universität innsbruck

Merging standardized and non-standardized gazetteers

Claudia Posch, Elisabeth Gruber, Gerald Hiebel, Eva Zangerle, Gerhard Rampl
University of Innsbruck / Department for Languages and Literature: Linguistics

Content

What and why?

- Named entity recognition (NER) in texts of the Austrian Alpine Journal
- Named entity linking (NEL)

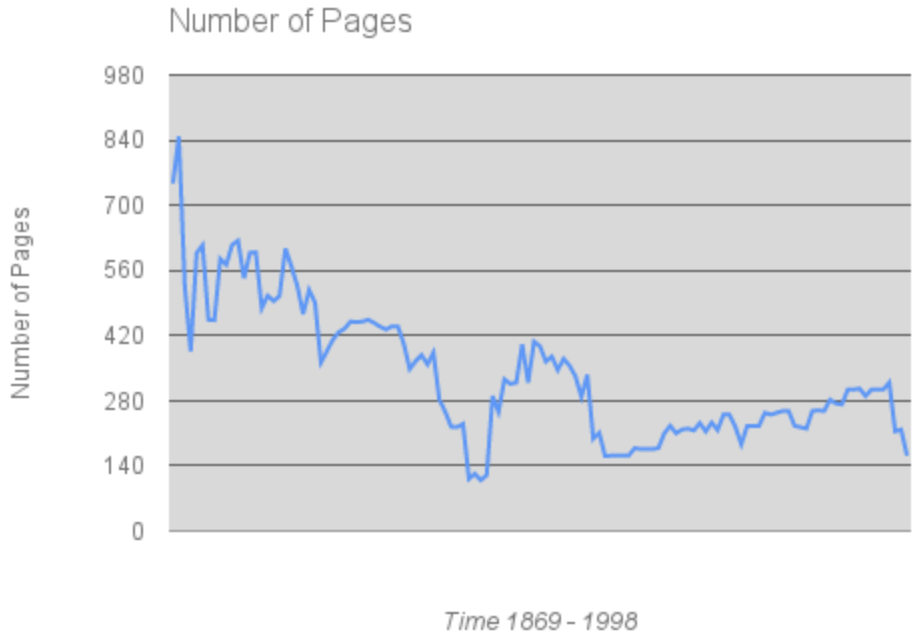
How?

- Problems
- Solution strategy

What? Corpus 'Alpenwort'

- corpus of the "Zeitschrift des Österreichischen Alpenvereins"
- 1869-1998; 126 years, 122 volumes)
- total: 43.383 pages
- ca. 18 mil. tokens

- 1915 – 1961 gothic script



Why?

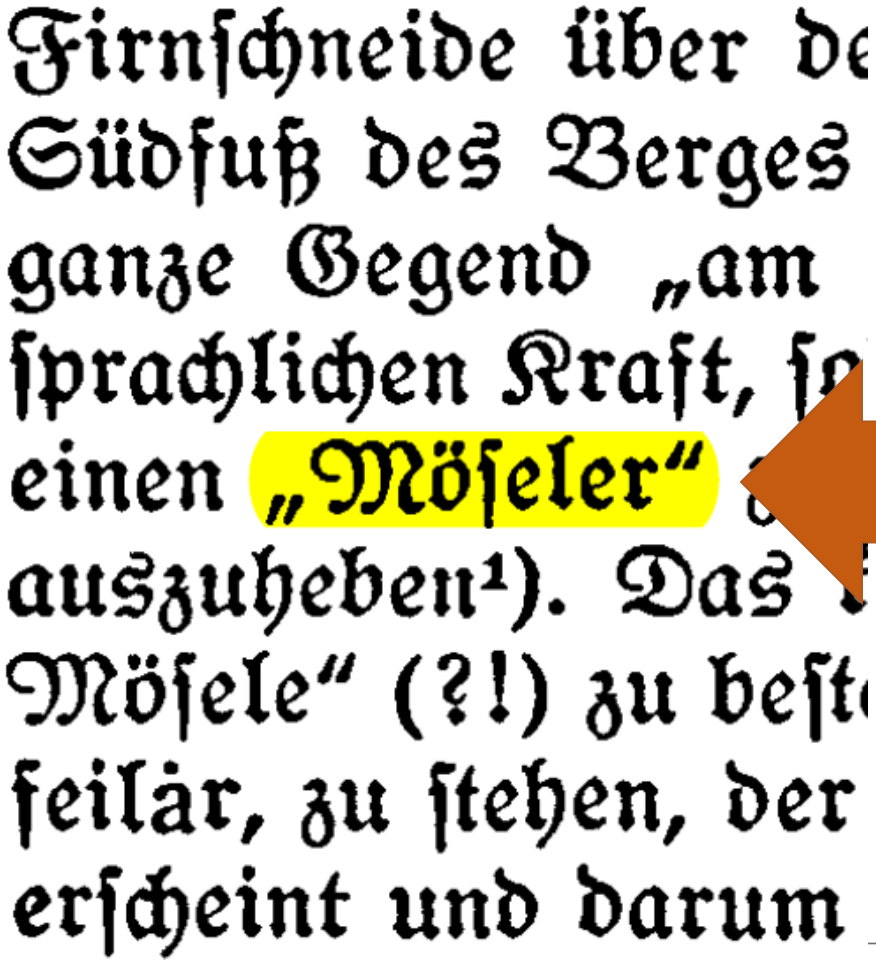
- The Alpine Club played an important role in the early exploration of the Alps
- Lots of alpine names (esp. mountain names)
- Lots of first records (first ascents etc.)

