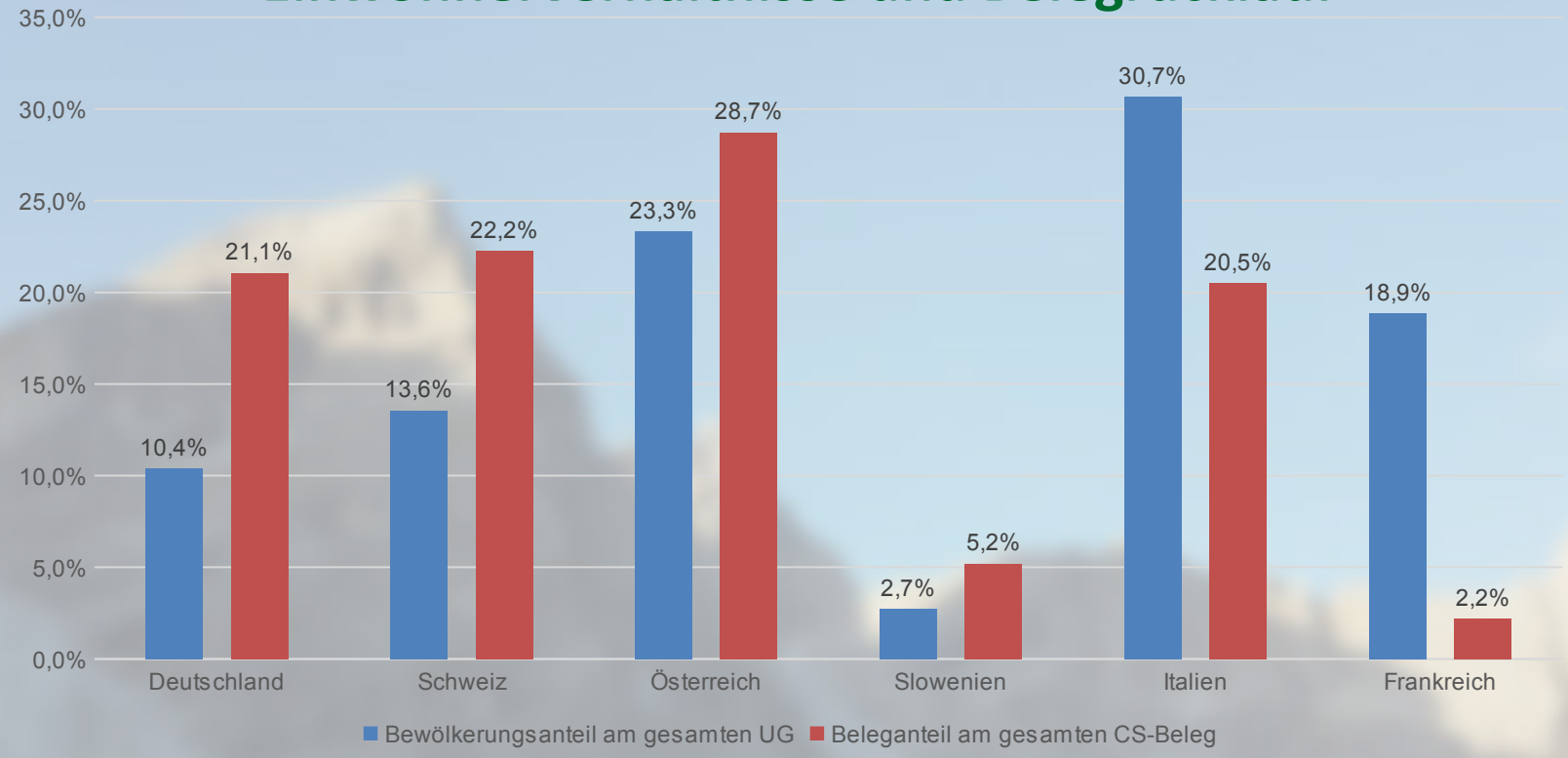
Datenerhebung

Crowdsourcing Kontakte (Februar – Mai 2017)

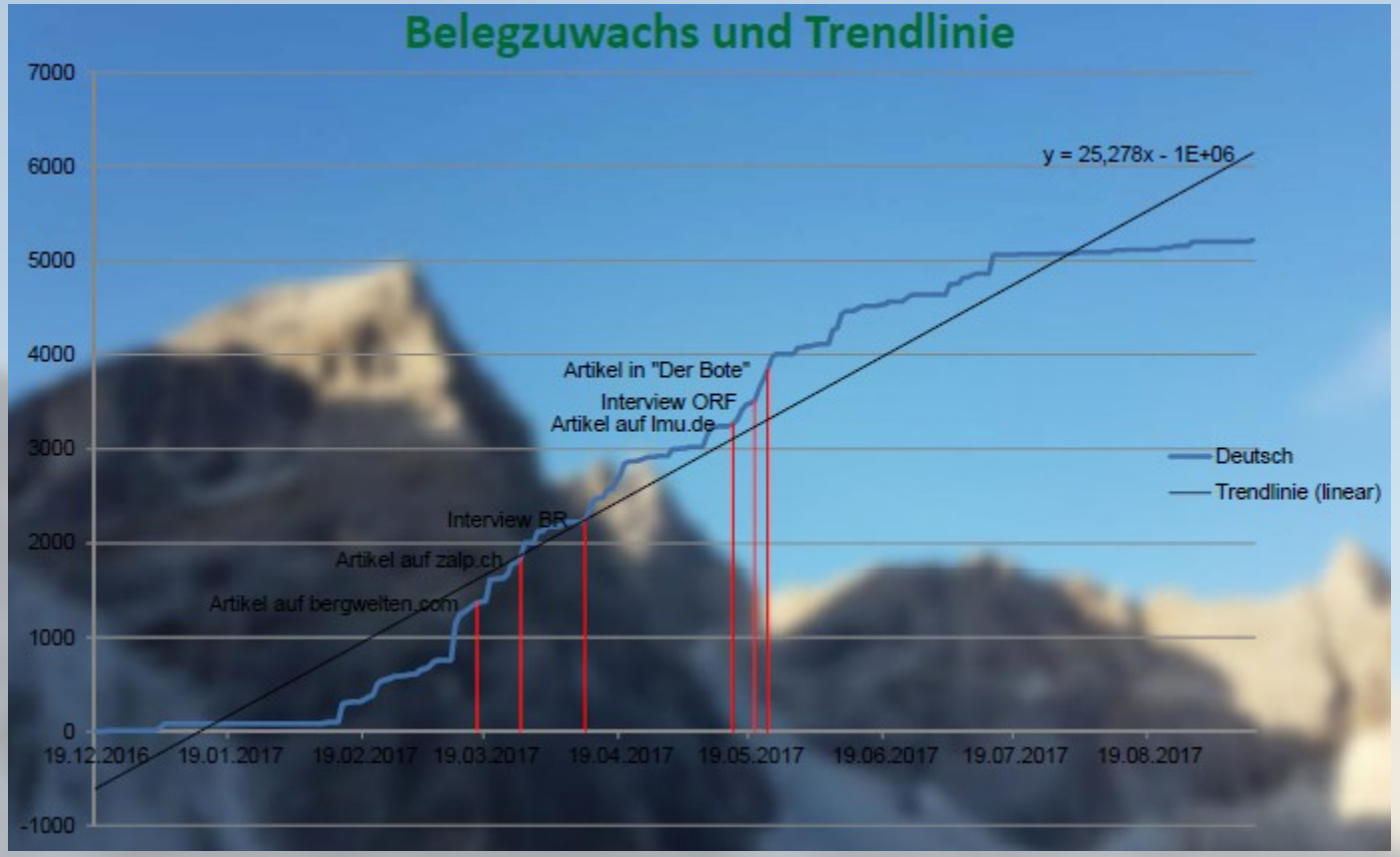


Datenerhebung

Crowdsourcing Mailingaktion (Februar – Mai 2017)
Einwohnerverhältnisse und Belegrücklauf



