Diagnosing Incompleteness in Wikidata with The Missing Path

Marie Destandau, Jean-Daniel Fekete

Diagnosing Incompleteness in Wikidata with The Missing Path

Marie Destandau
Jean-Daniel Fekete
Marie.Destandau@inria.fr
Jean-Daniel.Fekete@inria.fr
Université Paris-Saclay, CNRS, Inria, LRI
Orsay, France

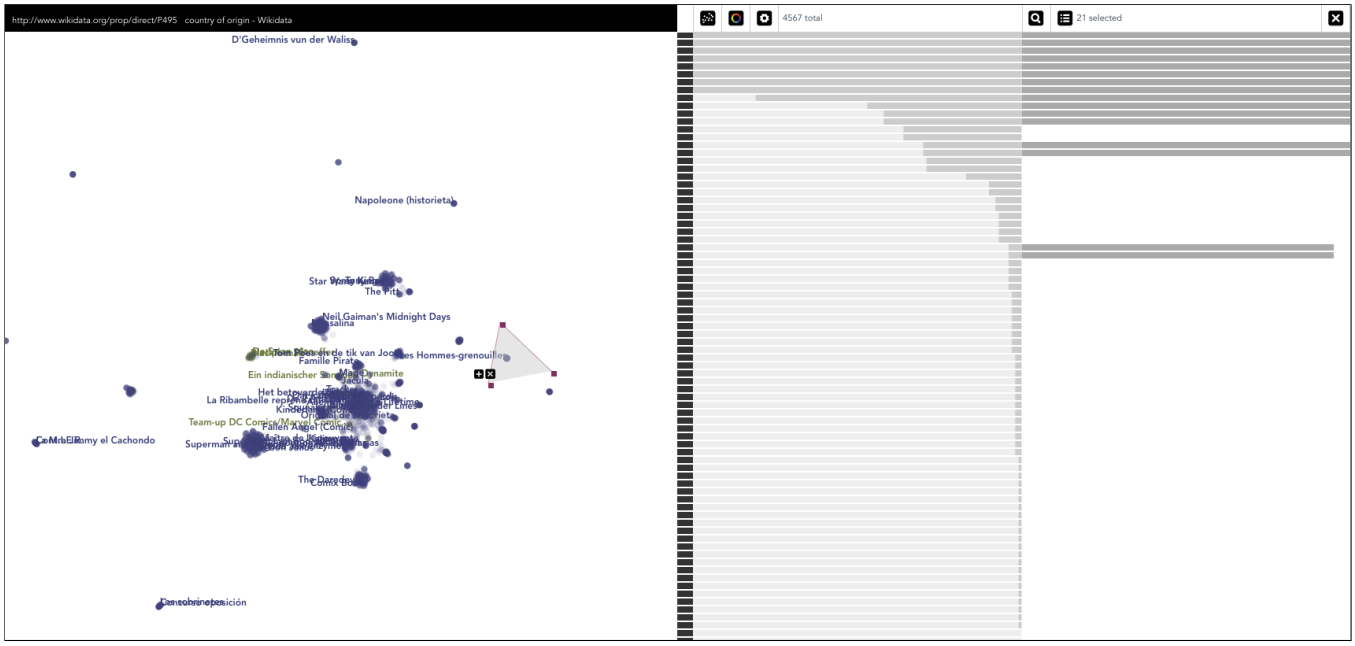

Figure 1: The Missing Path lets users identify subsets of items missing the same attributes and inspect them to find the cause of incompleteness. Dense groups of points (clusters) represent entities that share a large number of paths, meaning that they are structurally similar up to some specified path length (2 in this example). Isolated entities are therefore structurally different than others and probably inconsistently encoded. The presence of multiple clusters indicate the existence of groups of entities that are slightly different structurally, either for good reasons or because they have been entered in an inconsistent manner. With our tool, all these issues can be checked and collected for further corrections.

ABSTRACT

To make their data usable, Linked Data producers need to provide a minimum level of completeness. But the task of finding the missing attributes for a specific list of entities is notoriously difficult. We make the hypothesis that identifying subsets of entities with a similar structure can help finding the cause of incompleteness and decide if and how it could to be solved. We contribute with our visualisation tool: "The Missing Path", relying on dimensional reduction techniques to create a map of the entities based on missing properties, revealing clusters. Users can alternate their focus with the second coordinated visualisation that shows the distribution of properties, and of their values, datatypes and languages. We describe the evaluation and iterative design process we have planned with Wikidata contributors.

Wiki Workshop, 2020, Taipei

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Computer systems organization → Embedded systems**.

