A Proposal for a Two-way Journey on Validating Locations in Unstructured and Structured Data

Ilkcan Keles [1], **Omar Qawasmeh** [2], Tabea Tietz [3],
Ludovica Marinucci [4], Roberto Reda [5], and Marieke van Erp [6]

[1] Aalborg University- Dept. of Computer Science, [2] Université de Lyon, CNRS, Lab. Hubert Curien,
[3] FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, [4] Semantic Technology Laboratory (STLab), CNR,
[5] University of Bologna- Dept. of Computer Science, [6] KNAW Humanities Cluster, DHLab

LDK, May 2019

Introduction

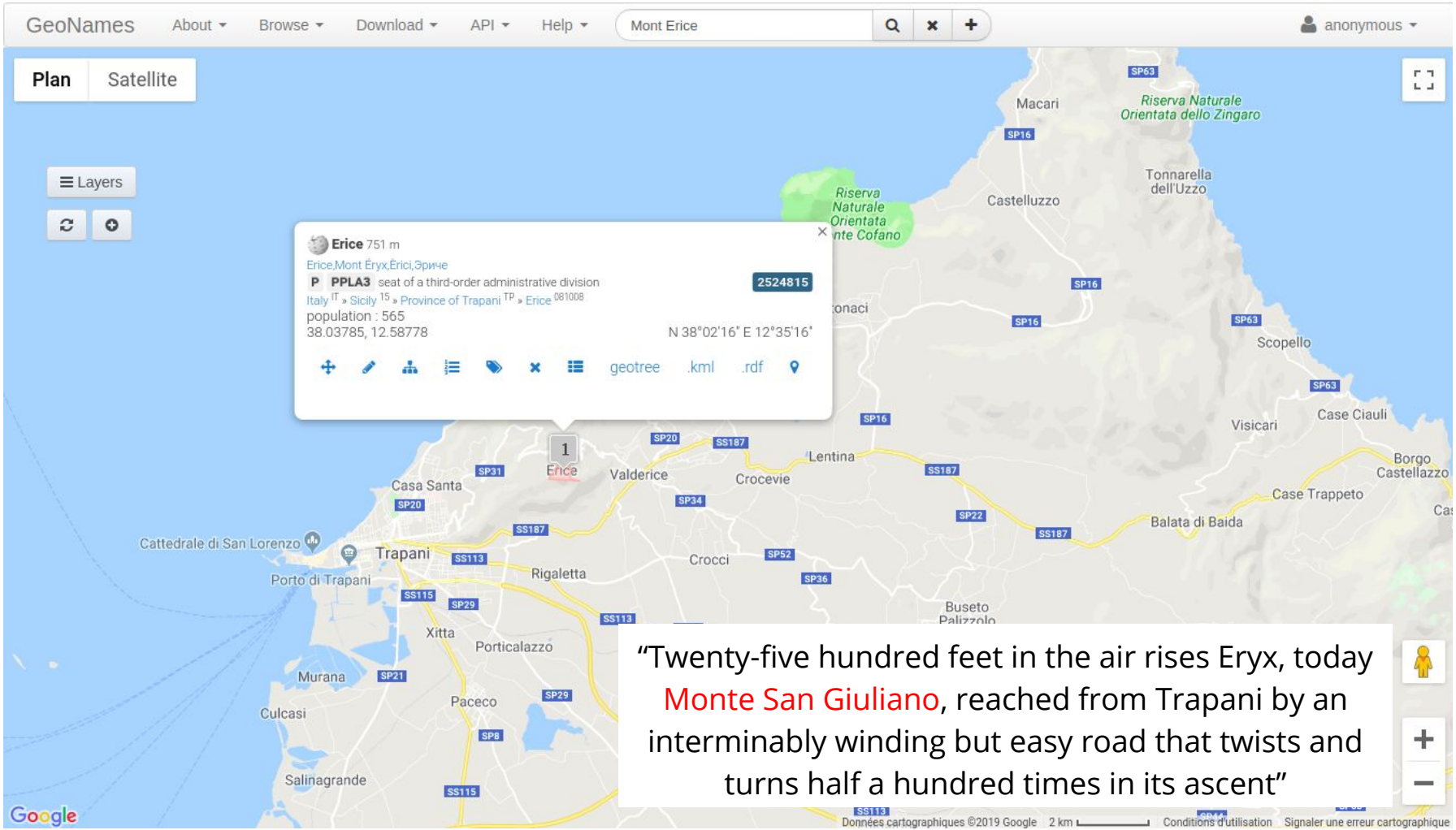
- NLP includes a variety of techniques for the automatic analysis and representation of human language
  - E.g. extract structured datasets from unstructured textual documents
    - Can be used to enrich, compare and/or match with existing Linked Data sets

- Problems?
  - NLP systems are not without errors, and neither is Linked Data
  - There is a need to ensure that information contained in structured datasets is valid