Crowdsourcing



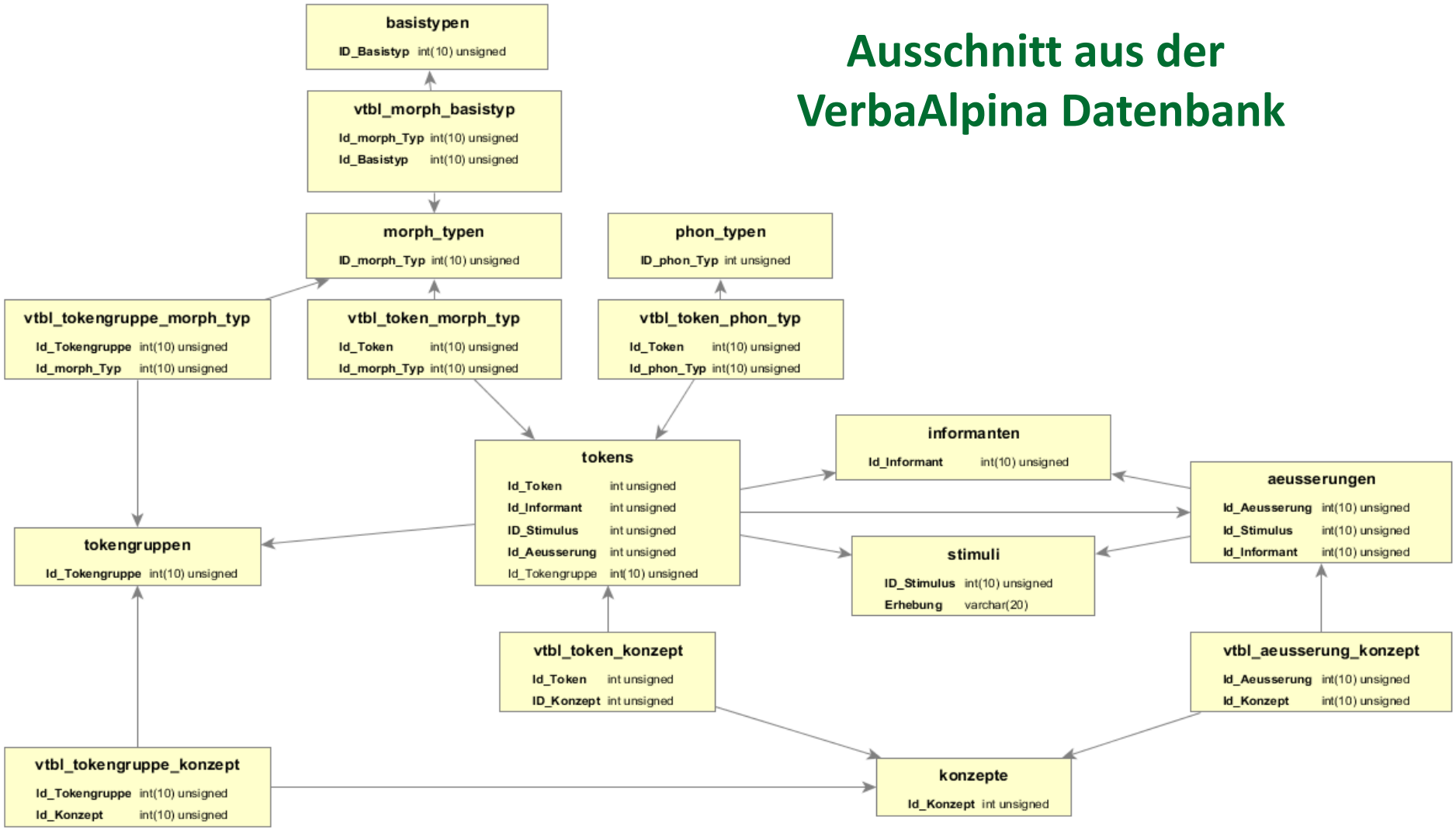
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Eingliederung in den Datenbestand

Markus Kunzmann | Florian Zacherl

Eingliederung

Ausschnitt aus der VerbaAlpina Datenbank



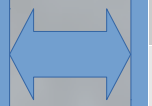
Die VerbaAlpina Datenbank

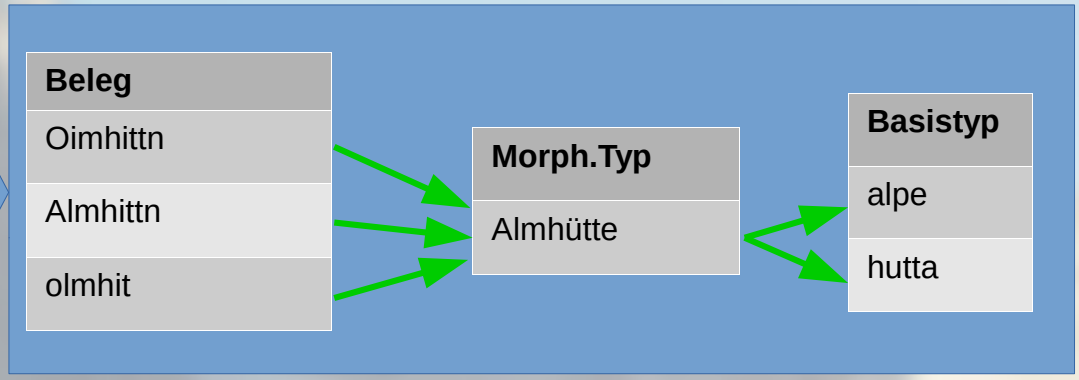
Normalisierte Datenbank-Struktur

Vorteil

- Größtmögliche Vermeidung von inkonsistenten Daten

Beleg	Morph. Typ	Basistyp
Oimhittn	Almhütte	alpe, hutta
Almhittn	Almhütte	alpe, hutta
olmhit	Almhütte	alpe, hutta





