


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

**Ilkcan Keles** 

Aalborg University, Dept. of Computer Science, Denmark  
ilkcan@cs.aau.dk

**Omar Qawasmeh** 

Univ. Lyon, CNRS, Lab. Hubert Curien UMR 5516, F-42023 Saint-Étienne, France  
omar.alqawasmeh@univ-st-etienne.fr

**Tabea Tietz** 

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany  
Karlsruhe Institute of Technology, Germany  
tabea.tietz@fiz-karlsruhe.de

**Ludovica Marinucci** 

Semantic Technology Laboratory (STLab), Istituto di Scienze e Tecnologie della  
Cognizione-Consiglio Nazionale delle Ricerche (ISTC-CNR), Rome, Italy  
ludovica.marinucci@istc.cnr.it

**Roberto Reda** 

Department of Computer Science and Engineering, University of Bologna, Italy  
roberto.reda@unibo.it

**Marieke van Erp** 

KNAW Humanities Cluster, DHLab, The Netherlands  
marieke.van.erp@dh.huc.knaw.nl

---

## Abstract

The Web of Data has grown explosively over the past few years, and as with any dataset, there are bound to be invalid statements in the data, as well as gaps. Natural Language Processing (NLP) is gaining interest to fill gaps in data by transforming (unstructured) text into structured data. However, there is currently a fundamental mismatch in approaches between Linked Data and NLP as the latter is often based on statistical methods, and the former on explicitly modelling knowledge. However, these fields can strengthen each other by joining forces. In this position paper, we argue that using linked data to validate the output of an NLP system, and using textual data to validate Linked Open Data (LOD) cloud statements is a promising research avenue. We illustrate our proposal with a proof of concept on a corpus of historical travel stories.

**2012 ACM Subject Classification** Computing methodologies → Natural language processing

**Keywords and phrases** data validity, natural language processing, linked data

**Digital Object Identifier** 10.4230/OASICS.LDK.2019.13

**Category** Short Paper

**Acknowledgements** This work was made possible by the *International Semantic Web Research Summer School* in Bertinoro, July 2018. The authors would like to thank the Summer School directors, Valentina Presutti and Harald Sack, as well as the tutors, the organizing team and the fellow students, in particular Amanda Pacini de Moura, Amr Azzam and Amina Annane for their suggestions and input.



© Ilkcan Keles, Omar Qawasmeh, Tabea Tietz, Ludovica Marinucci, Roberto Reda, and Marieke van Erp;

licensed under Creative Commons License CC-BY

2nd Conference on Language, Data and Knowledge (LDK 2019).

Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 13; pp. 13:1–13:8



Open Access Series in Informatics

OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

Even today, most of the content on the Web is available only in unstructured format, and in natural language text in particular. As large volumes of non-electronic textual documents, such as books and manuscripts in libraries and archives, are being digitised, undergoing optical character recognition (OCR) and made available online [12], we are presented with a huge potential of unstructured data that could feed the growth of the Linked Data Cloud.<sup>1</sup>

To integrate this content into the Web of Data, we need effective and efficient techniques to extract and capture the relevant data [5]. Natural Language Processing (NLP) encompasses a variety of computational techniques for the automatic analysis and representation of human language. As such, NLP can arguably be used to produce structured datasets from unstructured textual documents, which in turn could be used to enrich, compare and/or match with existing Linked Data sets. However, NLP systems are not without errors, and neither is Linked Data. We therefore need to ensure that information contained in structured datasets is valid.

This raises two main issues for data validity: **Textual Data Validity**, defined as the validity of information contained in texts, and **Linked Data validity**, defined as the validity of information contained in structured datasets, e.g. DBpedia or GeoNames. Textual data validity corresponds to the case whether one is not sure regarding whether the text contains correct or up-to-date information. Texts are not always written to be updated, for example a travel diary of a person provides his/her experiences during a specific time period using the information valid at that time. Unless particularly interested in providing a travel guide for future travellers, authors often do not return to their original text to add updates. For example, the updated location names remained unchanged in the text. By connecting information in such a publication to more recently updated information, such as a gazetteer that contains information on changes of location names, we can find out the place the author mentions in the text. To illustrate, if the text contains the name of ‘Monte San Giuliano’, we can infer that it corresponds to the contemporary location named ‘Erice’.<sup>2</sup> On the other hand, linked data validity corresponds to the case where the validity of the structured datasets is under question since not all structured datasets contain correct information. For this reason, by connecting a dataset to a text, for example to the original source material, statements in a database can be checked with respect to the information provided by the text. A schematic overview of this process is presented in Figure 1.