- Linking texts to names-database would be interesting (scientifically and for public)
- But: texts are unstructured data

NER

Firnſchneide über dem Wareckkees den unbegreiflichen
Südfuß des Berges findet ſich eine kleine ſumpfige St
ganze Gegend „am Möſele“ benannten! Immerhin
ſprachlichen Kraft, ſolch einen weither geholten Namen
einen „Möſeler“ zu machen, den Berg als „Perſönlich
auszuheben[1]). Das klingt doch noch beſſer, als wenn d
Möſele“ (?!) zu beſteigen. — Ganz ähnlich ſcheint es un
feilär, zu ſtehen, der bei Anich (1774), Staffler und So
erſcheint und darum kaum eine urſprüngliche Gipfelben

NER: probabilistic and text matching

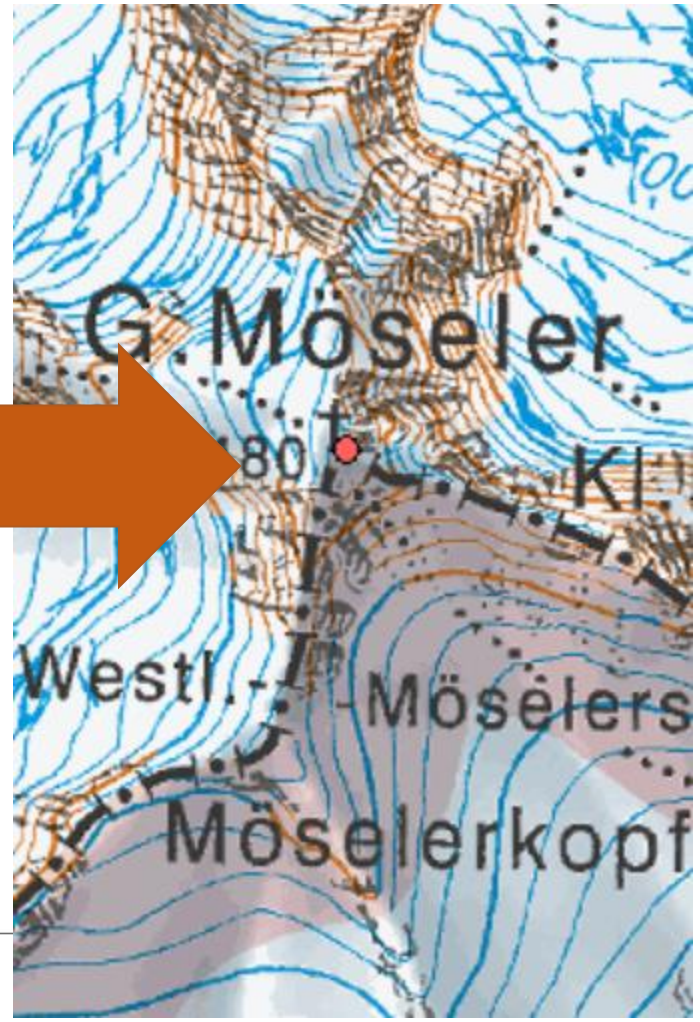

Firnſchneide über de
Südfuß des Berges
ganze Gegend „am
ſprachlichen Kraft, ſo
einen „Möſeler“
auszuheben[1]). Das
Möſele“ (?!) zu beſt
feilär, zu ſtehen, der
erſcheint und darum