Textual vs Linked data validity

- Textual Data validity
  - The validity of information contained in texts, where someone is not sure about correct or up-to-date information (e.g. travel diaries)

- Linked Data validity
  - The validity of information contained in structured datasets (e.g. DBpedia or GeoNames)



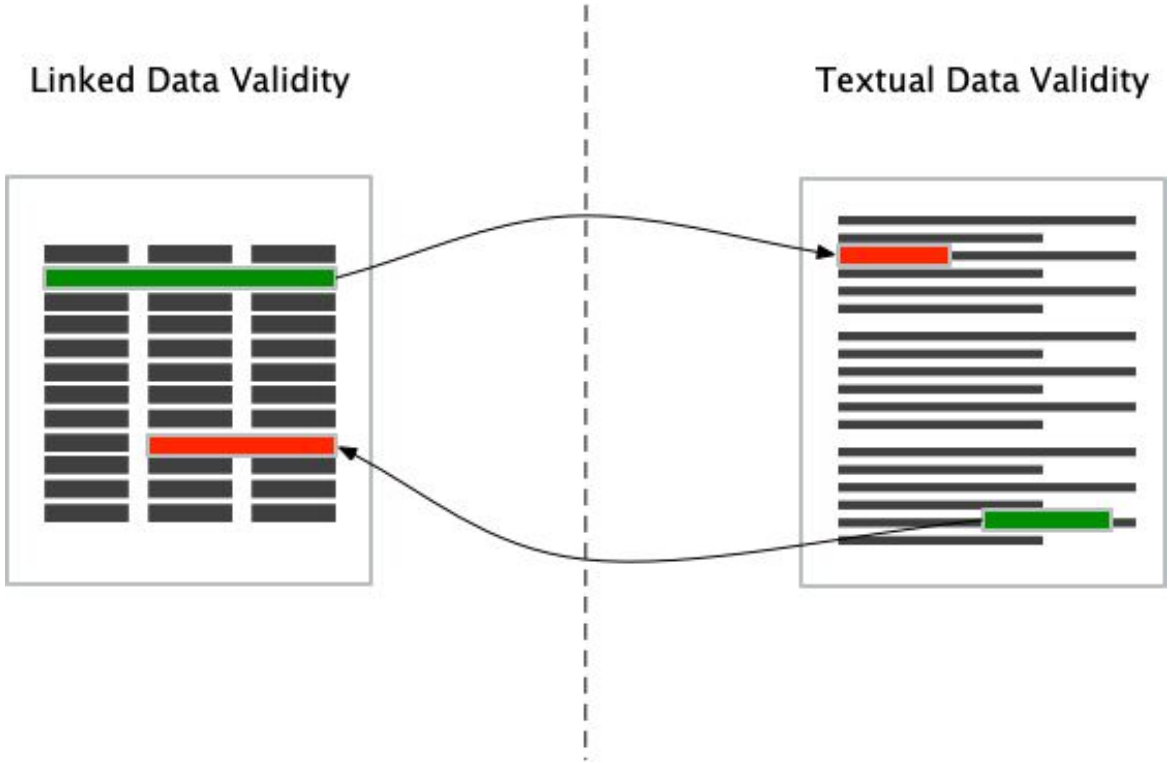
"Twenty-five hundred feet in the air rises Eryx, today Monte San Giuliano, reached from Trapani by an interminably winding but easy road that twists and turns half a hundred times in its ascent"

Definition: Data validity

- We consider the data element as valid whenever:
  - an extracted entity (with its properties) is referring to an entity in a trusted Linked Data database, or
  - an entity exists in a linked dataset (with its properties) is referring to an entity described in a trusted text

- Generate RDF triples from texts using an NLP pipeline
  - Match the RDF triples based on our assumption
    - If the information is consistent → RDF triple is valid (according to the textual data)

Process overview



Process overview (cont)

Example:

- DBpedia contains an RDF triple (dbr:Istanbul dbo:populationMetro 14,657,434).

However, in:

- “The most populated province was İstanbul with 15 million 29 thousand 231 inhabitants, constituting 18.6% of Turkey’s population”
  - If we can extract the RDF triple (dbr:Istanbul dbo:populationMetro 15,029,231) from this text and compare it to the triple present in DBpedia
    - we can assess that based on this review, the population size of Istanbul is 15,029,231 and that the old value is not valid anymore.