KEYWORDS

Linked Data, Semantic Web, Incompleteness, Wikidata

ACM Reference Format:
Marie Destandau and Jean-Daniel Fekete. 2020. Diagnosing Incompleteness in Wikidata with The Missing Path. In *Proceedings of Wiki Workshop.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The *Semantic Web* enables communities, institutions, research laboratories, and companies to combine and share data coming from different sources, and to query them jointly. It becomes possible to get, with a unique query, answers that would otherwise have requested access to several database, each with its own technical stack and data model. This opens new perspectives for data journalists, researchers, librarians, or any user or application concerned with collecting pieces of information from various sources. However this new format, coming along with new interfaces and new applications, also raises new problems. In particular, the issue of *completeness* is identified as a critical concern regarding Linked Data (LD) quality [8, 14]. For example, when querying the datasets `Nobel` and `Dbpedia` jointly to display the institutions of Nobel prize recipients on a map, the query will have to follow a chain or properties including laureates, institutions, and geographic coordinates. Each of these properties are important for the completeness of the data, but the last one related to geographic coordinates also raises consistency problems since there are multiple competing properties to describe geographic coordinates. For displaying the institutions on a map, a query needs to make a choice on the ontology, and the data provider needs to make sure the information is provided consistently. Failing to do so will produce unreliable results, in our example hiding important institutions for frivolous reasons.

We present *The Missing Path*, a tool to identify missing information related to groups of entities, to inspect them for diagnosing the reason why they miss, and to export instructions and information to support actions to remedy their absence.

For data producers, the task of finding the erroneous or missing attributes for a specific list of entities is difficult to complete. Although there are tools and methods to assess the rate of completeness of a specific property for a set of entities they do not help to identify meaningful subsets that require specific actions. For instance, checking for the death date of the Nobel laureates, the information might be missing because they are still alive, because the information is provided with an unexpected ontology, or simply because it has never been entered. Exporting the full list of laureates without birthdates would then require to check them individually to decide if they need to be fixed.

We posit that identifying groups of items sharing a similar structure can help finding the cause of incompleteness, and decide if and how it has to be resolved. We contribute "The Missing Path", a visualisation tool based on two coordinated visualisations. The first uses dimensional reduction techniques to create a 2D map of the entities based on missing attributes. The map reveals clusters or entities with similar missing structures. The tool lets users inspect these clusters in a coordinated view for diagnosing the reason why they miss. The second show a distribution of properties and allows to inspect their distribution and values. We consider not only direct properties but also chains of properties, that we name *paths*, building on previous work showing that they might convey first-order information [5]. We describe the evaluation and iterative design process we have planned with Wikidata contributors, following a methodology inspired by Multi-Dimensional in-Depth Long-Term Case Studies (MILCS) by Shneiderman & Plaisant [21].

2 RELATED WORK

2.1 Incompleteness

Though the definition of quality can have many acceptations, most of them mention the problem of completeness [3, 14, 20, 24]. Tools enabling to compute and see completeness usually consider the completeness of a property for the whole set [1] or regarding a set of entities sharing the same `rdf:type` [2, 11]. Some take into account chains of properties [5], or more elaborate clustering patterns [22]. Completeness can also be evaluated at retrieval time, as a contextual indicator to interpret a query [19].

In our work, we evaluate completeness using chains of properties, and allow interactive exploration to let users of our tool judge whether a missing path represents an issue or makes sense in a specific context.

2.2 Dimensional Reduction Techniques

Dimensional Reduction Techniques have many applications in Linked Data. They have been used to analyse the content of datasets [23], perform learning [10], estimate the similarity of items [9], support recommendation [7, 15], and evaluate the distance between ontologies [4]. Node2Vec focuses on exploring neighbourhood in graphs [6] and was also applied for item recommendation [17]. Paulheim advocates for vectors that preserves semantic and are interpretable [18].

There is a large number of dimensionality reduction techniques available with different properties [16]. We use the UMAP technique [12] which is fast and has excellent properties related to clustering.

3 THE MISSING PATH

The Missing Path supports the identification of groups of items sharing a similar structure, in order to inspect them, identify the causes of incompleteness, and decide if and how it shall be resolved.

3.1 Prior analysis

The prior analysis is done through an API taking as parameters a SPARQL endpoint URL, a similarity criteria for the collection and a maximum depth of paths of properties to analyse. The similarity criteria can be expressed in SPARQL, allowing for complex definitions. The analysis retrieves information about the entities complying with the similarity criteria, the associated properties forming the paths up to the specified depth, the values at their end, their datatype and their language.

Paths are used to populate high-dimensional vectors for each entities with boolean values indicating if each path exists. Then, the vectors are projected into a 2-dimensional (2D) map using the multidimensional projection algorithm UMAP [13] and the reduction function *russellrao* for boolean vectors.