name_id bigint	name text
	Möseler
344373	Großer Möseler
550290	Kleiner Möseler

NEL: linking NE to real world entities

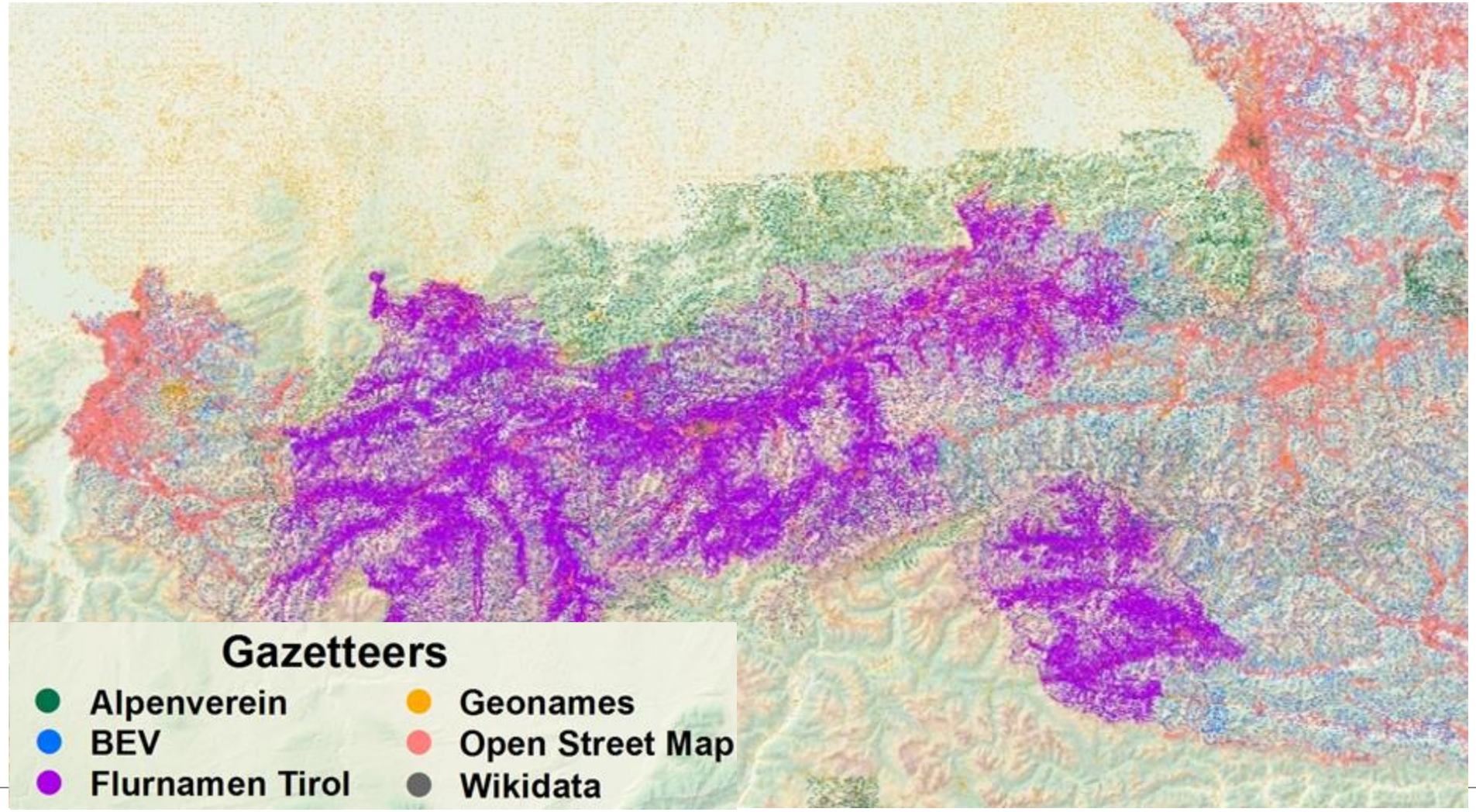
Firnſchneide über de
Südfuß des Berges
ganze Gegend „am
ſprachlichen Kraft, ſo
einen „Möſeler“
auszuheben[1]). Das
Möſele“ (?!) zu beſt
feilär, zu ſtehen, der
erſcheint und darum