Die VerbaAlpina Datenbank

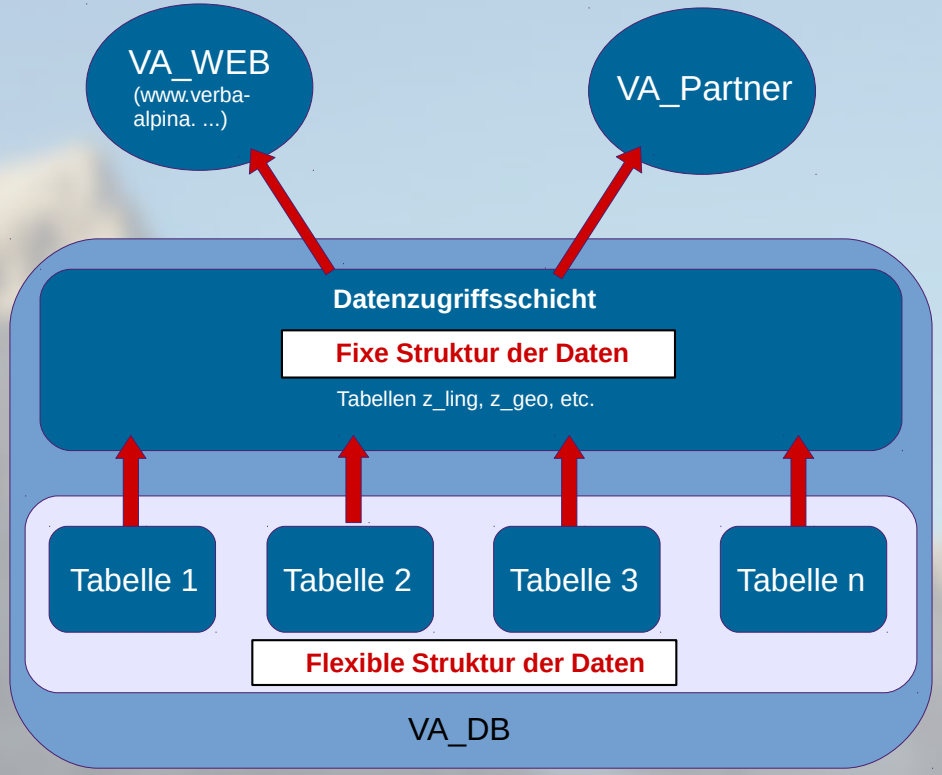
Normalisierte Datenbank-Struktur

Nachteile

- Komplexe SQL-Anfragen
- Langsame Datenbankabfragen

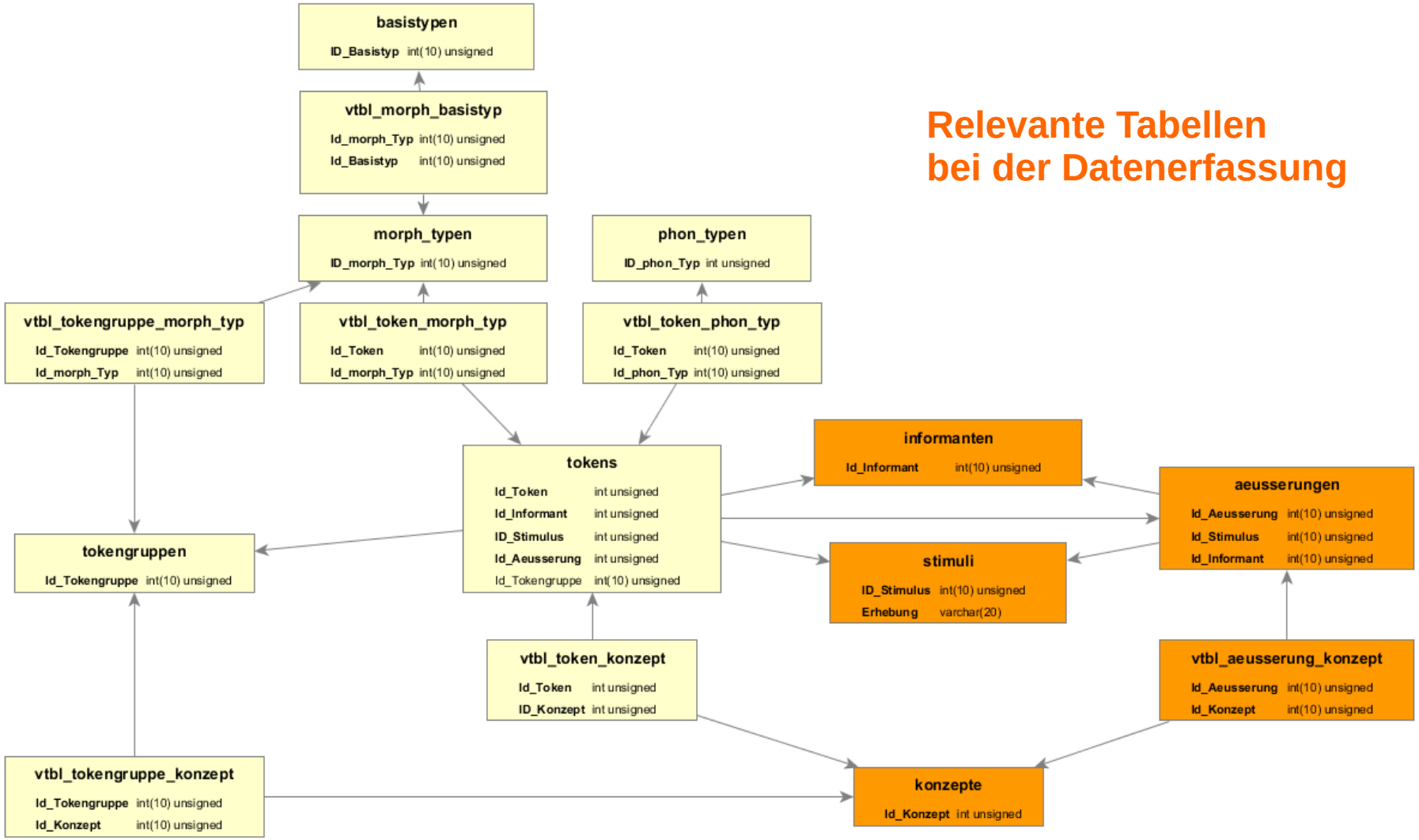
=> Datenzugriffsschicht

Datenzugriffsschicht



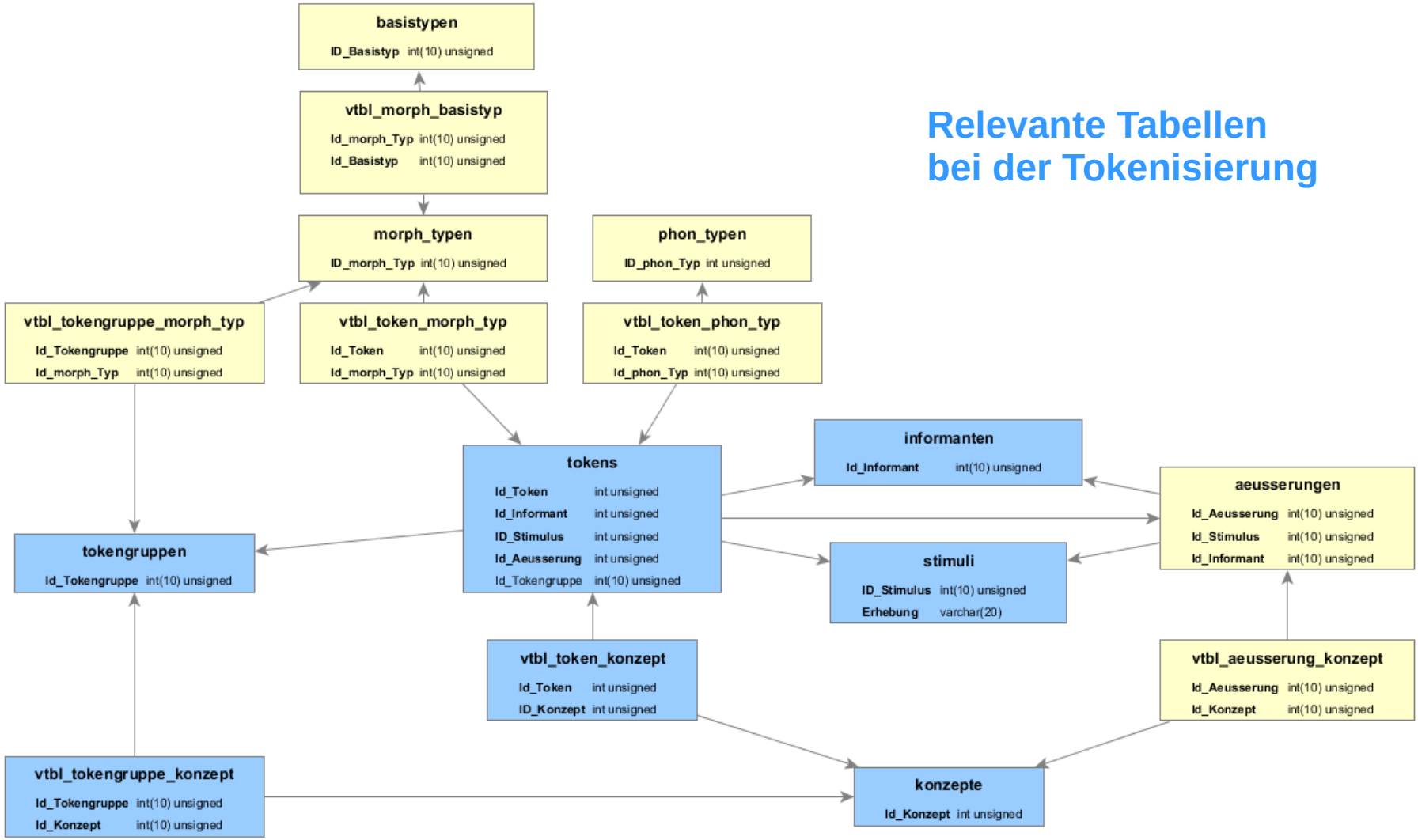
Eingliederung

Relevante Tabellen bei der Datenerfassung



Eingliederung

Relevante Tabellen bei der Tokenisierung



Eingliederung

Tokenisierung

Aeusserungen

Informant	Stimulus	Aeusserung	Konzept
anonymousCrowder_58	CROWD_1_1	Milch ohseichn	MILCH SEIHEN