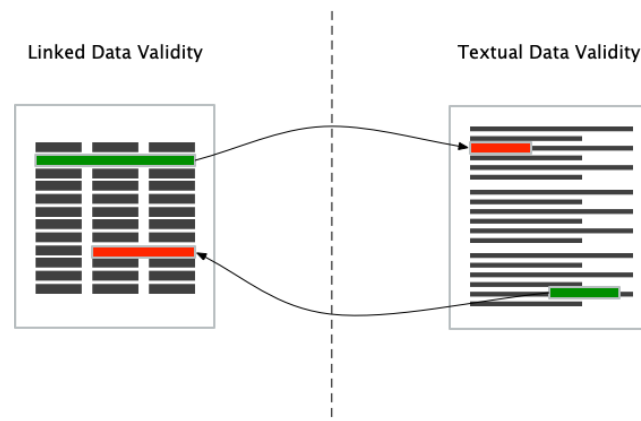
We propose that structured data extracted from text through NLP is a fruitful approach to address both issues, depending on the case at hand: structured data from reliable sources could be used to validate data extracted with NLP, and reliable textual sources could be processed with NLP techniques to be used as a reference knowledge base to validate Linked Data sets. This leads us to our definition of validity that covers both cases from an NLP perspective: We assess the data element as valid

- whenever an entity is extracted from a text and refers to an entity in a trusted Linked Data dataset and the entity’s properties extracted from text are aligned with the trusted dataset, or
- when an entity is present in a structured dataset, refers to an entity described in a trusted text and the entity’s properties are aligned with the information extracted from the trusted text.

---

<sup>1</sup> Linked Open Data Cloud <http://lod-cloud.net/> Last retrieved 10 January 2019

<sup>2</sup> <https://en.wikipedia.org/wiki/Erice> Last retrieved: 10 January 2019



■ **Figure 1** Interplay between Linked Data Validity and Textual Data Validity where Linked Data can be used to validate information contained in text, and information contained in text can be used to validate information contained in Linked Data.

Trust in this sense refers to metadata quality (e.g. precision and recall) as well as intrinsic data qualities [1].

In order to demonstrate this, we performed an analysis on a corpus of Italian travel writings by native English speakers<sup>3</sup> to extract data on locations, and then matched the extracted data with the two structured open data sets on geographic locations.

The remainder of this paper is structured as follows: Section 2 presents related work. Section 3 presents the use case description, highlighting the issues with the current disconnect between linked data and text. Section 4 concludes this work.

## 2 Related Work

Our proposed approach relies on using external knowledge bases in order to validate the quality of locations' named entities in historical travel writings, thus placing it in the realm of entity linking [7]. Whilst entity linking can cover a variety of entity types, we focus on location linking, which presents a host of problems specific to the geographical information systems domain.

Existing approaches for identifying which location names refer to which localities are summarized in [11]. The article describes the positional uncertainties and extent of vagueness frequently associated with the place names and with the differences between common users perception and the representation of places in gazetteers. The article focuses on approaches from the search/information retrieval domain, which often cannot benefit from potentially rich background information that linked data sources can provide.

A venture into location linking using semantic web resources is presented in [10]. In this paper, Van Erp et al. propose an automatic approach for georeferencing textual localities identified in a database of animal specimens using GeoNames,<sup>4</sup> Google Maps and the Global Biodiversity Information Facility (GBIF) [8].

<sup>3</sup> <https://sites.google.com/view/travelwritingsonitaly/> Last retrieved 10 January 2019

<sup>4</sup> <https://geonames.org> Last retrieved 10 January 2019

## 13:4 Validating Textual and Linked Data

An approach for historical entity linking is presented in [3]. Two use cases are presented:

1. Histpop: the Online Historical Population Reports for Britain and Ireland (1801 to 1937) and
2. BOPCRIS: the Journals of the House of Lords (1688 to 1854).