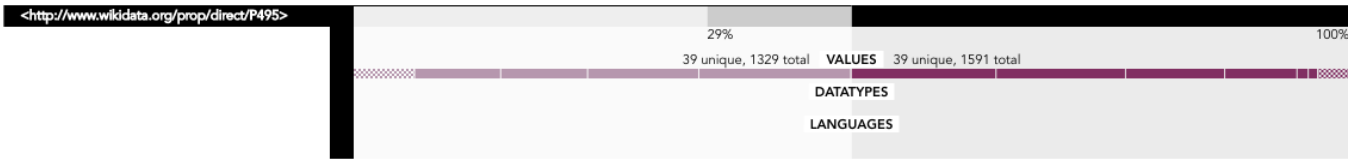

Figure 2: Detailed statistics for a path: the whole set is presented on the left, in comparison to the selection on the right

3.2 Structural overviews

The left part of the screen (Figure 1) presents a map giving an overview of all the entities in the collection. The coordinates are calculated by using the multidimensional projection of vectors indicating which properties or paths of properties are missing. By default, paths are ordered by coverage, and the 20 first that do not have a coverage of 100 percent (in which case they would not be discriminant) are considered, with a Boolean value indicating if the path is missing. The user can modify the list of paths and recompute the map, with a maximum of 50 paths considered.

Paths that contain ordinal values, or a limited number of categorical values, are candidates for color coding. By default, the best covered candidate is used to colour the entities. The user can choose another color if desired.

A second overview is proposed in the third fourth of the screen, displaying all paths ordered by coverage. Hovering one path displays the property or chain of properties of that path. Selecting it opens a box at the end of the path showing a summary of values, their data types, and languages, as show in Figure 2.

3.3 Inspection of a selection

To make sense of a *subset* of entities, users need to identify its distinctive features, what defines it in comparison to the whole set of entities. Our interface shows the summary of the set and subset facing each other. Each path can be inspected, comparing the statistics for the whole set and the subset, as detailed in Figure 2.

4 USE CASE

We describe the interface from the point of view of a contributor who wants to curate Wikidata entities of class comics (`Q1004 Comics`). She opens the tool, sees the map in Figure 1, and inspects a cluster in the selection panel. Among the well represented paths for the set, several are missing for the subset. She hovers them and finds out they represent important information for comics such as `P407 language of work or name`, `P495 country of origin`, `P123 publisher`. Looking at the summaries of values for the paths describing the cluster, she sees that 20 out of the 21 entities have the same `schema.org/description`: "stripverhaal van Robbedoes en Kwabernoot" ("comic strip Spirou & Fantasio" in Dutch), and that they were all modified on the same date. This hints that language, country, and maybe even editor information will be easy to fix consistently for this subset.

She selects another cluster and notices that `P179 part of the series` is completely missing for this subset, while `P361 part of`, which is rare, is present. Inspecting other properties for the selection in the right pane, she notices that the language of all descriptions is Luxembourger. She uses the statistical panel to replace the current selection with all comics having this `P361 part of`. Only one of them has a `P179`, which seems to confirm her guess that one is used in place of the other. In order to verify, she opens the list of items in the selection, and inspects the link to the first item. She exports those findings.

5 PLANNED EVALUATION AND ITERATIVE DESIGN

Using a methodology inspired by MILCS [21], we interviewed 6 Wikidata contributors, and asked them to describe how they contribute. Five of them were interested in trying to use the tool to visualise some of the data and try to solve some specific problems they have. The next step is a session to observe them discovering their data in the tool. Then we will leave the tool at their disposal, and ask them to report on usability problems. We will fix the tools as they report, with intermediate sessions where we watch them using it. We will report on the iterative design process as well as on the new errors they were able to identify and/or to diagnose. We also expect feedback from participants in the Wiki Workshop.

6 CONCLUSION AND FUTURE WORK

The Missing Path is a visualisation tool based on dimensional reduction techniques to create a 2D map of the entities based on missing attributes coordinated with a view showing the distribution and details of paths reachable by entities. The map reveals clusters or entities with similar missing structures. The tool lets users inspect these clusters for diagnosing the reason why they miss. This work is work in progress. We are currently conducting an evaluation and iterative design process with Wikidata contributors. We are at the beginning of the iterative design cycle, and the interface will evolve. We will evaluate how it enables to discover incomplete subsets and diagnoses the cause. The Missing Path will be available in open source form to allow the community to benefit from it and improve it. An important feature, that is not implemented yet, is the export of observations to support actions. When an action is decided, useful information to efficiently fix the problem is difficult to transmit in an understandable format to the concerned actors. We need to help users select information that are distinctive for a subset, and turn it into an exportable format.

REFERENCES

[1] Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. 2012. LODStats–an extensible framework for high-performance dataset analytics. In *International Conference on Knowledge Engineering and Knowledge Management*. Springer, 353–362.