Tokens

Informant	Stimulus	Token	Konzept	Morph. Typ
anonymousCrowder_58	CROWD_1_1	Milch	MILCH	Milch
anonymousCrowder_58	CROWD_1_1	ohseichn	ABSEIHEN	Abseihen

Tokengruppen

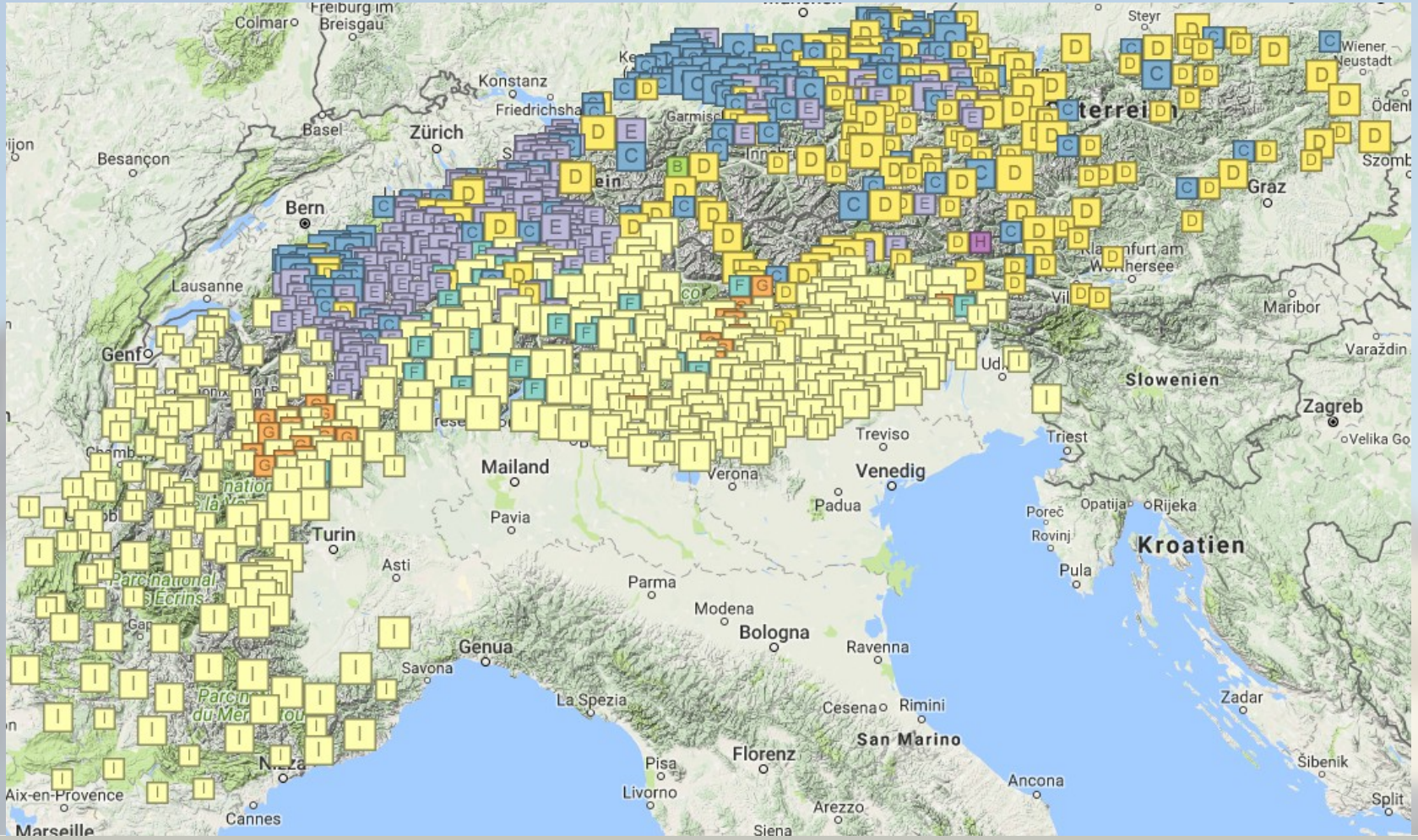
Informant	Stimulus	Tokengruppe	Konzept	Morph. Typ
anonymousCrowder_58	CROWD_1_1	Milch ohseichn	MILCH SEIHEN	Milch abseihen

