VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

A Polyvocal and Contextualised Semantic Web

van Erp, M.; de Boer, V.

published in
The Semantic Web
2021

DOI (link to publisher)
10.1007/978-3-030-77385-4_30

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

Link to publication in VU Research Portal

citation for published version (APA)
van Erp, M., & de Boer, V. (2021). A Polyvocal and Contextualised Semantic Web. In R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, & M. Alam (Eds.), The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings (pp. 506-512). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12731 LNCS). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-77385-4_30

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 28. Mar. 2024



A Polyvocal and Contextualised Semantic Web

Marieke van Erp[1(✉)] and Victor de Boer[2]

[1] KNAW Humanities Cluster, Amsterdam, The Netherlands
marieke.van.erp@dh.huc.knaw.nl
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
v.de.boer@vu.nl
https://mariekevanerp.com

Abstract. Current AI technologies and data representations often reflect the popular or majority vote. This is an inherent artefact of the frequency bias of many statistical analysis methods that are used to create for example knowledge graphs, resulting in simplified representations of the world in which diverse perspectives are underrepresented. With the use of AI-infused tools ever increasing, as well as the diverse audiences using these tools, this bias needs to be addressed in both the algorithms analysing data, as well as in the resulting representations. In this *problems to solve before you die* submission, we explain the implications of the lack of polyvocality and contextual knowledge in the semantic web. We identify three challenges for the Semantic Web community on dealing with various voices and perspectives as well as our vision for addressing it.

Keywords: Culturally aware AI · Polyvocality · Contextualisation · Bias

1 Introduction

Biases in data can be both explicit and implicit. Explicitly, 'The Dutch Seventeenth Century' and 'The Dutch Golden Age' are pseudo-synonymous and refer to a particular era of Dutch history. Implicitly, the 'Golden Age' moniker is contested due to the fact that the geopolitical and economic expansion came at a great cost, such as the slave trade. A simple two-word phrase can carry strong contestations, and entire research fields, such as post-colonial studies, are devoted to them. However, these sometimes subtle (and sometimes not so subtle) differences in voice are as yet not often found in knowledge graphs.

One of the reasons is that much of the knowledge found in knowledge graphs is mined automatically and current AI technologies (and their ensuing data representations) often reflect the popular or majority vote. This is an inherent artefact of the frequency bias of many statistical analysis methods that are used to create for example knowledge graphs, resulting in simplified representations

Supported by cultural-ai.nl.

© Springer Nature Switzerland AG 2021
R. Verborgh et al. (Eds.): ESWC 2021, LNCS 12731, pp. 506–512, 2021.
https://doi.org/10.1007/978-3-030-77385-4_30

of the world in which diverse perspectives are underrepresented. With the use of AI-infused tools ever increasing, as well as the diverse audiences using these tools, this bias needs to be addressed in both the algorithms analysing data, as well as in the resulting representations.

Conversations around data bias and polyvocality are taking place in for example the cultural heritage domain (cf. [14,15]) and computational linguistics communities (cf. [8,11]), but do not yet seem mainstream in the Semantic Web discourse. In 2010, Hendler and Berners-Lee already recognised that current knowledge mining mechanisms can bias results and recommend 'Making the different ontological commitments of competing interpretations explicit, and linked together', as this 'can permit different views of data to be simultaneously developed and explored.' [9]. Veltman noted in her 2004 paper that '[Dynamic knowledge will] allow us to trace changes of interpretation over time, have new insights and help us to discover new patterns in knowledge' [20]. Context was also flagged as the next frontier in knowledge representation at the 'Knowledge graphs: New directions for knowledge representation on the semantic web' Dagstuhl seminar in 2018 [5]. While it is known that large knowledge graphs are not always balanced (cf. [12]), that links between linked data resources contain biases (cf. [19]), and that research based on such resources might favour a Western perspective (cf. [6]), the creation of contextualised and polyvocal knowledge graphs has only gained modest traction.