A ranking system to validate the extracted places by taking advantage of GeoNames and Wikipedia is presented. However, the authors do not make any assumptions about whether the data in GeoNames or the sources from which they extract information is valid or not.

Ceolin et al. [2] propose an approach to address the uncertainty of categorical Web data. They used Beta-Binomial, Dirichlet-Multinomial and Dirichlet Process models in order to handle the validity issue. The authors focus on two validity issues, which are the validity of multi-authoring (i.e. the nature of the web data) and the time variability. In this paper, we address the general validity without focusing on the possible sources of invalidity.

### 3 Use case: Historical Travel Writings

In this section, we describe our use case through a corpus of historical travel writings which we try to validate against several widely used knowledge bases.

#### 3.1 Resource

We have chosen to work with a corpus of historical writings regarding travel itineraries named as ‘Two days we have passed with the ancients... Visions of Italy between XIX and XX century’ [9].<sup>5</sup> We propose that this dataset provides rich use cases for addressing the textual data validity defined in Section 1.

1. It contains 57 books that correspond to the accounts written by travelers who are native English speakers traveling in Italy.
2. The corpus consists of the accounts of travelers who have visited Italy within the period of 1867 and 1932. These writings share a common genre, namely ‘travel writing’. Therefore, we expect to extract location entities that are valid during the time of the travelling. However, given that the corpus covers a span of 75 years, it potentially includes cases of contradicting information due to various updates on geographical entities.
3. The corpus might also contain missing or invalid information due to the fact that the travelers included in the dataset are not Italian natives, and therefore we cannot assume that they are experts on the places they visited.
4. The corpus also contains pieces of non-factual data, such as the travelers’ opinions and impressions.

To validate the locations from the travel writings corpus, we chose structured data sources that deal with geographical entities: GeoNames<sup>4</sup> and DBpedia.<sup>6</sup> GeoNames is a database of geographical names that describes more than 11 million location entities. The project was initiated by geographical information retrieval researchers. The core database is provided by official government sources and users are able to update and improve the database by manually editing its information. Ambassadors from all continents contribute to the GeoNames dataset with their specific expertise.

---

<sup>5</sup> Italian Travel Writings Corpus <https://sites.google.com/view/travelwritingsonitaly/> Last retrieved 10 January 2019

<sup>6</sup> <https://dbpedia.org>

In addition to a dedicated geographical dataset, we selected DBpedia, the structured database based on Wikipedia, the crowdsourced encyclopaedia. The current version of DBpedia contains around 735,000 places. Information in DBpedia is not updated live, but around twice a year, thus, it is not sensitive to live information, e.g. an earthquake in a certain location or a sudden political conflict between states. However, since working with historical data in this case study and not with live events, we pose that it is reasonable to include geographical information from DBpedia. An added feature of DBpedia over Geonames is that it contains more contextual information about a location which may help the validation process.

### 3.2 Approach

Textual data validity is difficult to separate from the information extraction process from text, as in that process often background resources are also used. However, to validate an extracted piece of information from text, we propose that deeper background knowledge is used than is customary. Many approaches such as DBpedia spotlight [6] utilize some information from the Wikipedia abstract as well as general information on the knowledge resource. Ideally, multiple resources are used, as well as domain-specific resources and reasoning over the domain, as laid out in [4].

Linked Data validity refers to the validation of Linked Data. To identify whether a given RDF triple is valid or not, we propose to find evidence for a given triple in texts. We propose to generate RDF triples from texts using an NLP pipeline, then match these to RDF triple whose validity we aim to assess. If the information is consistent between the input and extracted relations, we conclude that the RDF triple is valid according to the textual data. Moreover, the proposed method can also be employed in order to find out the missing information related to the entities that are part of the structured data set. For instance, DBpedia contains an RDF triple (`dbr:Istanbul dbo:populationMetro 14,657,434`). However, we have a document that is published recently that has a statement ‘The most populated province was İstanbul with 15 million 29 thousand 231 inhabitants, constituting 18.6% of Turkey’s population’<sup>7</sup> 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 as of 31 December 2017, the population size of Istanbul was 15,029,231 and that the old value is not valid anymore.