	name text
	Möseler
3	Großer Möseler
)	Kleiner Möseler



universität innsbruck

Gazetteers: need for border crossing merging



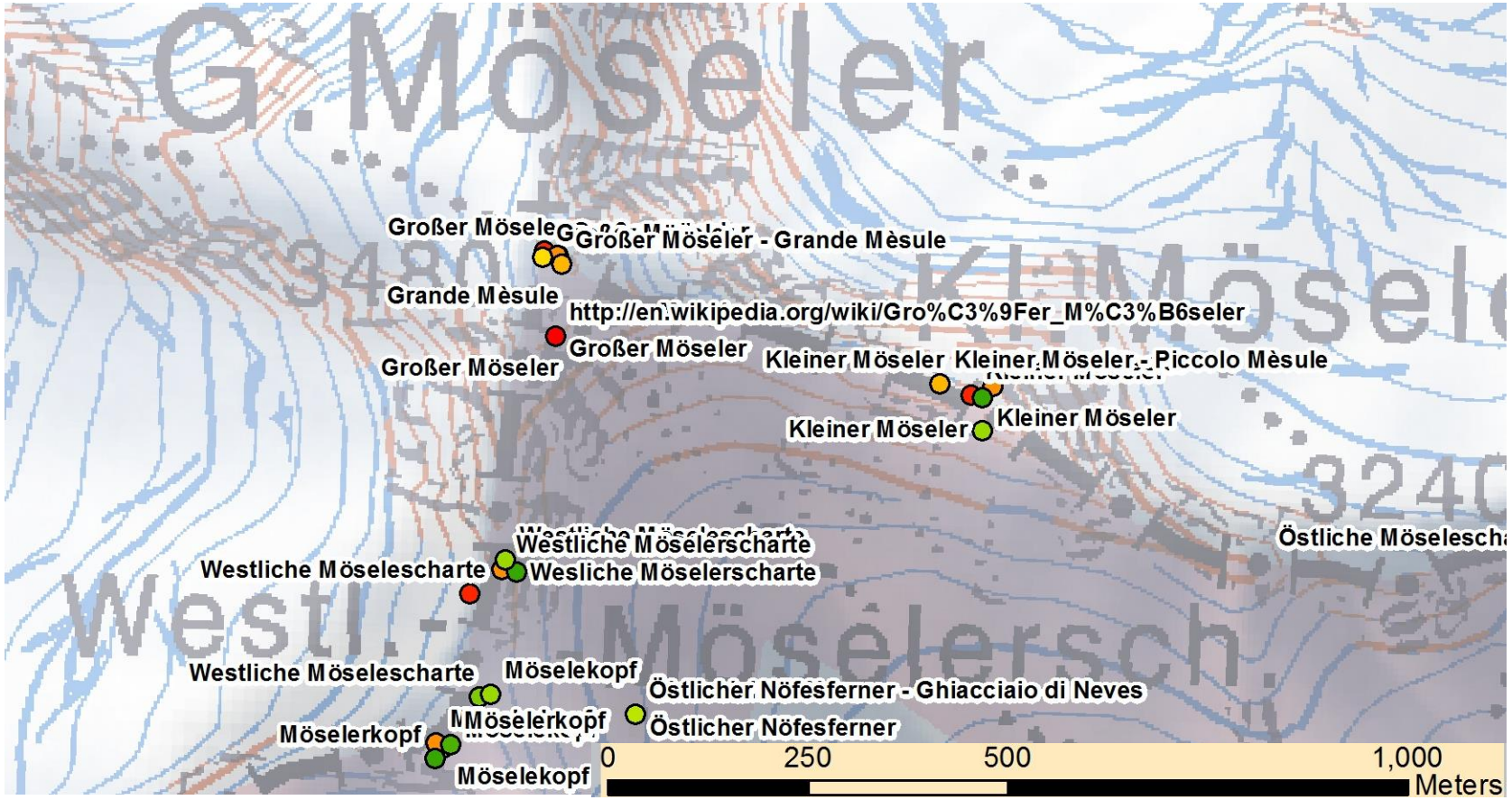
universität innsbruck

Problem 1: Text matching

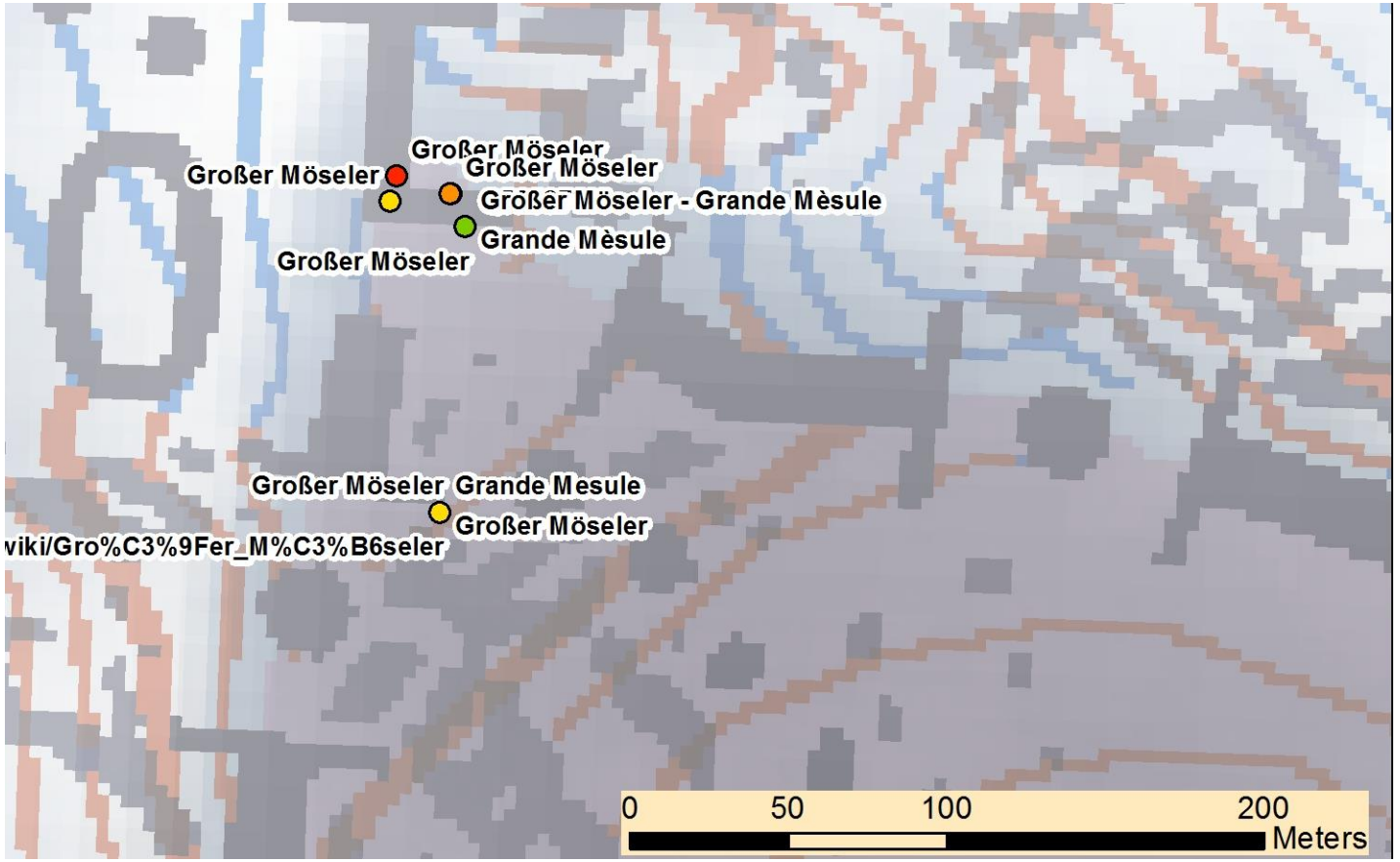
- Mairspitz : Mairspitz**e** : Ma**y**rspitz**e**
- Dreiherrnspitz : Dreiherrnspitz**e** : Dreiherr**e**nspitz**e**

- Dreiecker : Cima di Campo
- Großer Möseler : Gran M**è**sule : Gran M**e**sule : Großer Möseler **-** Gran Mesule…