[2] Vevake Balaraman, Simon Razniewski, and Werner Nutt. 2018. Recoin: Relative Completeness in Wikidata. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1787–1792. https://doi.org/10.1145/3184558.3191641

[3] Christian Bizer and Richard Cyganiak. 2009. Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics* 7, 1 (2009), 1 – 10. https://doi.org/10.1016/j.websem.2008.02.005 The Semantic Web and Policy.

[4] Jérôme David and Jérôme Euzenat. 2008. Comparison between Ontology Distances (Preliminary Results). In *The Semantic Web - ISWC 2008*, Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin, and Krishnaprasad Thirunarayan (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 245–260.

[5] Marie Destandau, Olivier Corby, and Alain Giboin. 2020. Path Outlines: Browsing Path-Based Summaries of Linked Open Datasets. *CoRR* abs/XXXX.YYYY (2020). arXiv:XXXX.YYYY

[6] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[7] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2013. Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems. In *On the Move to Meaningful Internet Systems: OTM 2013 (Lecture Notes in Computer Science)*, Robert Meersman, Hervé Panetto, Tharam Dillon, Johann Eder, Zohra Bellahsene, Norbert Ritter, Pieter Leenheer, , and Deijing Dou (Eds.), Vol. 8185. Springer Berlin Heidelberg, Graz, Austria, 606–615. https://doi.org/10.1007/978-3-642-41030-7_44

[8] Andreas Harth and Sebastian Speiser. 2012. On completeness classes for query evaluation on linked data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[9] Aidan Hogan, Axel Polleres, Jürgen Umbrich, and Antoine Zimmermann. 2010. Some entities are more equal than others: statistical methods to consolidate linked data. In *4th International Workshop on New Forms of Reasoning for the Semantic Web: Scalable and Dynamic (NeFoRS2010)*.

[10] Yi Huang, Volker Tresp, Maximilian Nickel, Achim Rettinger, and Hans-Peter Kriegel. 2014. A scalable approach for statistical learning in semantic graphs. *Semantic Web* 5, 1 (2014), 5–22.

[11] Subhi Issa, Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. 2019. Revealing the Conceptual Schemas of RDF Datasets. In *International Conference on Advanced Information Systems Engineering*. Springer, 312–327.

[12] L. McInnes, J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* (Feb. 2018), 51. arXiv:stat.ML/1802.03426

[13] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[14] Pablo N Mendes, Hannes Mühleisen, and Christian Bizer. 2012. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. Citeseer, 116–123.

[15] Cataldo Musto. 2010. Enhanced Vector Space Models for Content-Based Recommender Systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 361–364. https://doi.org/10.1145/1864708.1864791

[16] Luis Gustavo Nonato and Michael Aupetit. 2018. Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment. *IEEE Transactions on Visualization and Computer Graphics* 25, 8 (Aug 2018), 2650–2673. https://doi.org/10.1109/TVCG.2018.2846735

[17] Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, Elena Baralis, Michele Osella, and Enrico Ferro. 2018. Knowledge Graph Embeddings with node2vec for Item Recommendation. In *The Semantic Web: ESWC 2018 Satellite Events*, Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 117–120.

[18] Heiko Paulheim. 2018. Make embeddings semantic again!. In *International Semantic Web Conference (P&D/Industry/BlueSky)*.

[19] Radityo Eko Prasojo, Fariz Darari, Simon Razniewski, and Werner Nutt. 2016. Managing and consuming completeness information for wikidata using COOL-WD. CEUR-WS. org.

[20] Filip Radulovic, Nandana Mihindukulasooriya, Raúl García-Castro, and Asunción Gómez-Pérez. 2018. A comprehensive quality model for linked data. *Semantic Web* 9, 1 (2018), 3–24.

[21] Ben Shneiderman and Catherine Plaisant. 2006. Strategies for Evaluating Information Visualization Tools: Multi-Dimensional in-Depth Long-Term Case Studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (Venice, Italy) *(BELIV '06)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/1168149.1168158

[22] Avicenna Wisesa, Fariz Darari, Adila Krisnadhi, Werner Nutt, and Simon Razniewski. 2019. Wikidata Completeness Profiling Using ProWD. In *Proceedings of the 10th International Conference on Knowledge Capture*. 123–130.

[23] Amrapali Zaveri, Joao Ricardo Nickenig Vissoci, Cinzia Daraio, and Ricardo Pietrobon. 2013. Using Linked Data to Evaluate the Impact of Research and Development in Europe: A Structural Equation Model. In *The Semantic Web – ISWC 2013*, Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 244–259.

[24] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer, and Pascal Hitzler. 2013. Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal* (2013).