

1 A Proposal for a Two-way Journey on Validating 2 Locations in Unstructured and Structured Data

3 **Ilkcan Keles**

4 Aalborg University, Dept. of Computer Science, Denmark

5 ilkcan@cs.aau.dk

6  <https://orcid.org/0000-0002-1424-5223>

7 **Omar Qawasmeh**

8 Univ. Lyon, CNRS, Lab. Hubert Curien UMR 5516, F-42023 Saint-Étienne, France

9 omar.alqawasmeh@univ-st-etienne.fr


10  <https://orcid.org/0000-0002-3461-4698>

11 **Tabea Tietz**

12 FIZ Karlsruhe - Leibniz Institute for Information Infrastructure & Karlsruhe Institute of

13 Technology, Germany

14 tabea.tietz@fiz-karlsruhe.de

15  <https://orcid.org/0000-0002-1648-1684>

16 **Ludovica Marinucci**

17 Semantic Technology Laboratory (STLab), Istituto di Scienze e Tecnologie della

18 Cognizione-Consiglio Nazionale delle Ricerche (ISTC-CNR), Italy

19 ludovica.marinucci@istc.cnr.it

20  <https://orcid.org/0000-0002-1605-8819>

21 **Roberto Reda**

22 Department of Computer Science and Engineering, University of Bologna, Italy


23 roberto.reda@unibo.it

24  <https://orcid.org/0000-0002-7566-8561>

25 **Marieke van Erp**

26 KNAW Humanities Cluster, DHLab, the Netherlands

27 marieke.van.erp@dh.huc.knaw.nl

28  <https://orcid.org/0000-0001-9195-8203>

29 — Abstract —

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

39 **2012 ACM Subject Classification** Computing methodologies → Artificial intelligence → Natural
40 language processing

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

42 **Digital Object Identifier** 10.4230/OASIS.LDK.2019.8



© 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 and Gerard de Melo; Article No. 8; pp. 8:1–8:8

OpenAccess Series in Informatics



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

43 **Category** Position paper

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

49 **1 Introduction**

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

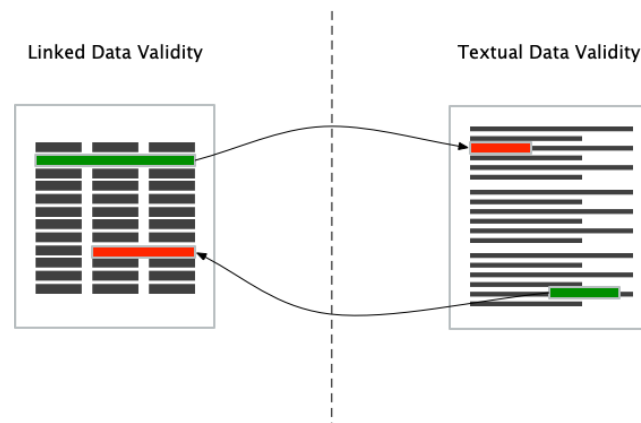
55 To integrate this content into the Web of Data, we need effective and efficient techniques to
56 extract and capture the relevant data [5]. Natural Language Processing (NLP) encompasses
57 a variety of computational techniques for the automatic analysis and representation of
58 human language. As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

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

81 We propose that structured data extracted from text through NLP is a fruitful approach
82 to address both issues, depending on the case at hand: structured data from reliable sources
83 could be used to validate data extracted with NLP, and reliable textual sources could be
84 processed with NLP techniques to be used as a reference knowledge base to validate Linked

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

² <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.

85 Data sets. This leads us to our definition of validity that covers both cases from an NLP
86 perspective: We assess the data element as valid

- 87 ■ whenever an entity is extracted from a text and refers to an entity in a trusted Linked
88 Data dataset and the entity's properties extracted from text are aligned with the trusted
89 dataset, or
- 90 ■ when an entity is present in a structured dataset, refers to an entity described in a trusted
91 text and the entity's properties are aligned with the information extracted from the
92 trusted text.

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

95 In order to demonstrate this, we performed an analysis on a corpus of Italian travel
96 writings by native English speakers³ to extract data on locations, and then matched the
97 extracted data with the two structured open data sets on geographic locations.

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

101 2 Related Work

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

107 Existing approaches for identifying which location names refer to which localities are
108 summarized in [11]. The article describes the positional uncertainties and extent of vagueness
109 frequently associated with the place names and with the differences between common users
110 perception and the representation of places in gazetteers. The article focuses on approaches

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

111 from the search/information retrieval domain, which often cannot benefit from potentially
 112 rich background information that linked data sources can provide.