Problem 2: location based matching

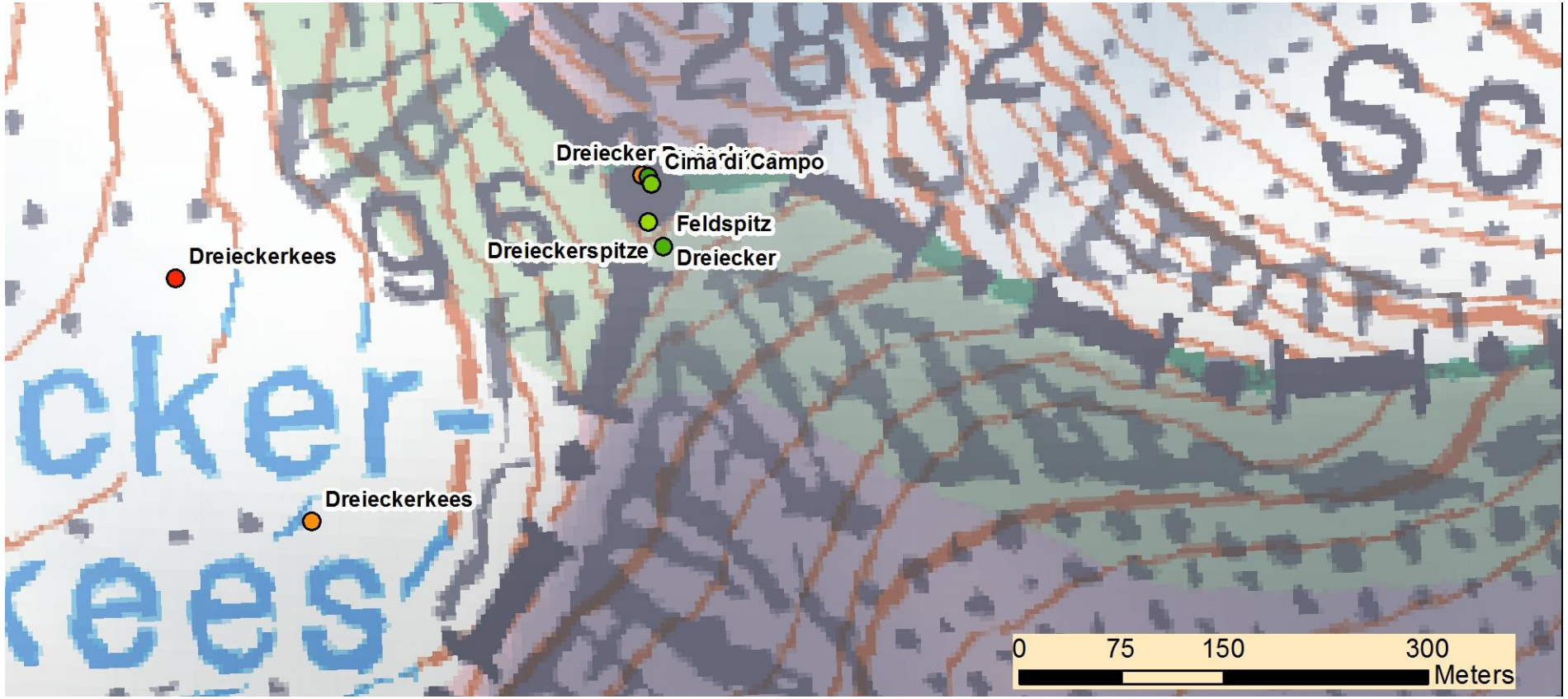


Problem 2: location based matching



Problem 2: location based matching

Problem 2: location based matching



Problem 3: feature type based matching

- Homonyms can often be separated by feature type
- Brandach: settlement name and field name

- OpenStreetMapp: 1749 types and subtypes
- Wikidata: 1076 types and subtypes
- geonames: 680 types and subtypes
- Austrian Map: 41 typesand subtypes
- Alpine Club maps: 18 types and subtypes

What we do

- Text identity matching (Levenshtein distance 0)
- Location based matching (buffer 200m)
- Type matching (11 supertypes, 11 subtypes)

- Text identity matching (Levenshtein distance 1 and 2) -> like above (buffer 50-100m)

- Location based matching (buffer max 50m) -> type matching

Questions

- Which methods are used by others?
- To which extent is merging done automatically? What/How much is done manually?
- Are there other components involved in the matching (besides name, type and location)?
- Which products/services do exist already we might not be aware of?

Thank you!

Feature supertypes

- topographic_feature
- undersea
- vegetation
- hydro_feature
- admin_area
- settlement
- buildings
- activity
- area
- path
- not_specified

- topographic_feature
  - mountain_peak
  - valley
  - mountain_range
  - natural_saddle
- vegetation
  - alpine_pasture
  - wood_area
- hydro_feature
  - stream
  - lake
  - glacier
- admin_area
  - country
- area
  - named_micro_area