As data-driven applications are permeating everyday life, it is necessary that these applications can serve as many and as diverse audiences as possible. A Dutch visitor to the Rijksmuseum in Amsterdam could recognise the Christian context of a painting depicting Saint Christopher even if s/he does not know that he is the patron saint of travellers. For a Japanese visitor, this might not be immediately apparent, and s/he needs more context to connect this to her/his cultural context, to for example recognise the parallels to Jizō, protector of travellers in some Buddhist traditions. For most other applications (e.g. in the domains of food, health, policy and transportation) a greater awareness of cultural and contextual frameworks of users would be crucial to increase engagement and understanding. Thus far, one of the ways machine learning driven approaches have dealt with this, is to add more data, at the risk of flattening nuances. At times, this has led to painful situations, for example when a Twitter trained chatbot started publishing racist, sexist and anti-semitic Tweets within a day of being released [23]. We argue that addressing this by more intelligently organised datasets to identify bias, and multiple perpectives and contexts is needed. Many of the puzzle pieces for a polyvocal and contextualised Semantic Web are already in place. The different local Wikipedias for example present different perspectives [4,24] with varying levels of in- and between group biases [1] but this is often still expressed implicitly.

In this *problems to solve before you die* submission, we explain the implications of the lack of polyvocality and contextual knowledge in the semantic web, as well as our vision for addressing it. The examples we use come from cultural heritage datasets as we work with those and some of the bias issues are amplified

in there due to their longue durée,[1] but these issues and our proposed approach also apply to other domains.

2 What Is a Voice?

Data is not objective, but rather created from a particular perspective or view, representing a *voice*. These perspectives can be informed by cultural, historical or social conventions or a combination of those. Often, these different perspectives are contiguous, rather than disjoint. A cultural view on for example the 'The Dutch Golden Age' is that it was an era in which the Dutch economy and scientific advancements were among the most acclaimed in the world, laying the foundations for the first global multinational corporation and shaping Dutch architecture. A historical view could be that it was an era of unbridled opportunity and wealth for many in the Dutch Republic. A social view could focus on the pride associated with the achievements of this era. However, none of these views tells the entire story. The Dutch Republic maintained several colonies in Africa and Asia and was heavily involved in slave trading; the other side of the coin of wealth and pride.

These different views often come to us through objects in cultural heritage collections which can often be different things throughout their life. For example, many cult figures and other objects were taken and brought back home by missionaries -in some cases to learn about 'the other' from a euro-centric view. Such items were often (dis)qualified as "fetishes" or "idols" and Christian converts were expected to stay away converted from them. Later, these objects were removed from their original ritual settings and became part of ethnographic collections in missionary exhibitions or were sold as works of art. These reframings represent new perspectives on the same object and depend on time, culture and the object provenance. Data models that are perspective-aware should be able to trace the various changes that objects undergo in their trajectory from their original uses in indigenous religious practices into museum collections [13].

A polyvocal Semantic Web provides opportunities, models and tools to identify, represent and show users different perspectives on an event, organisation, opinion or object. Furthermore, identified individual perspectives need to be connected and clustered to be lifted from an individual perspective to a (representative) voice of a group. Identifying, representing and using such groups with fuzzy boundaries that change over time is one of the core challenges.

3 How Can the Semantic Web Deal with Multiple Perspectives and Interpretations?

We identify three main challenges to support polyvocality on the Semantic Web: identification, representation and usage. We describe these challenges below, along with the current state of the art and promising research directions for addressing them.

[1] Some collections we work with have been gathered over a span of over 200 years, and the objects they describe go back even longer.

3.1 Identifying Polyvocality in Data Sources

Data sources of various modalities can represent a singular 'voice', and provide a specific perspective on the world. However, within and across data sources, there can also be multiple voices present. Moreover, when aggregating datasets, such voices can become lost in the process. The challenge here is to identify such voices. For natural language data sources, as well as for methods that extract information from structured sources, this requires Information Extraction methods that are 'bias' or 'voice'-aware. In cases where datasets are the result of aggregations, such methods should be able to (re-)identify the separate voices in the combined source.