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

117 An approach for historical entity linking is presented in [3]. Two use cases are presented:
 118 **1.** Histpop: the Online Historical Population Reports for Britain and Ireland (1801 to 1937)
 119 and **2.** BOPCRIS: the Journals of the House of Lords (1688 to 1854). A ranking system to
 120 validate the extracted places by taking advantage of GeoNames and Wikipedia is presented.
 121 However, the authors do not make any assumptions about whether the data in GeoNames or
 122 the sources from which they extract information is valid or not.

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

128 **3 Use case: Historical Travel Writings**

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

131 **3.1 Resource**

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

- 136 **1.** It contains 57 books that correspond to the accounts written by travelers who are native
 137 English speakers traveling in Italy.
- 138 **2.** The corpus consists of the accounts of travelers who have visited Italy within the period of
 139 1867 and 1932. These writings share a common genre, namely “travel writing”. Therefore,
 140 we expect to extract location entities that are valid during the time of the travelling.
 141 However, given that the corpus covers a span of 75 years, it potentially includes cases of
 142 contradicting information due to various updates on geographical entities.
- 143 **3.** The corpus might also contain missing or invalid information due to the fact that the
 144 travelers included in the dataset are not Italian natives, and therefore we cannot assume
 145 that they are experts on the places they visited.
- 146 **4.** The corpus also contains pieces of non-factual data, such as the travelers’ opinions and
 147 impressions.

148 To validate the locations from the travel writings corpus, we chose structured data
 149 sources that deal with geographical entities: GeoNames⁴ and DBpedia.⁶ GeoNames is a

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

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

⁶ <https://dbpedia.org>

150 database of geographical names that describes more than 11 million location entities. The
151 project was initiated by geographical information retrieval researchers. The core database
152 is provided by official government sources and users are able to update and improve the
153 database by manually editing its information. Ambassadors from all continents contribute to
154 the GeoNames dataset with their specific expertise.

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

164 3.2 Approach

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

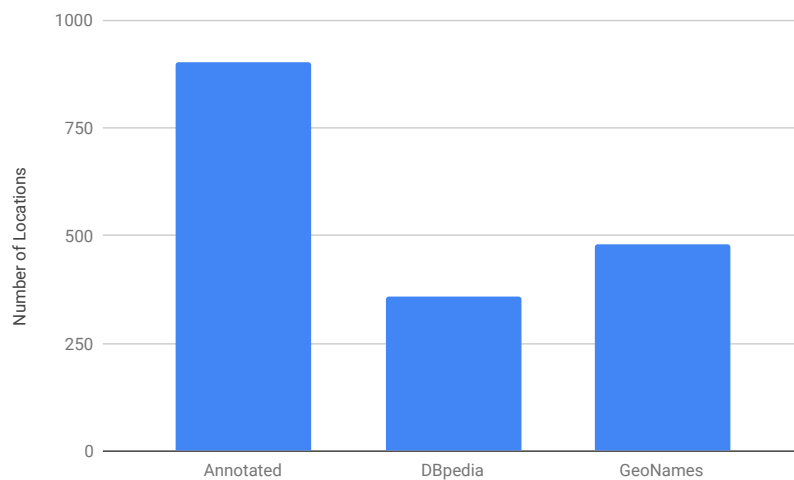
172 Linked Data validity refers to the validation of Linked Data. To identify whether a
173 given RDF triple is valid or not, we propose to find evidence for a given triple in texts. We
174 propose to generate RDF triples from texts using an NLP pipeline, then match these to RDF
175 triple whose validity we aim to assess. If the information is consistent between the input
176 and extracted relations, we conclude that the RDF triple is valid according to the textual
177 data. Moreover, the proposed method can also be employed in order to find out the missing
178 information related to the entities that are part of the structured data set. For instance,
179 DBpedia contains an RDF triple (`dbr:Istanbul dbo:populationMetro 14,657,434`). However,
180 we have a document that is published recently that has a statement “The most populated
181 province was İstanbul with 15 million 29 thousand 231 inhabitants, constituting 18.6% of
182 Turkey’s population”⁷ If we can extract the RDF triple (`dbr:Istanbul dbo:populationMetro`
183 `15,029,231`) from this text and compare it to the triple present in DBpedia, we can assess
184 that as of 31 December 2017, the population size of Istanbul was 15,029,231 and that the old
185 value is not valid anymore.

186 3.3 Validating extractions

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

190 We tried to automatically disambiguate each location name using GeoNames and DBpedia
191 knowledge bases based on string matching and DBpedia spotlight [6], respectively. Figure 2

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



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

192 displays the number of location entities, the number of entities linked using GeoNames and
 193 the number of entities linked using DBpedia. As the graph shows, we only find links for fewer
 194 than half the entities in either resource, with GeoNames having a slightly better coverage.
 195 This indicates gaps in the linked data resources preventing us from using the linked data
 196 resource to validate information from texts, or to further enrich them. It should be noted
 197 here that we only look at recall here, and precision is not evaluated formally so the actual
 198 number of correctly disambiguated entities is very likely lower.