### 3.3 Validating extractions

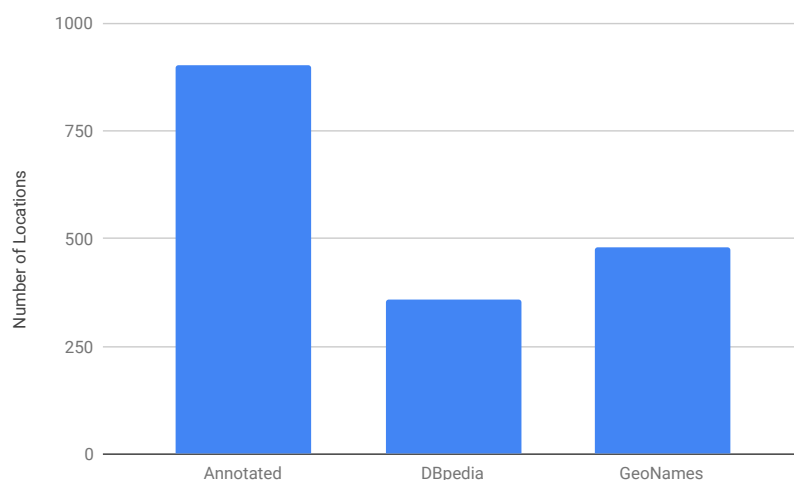
In the 57 books that comprise the travel writings on Italy corpus, 2,226 location entities are annotated, but some locations are mentioned more than once, so we identified 903 unique location strings.

We tried to automatically disambiguate each location name using GeoNames and DBpedia knowledge bases based on string matching and DBpedia spotlight [6], respectively. Figure 2 displays the number of location entities, the number of entities linked using GeoNames and the number of entities linked using DBpedia. As the graph shows, we only find links for fewer than half the entities in either resource, with GeoNames having a slightly better coverage. 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. It should be noted

---

<sup>7</sup> <http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=27587>. Last retrieved 8 January 2019.

## 13:6 Validating Textual and Linked Data



■ **Figure 2** Number of entities and entities linked from GeoNames and DBpedia.

here that we only look at recall here, and precision is not evaluated formally so the actual number of correctly disambiguated entities is very likely lower.

An example of a recall issue is a mention of the ‘chapel of San Giuliano’, between ‘Val di Genova’ and ‘Val di Borzago’<sup>8</sup> Many towns have chapels dedicated to Saint Julian, but this is a particular church located in the hills north of Trento. On current-day maps, this is called Rifugio San Giuliano, and neither the chapel, nor Val di Genova or Val di Borzago occur in Geonames or DBpedia. Deep NLP could help create linked data that encodes this information, although to georeference the exact locations, detailed maps, gazetteers and/or GIS sources would still be needed.

A big issue related to precision is that some location names are not unique; in the corpus, we find locations such as ‘Piazza’, which is used to denote the town square and can only be disambiguated in the context of knowing which town the author is talking about.

Location names are also often reused. ‘Poggio’, for example, as it is mentioned in ‘Italian Days and Ways’<sup>9</sup> probably refers to Poggio San Remo because nearby in the text Taggia and San Remo are mentioned. However, in general Poggio can refer to many different places scattered around the country.<sup>10</sup>

In order to distinguish between different locations with the same name, entity disambiguation methods need to expand the context that they take into account and go beyond sentence or paragraph barriers (as humans do). There are efficiency concerns here, as this can be computationally expensive, but we consider this a prerequisite for true deep language understanding.

An example of a location name that is both valid in only certain contexts and ambiguous as to what it exactly refers to, is ‘Monte S. Giuliano’. In the travel writings corpus, this location is described in ‘Diversions of Sicily’<sup>11</sup> as ‘This mountain, formerly world-renowned

<sup>8</sup> ‘Italian Alps Sketches in the Mountains of Ticino, Lombardy, the Trentino, and Venetia’ by Douglas William Freshfield <http://www.gutenberg.org/ebooks/45972>. Last retrieved 10 January 2019

<sup>9</sup> By A. Hollingsworth Wharton source: <https://www.gutenberg.org/ebooks/44418> Last retrieved 10 January 2019

<sup>10</sup> <https://en.wikipedia.org/wiki/Poggio> Last retrieved 10 January 2019