Visualisierung

Markus Kunzmann | Florian Zacherl

Visualisierung

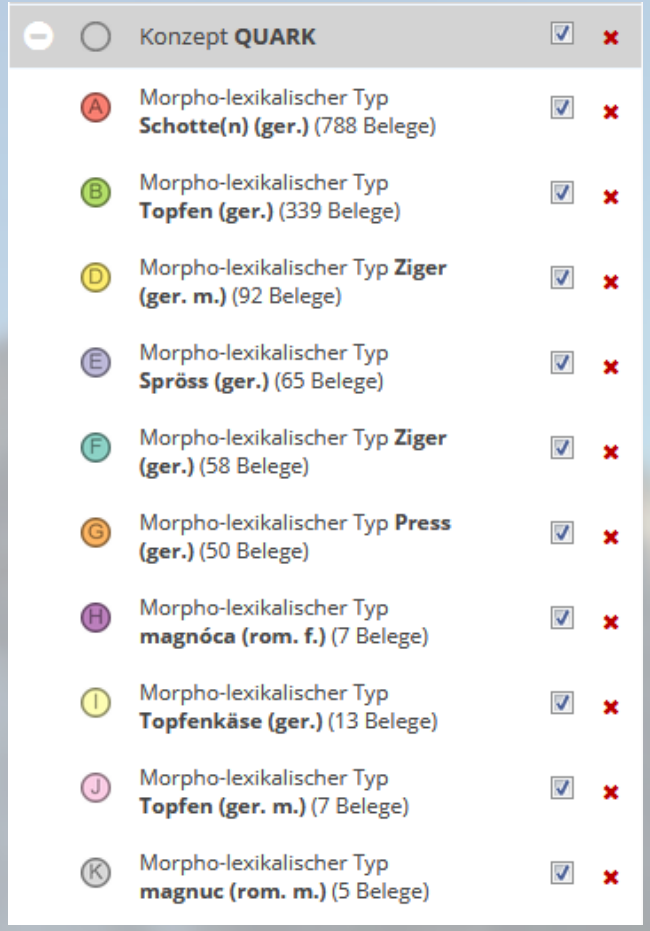
Qualitative Karte

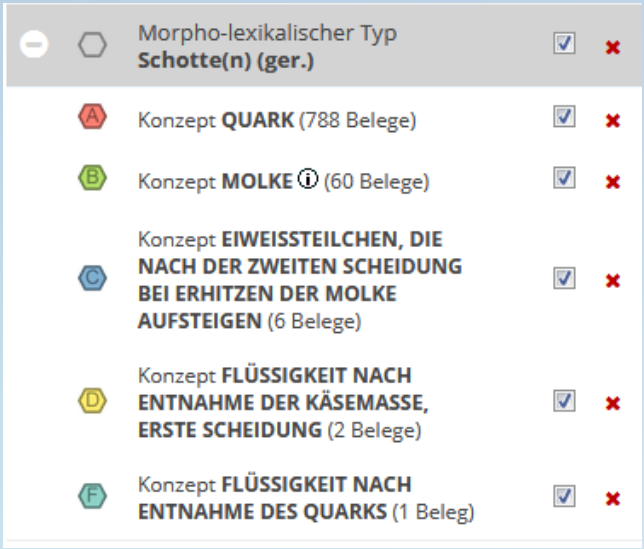


Markus Kunzmann | Florian Zacherl

Visualisierung

Onomasiologisch vs. Semasiologisch





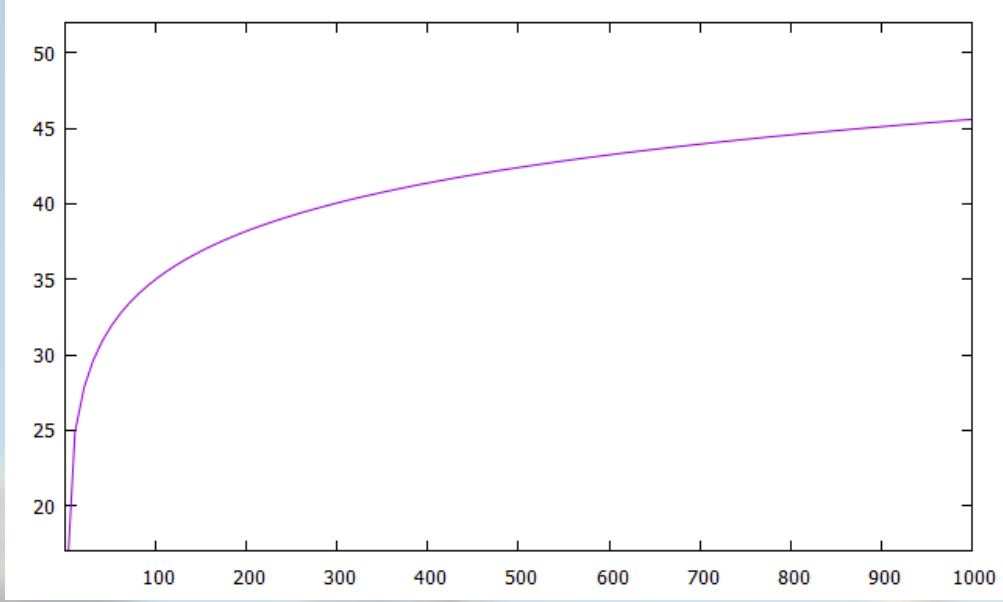
Visualisierung

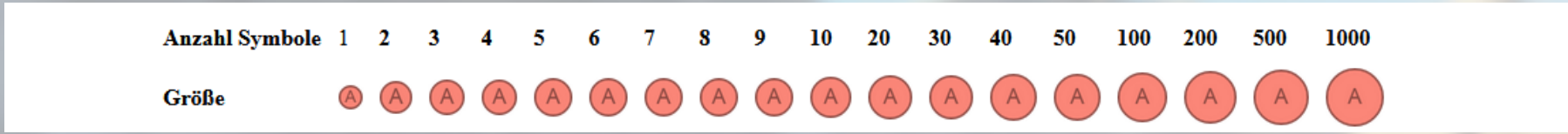
Symbole

Prinzipiell: Ein Symbol pro Beleg

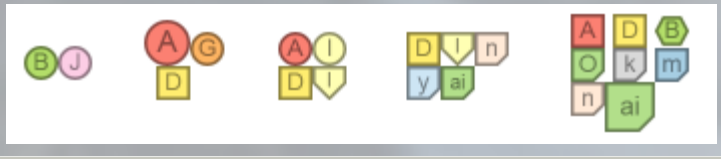
Referenzgröße: Gemeinde
=> gleichartige Symbole innerhalb einer Gemeinde werden zusammengefasst