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

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

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

213 In order to distinguish between different locations with the same name, entity disam-
 214 biguation methods need to expand the context that they take into account and go beyond
 215 sentence or paragraph barriers (as humans do). There are efficiency concerns here, as this

⁸ “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

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

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

216 can be computationally expensive, but we consider this a prerequisite for true deep language
217 understanding.

218 An example of a location name that is both valid in only certain contexts and ambiguous
219 as to what it exactly refers to, is ‘Monte S. Giuliano’. In the travel writings corpus, this
220 location is described in ‘Diversions of Sicily’¹¹ as “This mountain, formerly world-renowned
221 as Mount Eryx, and still often called Monte Erice, is now Monte S. Giuliano and gives its
222 name both to the town on the top and to the commune of which that town is the chief place.”
223 According to Wikipedia,¹² the town was named back to Erice in 1934, but as “Diversions
224 of Sicily” was first published in 1909 and republished in 1920, the reversion back to the
225 old name was not in there. The history of name changes is not (yet) encoded in DBpedia,
226 GeoNames, or Pelagios¹³ although it is present in the the Wikipedia page listing renamed
227 places in Italy.¹⁴ Analysis of this page or deep text analysis of the Erice Wikipedia page and
228 its mention in the travel writings corpus could provide this.

229 **4 Discussion and Conclusion**

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

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

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

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

252 **References**

- 253 **1** Davide Ceolin, Valentina Maccatrozzo, Lora Aroyo, and T De-Nies. Linking trust to data
254 quality. In *4th International Workshop on Methods for Establishing Trust of (Open) Data*,

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

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

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

¹⁴ https://en.wikipedia.org/wiki/List_of_renamed_places_in_Italy Last retrieved: 8 January 2019

- 255 2015.
- 256 **2** Davide Ceolin, Willem Robert van Hage, Wan Fokkink, and Guus Schreiber. Estimating
257 uncertainty of categorical web data. In *Proceedings of the 7th International Workshop on*
258 *Uncertainty Reasoning for the Semantic Web (URSW 2011), Bonn, Germany, October 23,*
259 *2011*, pages 15–26, 2011. URL: <http://ceur-ws.org/Vol-778/paper2.pdf>.
- 260 **3** Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn,
261 and Julian Ball. Use of the edinburgh geoparser for georeferencing digitized historical
262 collections. *Philosophical Transactions of the Royal Society of London A: Mathematical,*
263 *Physical and Engineering Sciences*, 368(1925):3875–3889, 2010.
- 264 **4** Filip Ilievski, Piek Vossen, and Marieke van Erp. Hunger for contextual knowledge and a
265 road map to intelligent entity linking. In *International Conference on Language, Data and*
266 *Knowledge*, pages 143–149. Springer, 2017.
- 267 **5** Andrew McCallum. Information extraction: distilling structured data from unstructured
268 text. *ACM Queue*, 3(9):48–57, 2005. URL: <https://doi.org/10.1145/1105664.1105679>,
269 doi:10.1145/1105664.1105679.
- 270 **6** Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight:
271 shedding light on the web of documents. In *Proceedings of the 7th international conference*
272 *on semantic systems*, pages 1–8. ACM, 2011.
- 273 **7** Delip Rao, Paul McNamee, and Mark Dredze. Entity linking: Finding extracted entities in
274 a knowledge base. In *Multi-source, multilingual information extraction and summarization*,
275 pages 93–115. Springer, 2013.
- 276 **8** GBIF Secretariat. Gbif backbone taxonomy. *Global Biodiversity Information Facility*,
277 <http://www.gbif.org/species/2879175>, 2013.
- 278 **9** Rachele Sprugnoli. “Two days we have passed with the ancients...”: a Digital Resource of
279 Historical Travel Writings on Italy. *SocArXiv*, 2018.
- 280 **10** Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian van der Meij. Georeferencing
281 animal specimen datasets. *Trans. GIS*, 19(4):563–581, 2015. URL: [https://doi.org/10.](https://doi.org/10.1111/tgis.12110)
282 [1111/tgis.12110](https://doi.org/10.1111/tgis.12110), doi:10.1111/tgis.12110.
- 283 **11** Maria Vasardani, Stephan Winter, and Kai-Florian Richter. Locating place names from
284 place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–
285 2532, 2013. URL: <https://doi.org/10.1080/13658816.2013.785550>, doi:10.1080/
286 13658816.2013.785550.
- 287 **12** Iris Xie and Krystyna Matusiak. *Discover digital libraries: Theory and practice*. Elsevier,
288 2016.