<sup>11</sup> By H. Festing Jones source: <https://www.gutenberg.org/ebooks/24652> Last retrieved 10 January 2019

as Mount Eryx, and still often called Monte Erice, is now Monte S. Giuliano and gives its name both to the town on the top and to the commune of which that town is the chief place.' According to Wikipedia,<sup>12</sup> the town was named back to Erice in 1934, but as 'Divisions of Sicily' was first published in 1909 and republished in 1920, the reversion back to the old name was not in there. The history of name changes is not (yet) encoded in DBpedia, GeoNames, or Pelagios<sup>13</sup> although it is present in the the Wikipedia page listing renamed places in Italy.<sup>14</sup> Analysis of this page or deep text analysis of the Erice Wikipedia page and its mention in the travel writings corpus could provide this.

## 4 Discussion and Conclusion

Textual documents are rich sources of information which due to their unstructured nature cannot easily be validated or updated automatically. Alternatively, linked data may contain invalid instances which can be checked with information coming from textual sources. We posit that a combination of natural language processing and linked data provides interesting opportunities for quality evaluation of both types of data.

In this paper, we proposed definitions for validity of textual data and Linked Data. We illustrated different aspects of validity through an analysis of a corpus of travel writings from the 19th and 20th centuries.

In our work, we focused on an analysis of validity issues of location names, which, whilst most locations will stay inhabited for a while, names of towns change. We suggested a combination of NLP and linked data can be utilised to check the validity of information as well as difficulties for these approaches. Whilst combining NLP and linked data is not new, our use case illustrates that this topic deserves more attention. In future work, aspects of validity for different types of information can be investigated. We will connect our analyses to research on trust and provenance on the semantic web, to assess and model trust and reliability.

Furthermore, we plan to extend our experiments by enriching the dataset with entity links such that we can assess the precision and work towards automating data validation. As our initial linking experiment showed that both DBpedia and GeoNames have insufficient coverage for historical location names, we will consider more knowledge bases to compare with and include other domains. We will investigate which properties and historical information about the extracted locations are useful to further automate the validation process.

---

## References

- 1 Davide Ceolin, Valentina Maccatrozzo, Lora Aroyo, and T De-Nies. Linking Trust to Data Quality. In *4th International Workshop on Methods for Establishing Trust of (Open) Data*, 2015.
- 2 Davide Ceolin, Willem Robert van Hage, Wan Fokkink, and Guus Schreiber. Estimating Uncertainty of Categorical Web Data. In *Proceedings of the 7th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2011), Bonn, Germany, October 23, 2011*, pages 15–26, 2011. URL: <http://ceur-ws.org/Vol-778/paper2.pdf>.
- 3 Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889, 2010.

---

<sup>12</sup><https://en.wikipedia.org/wiki/Erice> Last retrieved 8 January 2019

<sup>13</sup><http://commons.pelagios.org/> Last retrieved 10 January 2019

<sup>14</sup>[https://en.wikipedia.org/wiki/List\\_of\\_renamed\\_places\\_in\\_Italy](https://en.wikipedia.org/wiki/List_of_renamed_places_in_Italy) Last retrieved: 8 January 2019

- 4 Filip Ilievski, Piek Vossen, and Marieke van Erp. Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking. In *International Conference on Language, Data and Knowledge*, pages 143–149. Springer, 2017.
- 5 Andrew McCallum. Information extraction: distilling structured data from unstructured text. *ACM Queue*, 3(9):48–57, 2005. doi:10.1145/1105664.1105679.
- 6 Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.
- 7 Delip Rao, Paul McNamee, and Mark Dredze. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer, 2013.
- 8 GBIF Secretariat. GBIF Backbone Taxonomy. *Global Biodiversity Information Facility*, 2013. URL: <http://www.gbif.org/species/2879175>.
- 9 Rachele Sprugnoli. “Two days we have passed with the ancients...”: a Digital Resource of Historical Travel Writings on Italy. *SocArXiv*, 2018.
- 10 Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian van der Meij. Georeferencing Animal Specimen Datasets. *Trans. GIS*, 19(4):563–581, 2015. doi:10.1111/tgis.12110.
- 11 Maria Vasardani, Stephan Winter, and Kai-Florian Richter. Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532, 2013. doi:10.1080/13658816.2013.785550.
- 12 Iris Xie and Krystyna Matusiak. *Discover digital libraries: Theory and practice*. Elsevier, 2016.