Use case: Historical Travel Writings (Textual resource)

- “ Two days we have passed with the ancients… Visions of Italy between XIX and XX century ”
  - 57 books that contain travel writings about Italy between (1867-1932)

- Some issues:
  - Contradicting information due to various updates on geographical entities
  - Missing or invalid information (not Italian natives, and not experts)
  - Contains non-factual data (travelers’ opinions and impressions)

Linked data resources

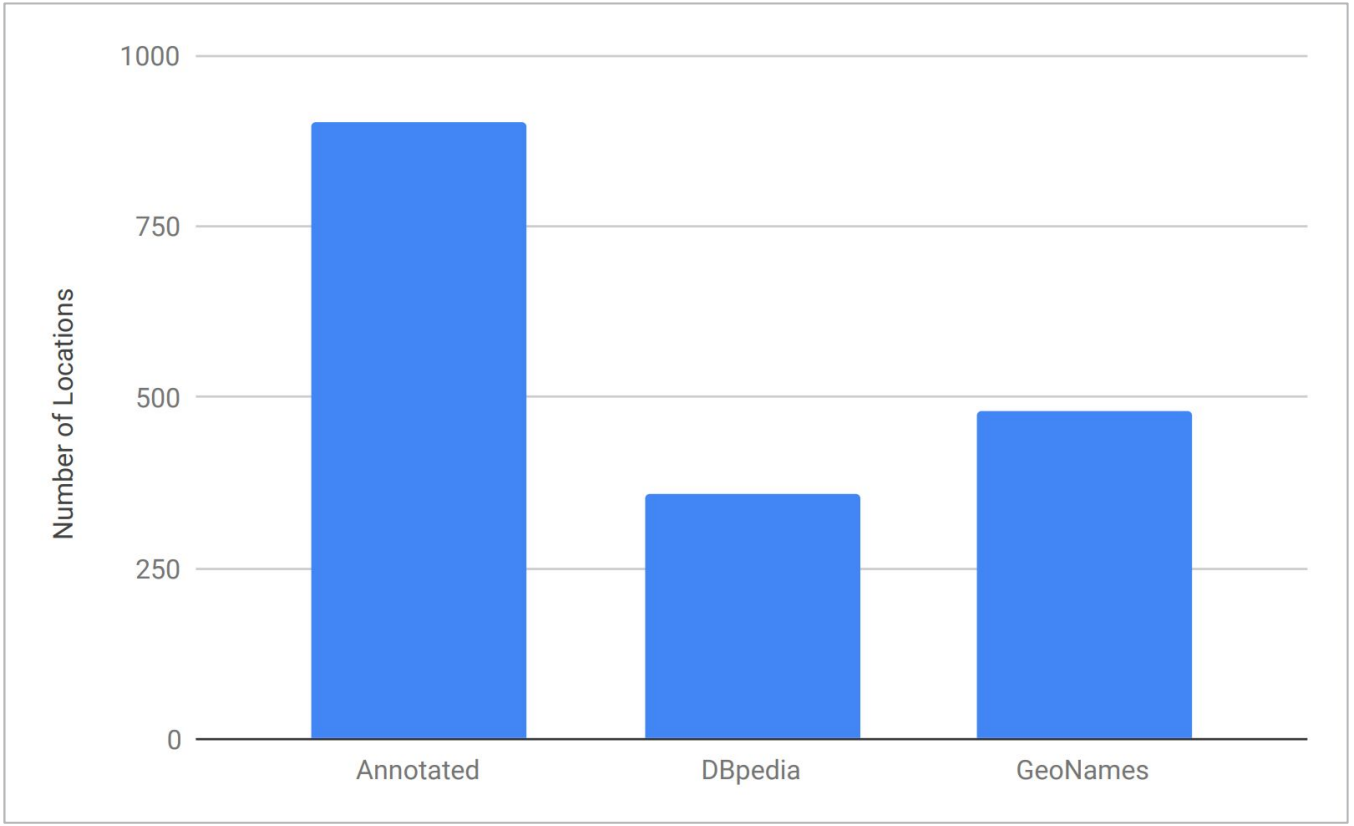
To validate the locations we rely on:

- GeoNames:
  - Database of geographical names that describes more than 11 million location entities
- DBpedia:
  - Database based on Wikipedia that contains around 735,000 location entities

Validating Extractions

- In the 57 books:
  - 2,226 locations are annotated (903 unique locations)
  - We disambiguate each location using GeoNames and DBpedia based on string matching
- We found links for fewer than half the entities in both DBpedia and GeoNames
  - This indicates gaps in the linked data resources preventing us from using the linked data resource to validate information from texts, or to further enrich them
  - We only look at recall here, and precision is not evaluated formally so the actual number of correctly disambiguated entities is very likely lower

Results


1000
750
500
250
0
Number of Locations
Annotated
DBpedia
GeoNames

Conclusion and Future Work

- We suggest a combination of NLP and linked data that can be used to check the validity of location entities
  - Whilst combining NLP and linked data is not new, our use case illustrates that this topic deserves more attention

- Future work:
  - Investigate different types (persons, organizations, etc.)
  - In case of evolution (changing information), we need to deal with the changes in order to facilitate the querying tasks
  - To have an automatic framework for data validation that combines both NLP and linked data

Acknowledgment

- This work was made possible by the International Semantic Web Research Summer School (ISWS) 2018 in Bertinoro, Italy.

- The authors would like to thank the summer school organizers, the tutors, and the fellow students in particular Amanda Pacini de Moura, Amr Azzam and Amina Annane for their suggestions and input.