Logarithmisches Wachstum bei identischen Symbolen



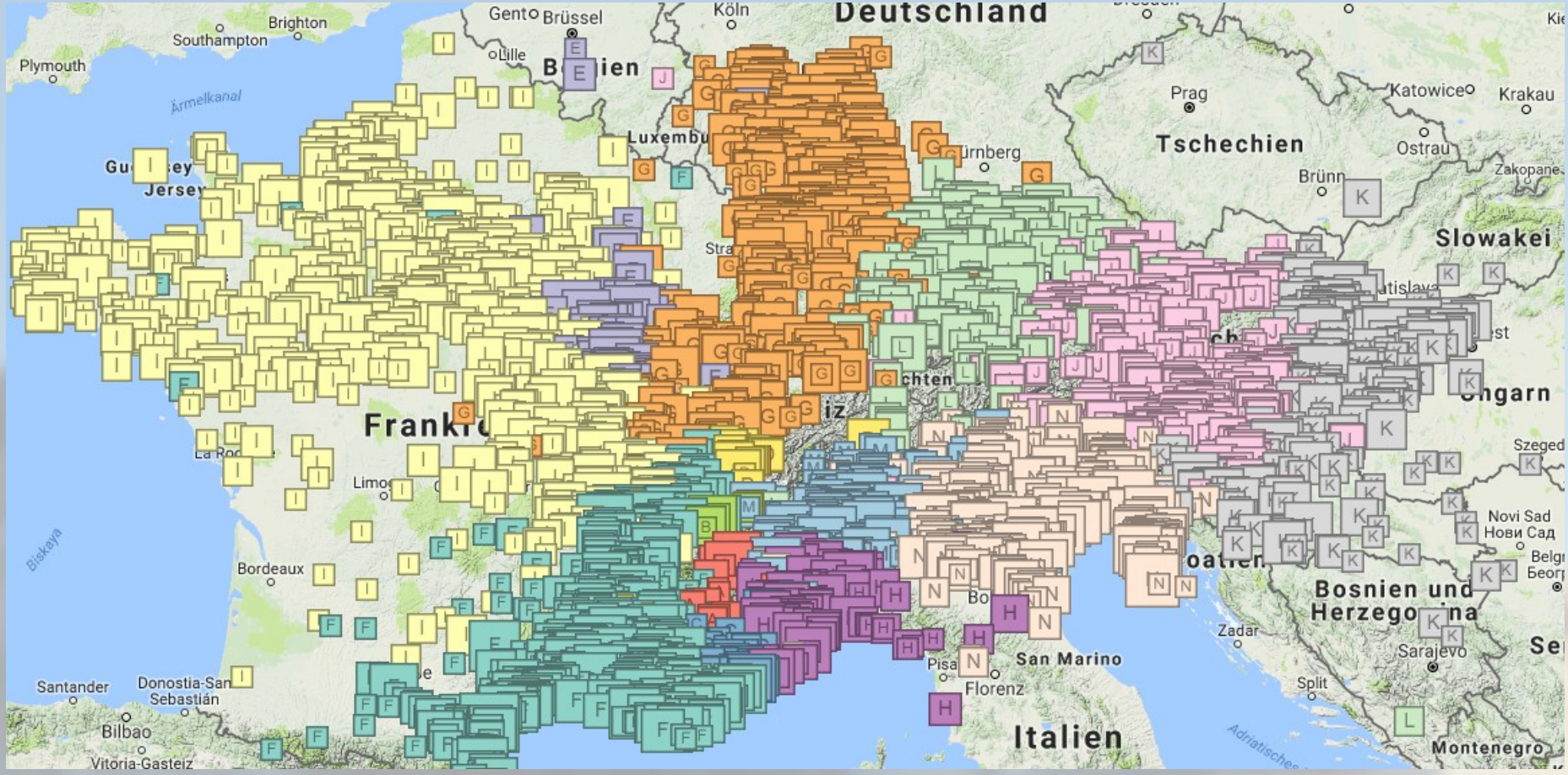


Kombinierte Symbole



Visualisierung

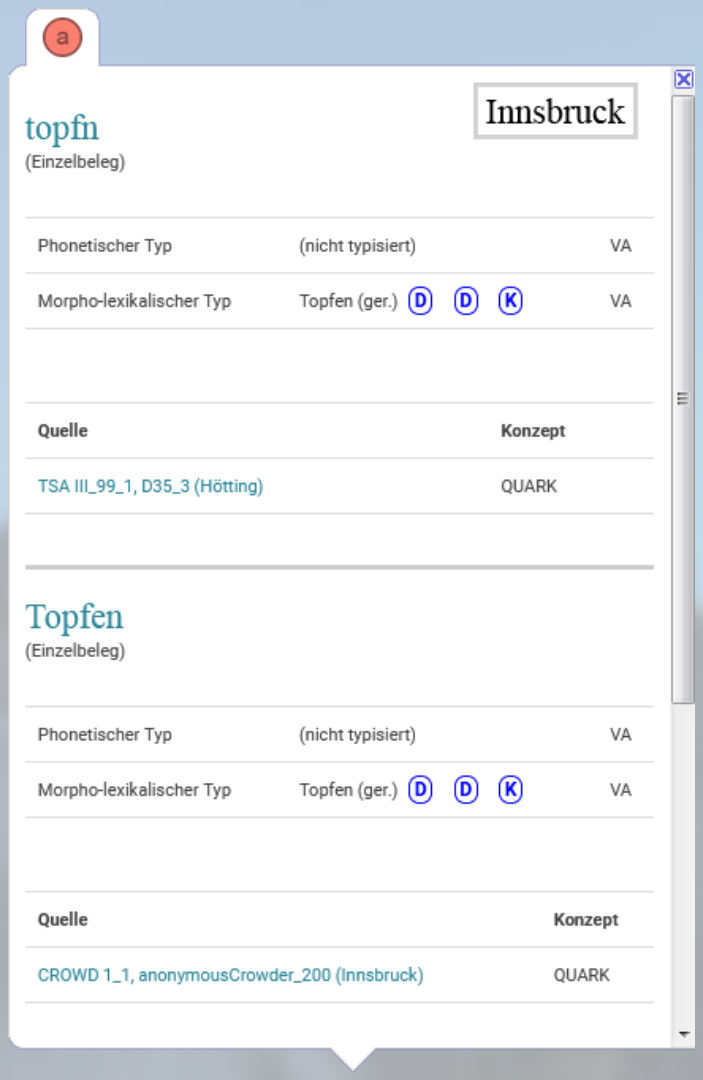
Große Anzahl an Belegen möglich

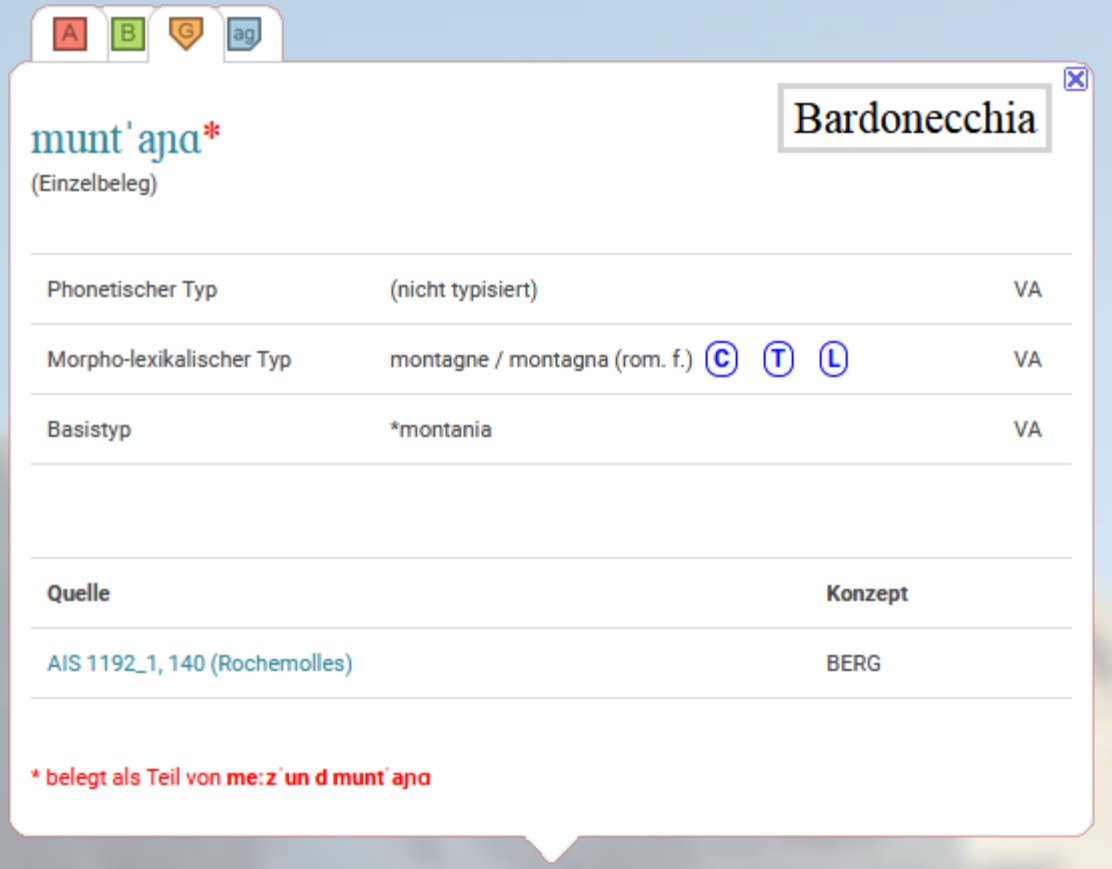


Karte mit 92890 Belegen

Visualisierung

Belegfenster

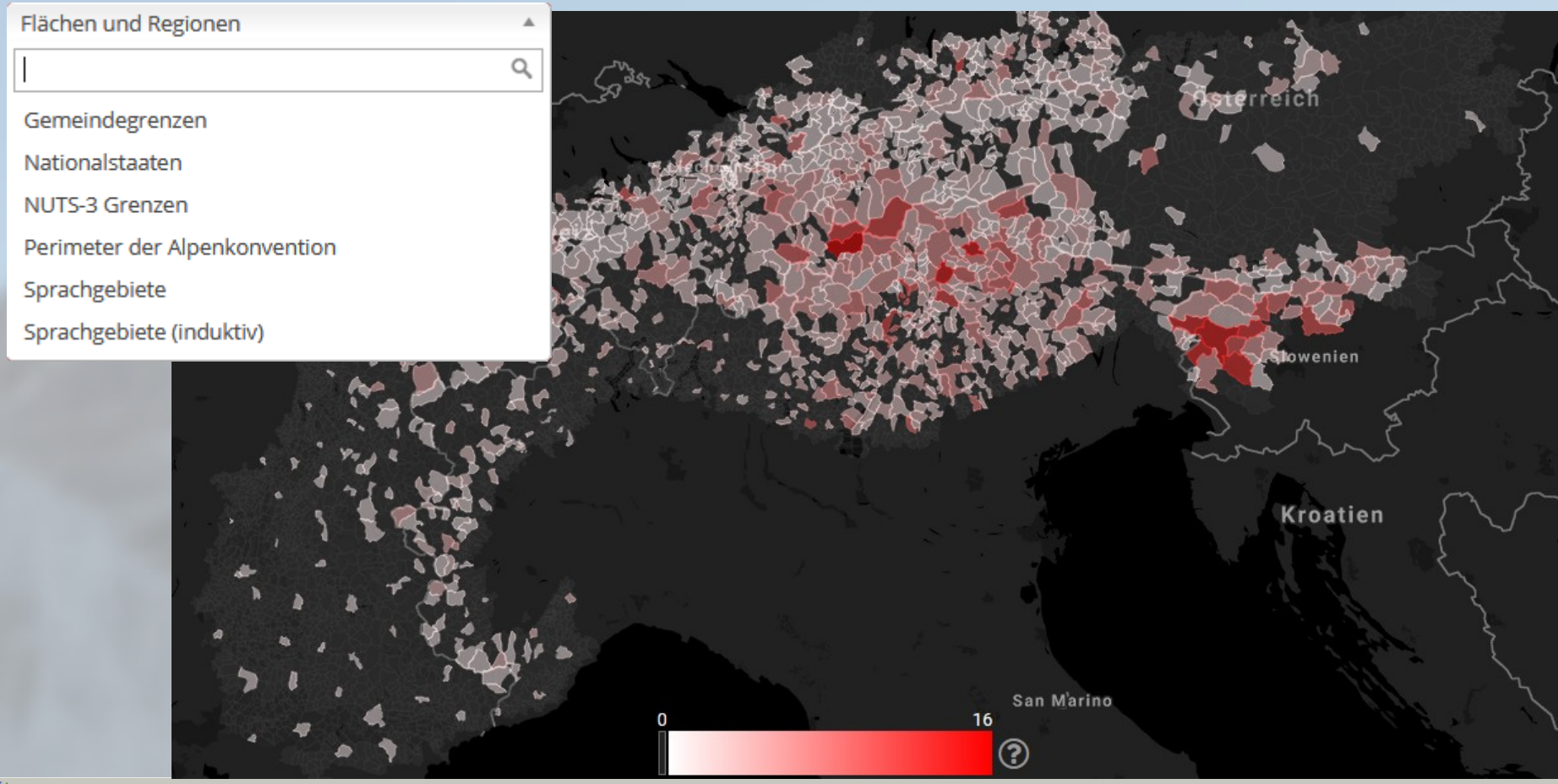




Markus Kunzmann | Florian Zacherl

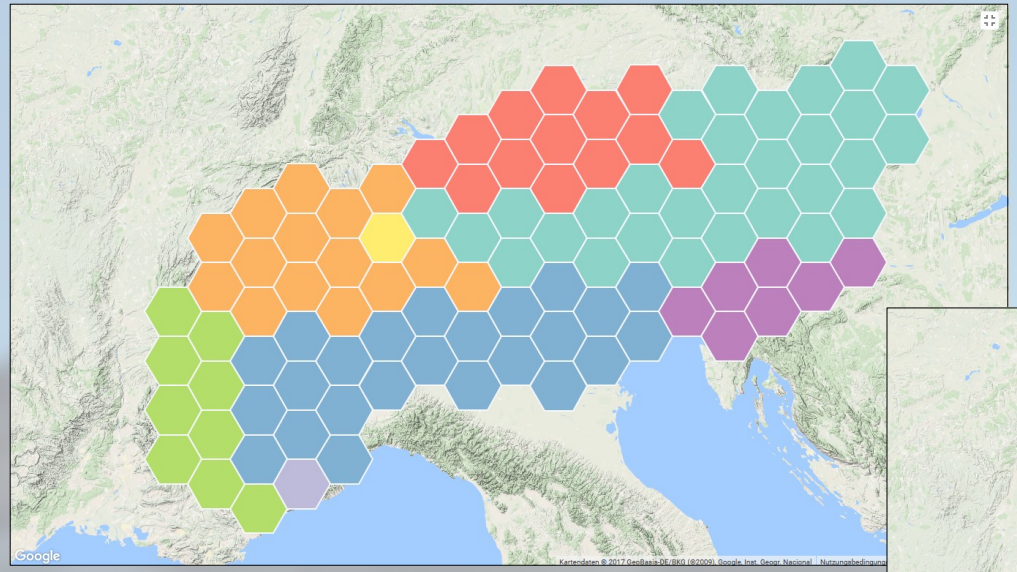
Quantitative Darstellung

Abbildung von Punktdaten auf Flächen



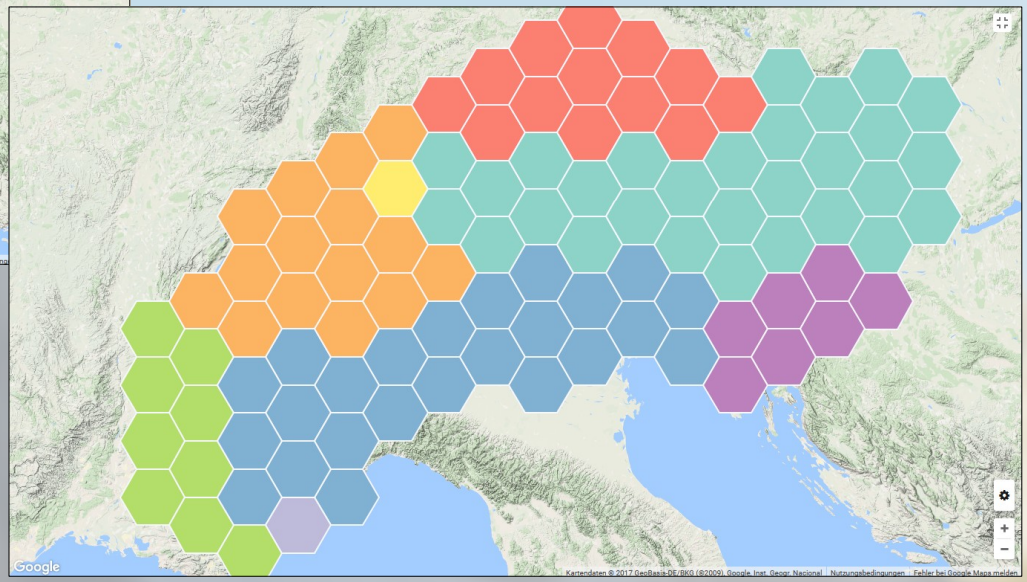
Visualisierung

Abstrahierte Karte



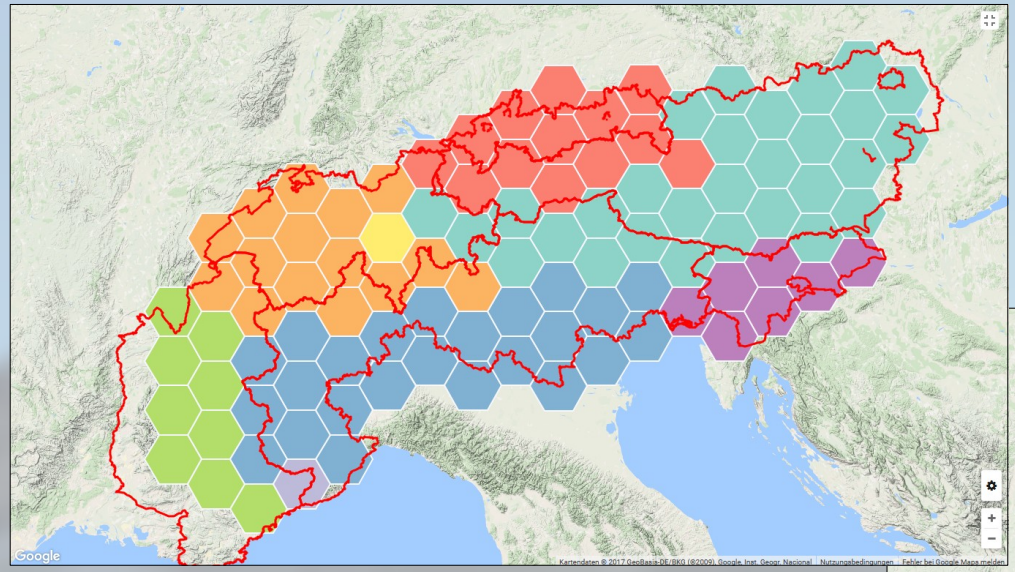
NUTS-3 Grenzen Geographisch

NUTS-3 Grenzen Schematisch



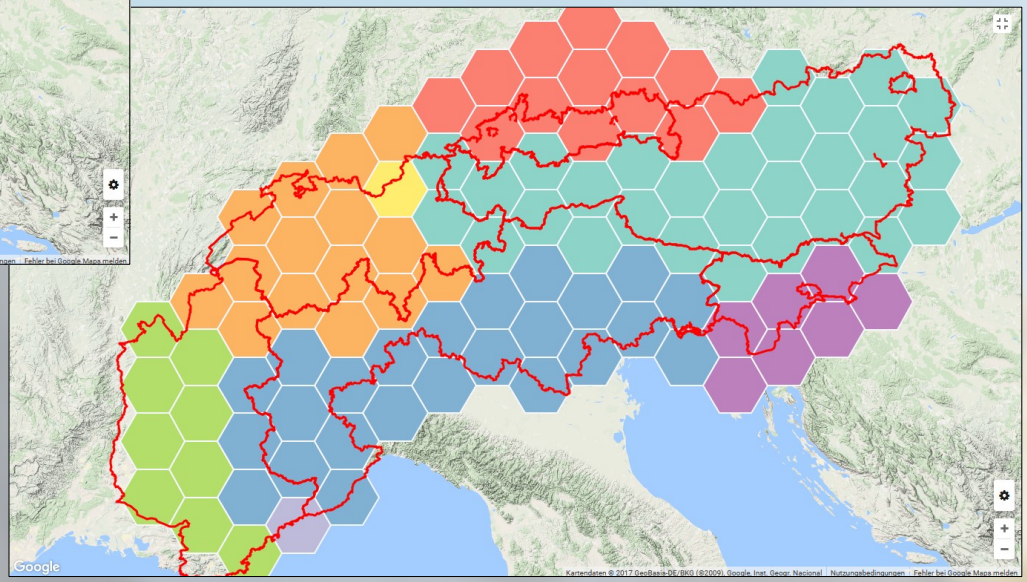
Visualisierung

Abstrahierte Karte



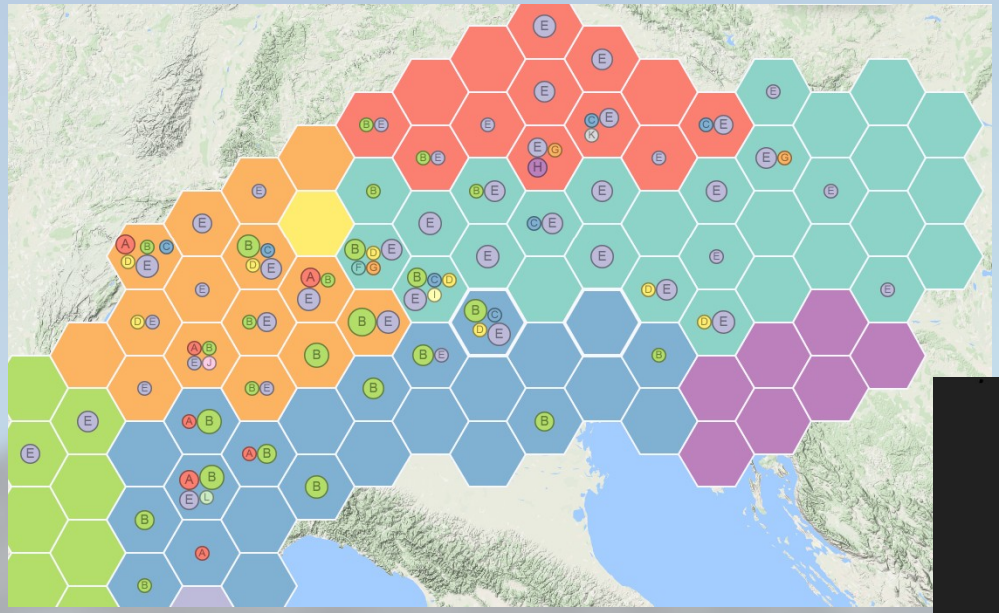
NUTS-3 Grenzen Geographisch

NUTS-3 Grenzen Schematisch



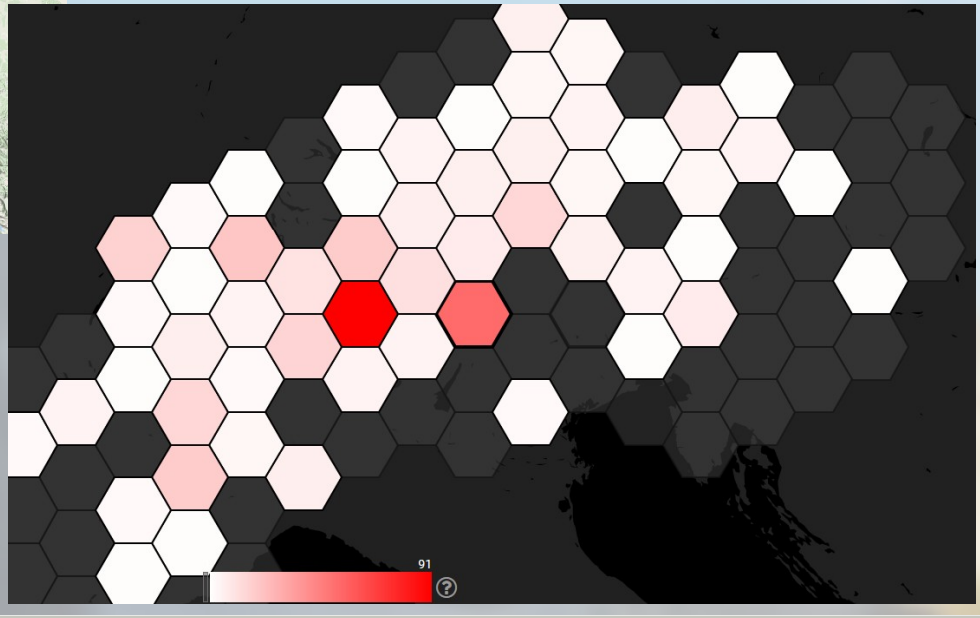
Visualisierung

Abstrahierte Karte



Qualitativ

Quantitativ

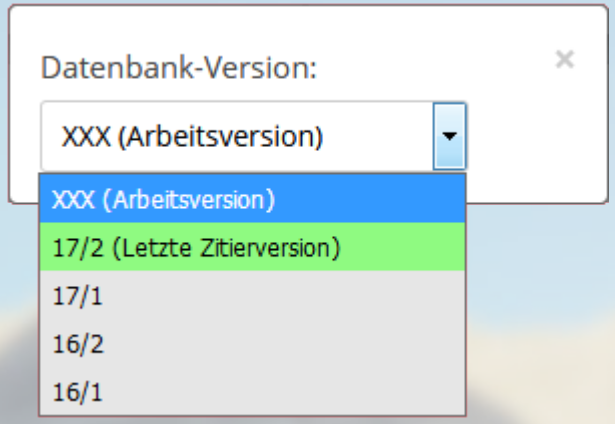


Nachhaltigkeit

Markus Kunzmann | Florian Zacherl

Versionierung

Vollständiger Datenbestand wird regelmäßig versioniert (ca. halbjährlich)
Grund: Zitierbarkeit
Technische Umsetzung: Kopie der vollständigen Datenbank



Vorteile

- Schneller Zugriff
- Geringe Fehleranfälligkeit

Nachteile

- Höherer Speicherbedarf
- Nachträgliche Strukturanpassungen

Versionierung

Außerdem Bereitstellung des vollständigen Programmcodes unter

https://github.com/VerbaAlpina

Plugin Interactive Map | **Plugin VerbaAlpina** | **Plugin CS-Tool** | **Theme VerbaAlpina**

Langfristige Nutzung

Probleme

Nicht rückwärtskompatible Software

Lösungsansatz: Zeitlich unbegrenzter Zugriff über einen Docker auf dem Server der Universitätsbibliothek: Eingefrorene Konfiguration ab Förderungsende (Prototyp Ende 2016)

Abhängigkeit von externen Diensten für die Kartierung

Lösungsansatz: Eigenes Hosten des Kartenmaterials über das Leibniz-Rechenzentrum (aktuell in Arbeit)

Archivierung der Datenbasis

Archivierung bei verschiedenen Institutionen

- IT-Gruppe Geistewissenschaften
- Clarin-D
- Evtl. Generic Research Data Infrastructure (GeRDI)