In cases where human data providers are involved, the voices of the individual users should be maintained. These persons can also represent cultural or societal groups and this information should be available for subsequent representations. As an example, consider a crowdsourcing effort to annotate a Polynesian object in a European museum. Here, annotations provided by the European public on the one hand and annotations provided by members of the source community on the other hand will represent various perspectives. Methods that aggregate such annotations should retain these voices. An example of a method that does this for individual annotators is CrowdTruth [2], which maintains so-called 'disagreement' between annotators. Such methods can be expanded to retain this disagreement to the extent to which it represents a 'voice'.

What holds for these non-professional annotators, also holds for professionals involved in metadating or interpreting information. Different perspectives provided by such professionals should also be maintained in the information processing pipeline. Again, in the cultural heritage domain, with movements towards a diversification of the museum professionals gaining traction [18,22], the different perspectives that such professionals provide should be maintained in the object metadata.

3.2 Representation of Polyvocality: Datamodels and Formalisms

Identifying, extracting and retaining polyvocal information in data sources is one aspect, but this is meaningless if the various voices cannot be represented in the data structures used. Luckily, the Knowledge Graph as a data model and the Semantic Web as an information architecture provide excellent opportunities for maintaining various viewpoints on one subject. Its network structure and distributed nature is well-equipped to deal with such viewpoints. One example is found in Europeana, where the Europeana Data Model [10] allows for multiple publishers to provide information about one cultural heritage object. Building on the OAI-ORE model[2], each data provider provides a "Proxy" resource that represents that object in the context of that provider and all metadata is attached to that Proxy resource, rather than the object representation itself. This allows to have multiple perspectives on a single real-world object. This model has

[2] http://www.openarchives.org/ore/1.0/datamodel.

also been applied to represent multiple and potentially disagreeing biographical descriptions of persons [16]. In the NewsReader project,[3] information on events was mined from newspaper articles. Multiple reports on the same event were identified via in- and cross-document coreference resolution and presented using named graphs to group statements about a particular event. Provenance information is then attached to these named graphs to represent the viewpoints of various providers [7,17].

Where a voice corresponds to the view of a data publishers, such models appear adequate. At the other end of the spectrum, we see provenance at the level of the individual annotator or even the individual statement. WikiData for example, represents each statement as a tuple which includes the metainformation about that statement [21]. This allows for the recording of very precise provenance of statements, to the individual level.

The view of a data provider and the view of the individual annotator represent two points on the spectrum that can inform how to represent the more collective and elusive social-cultural voice. Here, we need models and design patterns on how voices can be represented. This means that not only "what is expressed" should be represented, but also "which worldview does this represent". One direction is that of "data lenses" as deployed for example in the OpenPhacts project [3]. Here sets of statements are annotated with provenance information, which in turn represent the scientific worldview of different types of end-users. In our case, we would like such lenses to correspond to various voices regardless of the end-user.

3.3 Usage of Polyvocal Knowledge

Representing and maintaining polyvocal knowledge is meaningless if it cannot be used in end-user facing applications. With the voices being represented in the knowledge graphs, different viewpoints are available for such applications. How to provide access to this more complex information in a meaningful way is a third major research challenge. The Europeana, Wikidata and GRaSP data models are already complex to query, even for Semantic Web experts. Adding another layer, that of the voice or lenses, potentially further complicates this.

Personalisation can be supported by these different views. Developers of end-user applications should take care that such personalisation does not lead to 'filter bubbles'. Effective communication of the various views that exist on specific resources is crucial. One example is objects that are assigned to different categories depending on the cultural context. Many objects in ethnography museums are currently categorised based on older (colonial) classification schemes, that might not represent currently held views, and moreover, such views are likely to differ per culture. If such polyvocal information about the object is maintained, it can be shown to the users in digital interfaces as well as in physical locations.

What is needed are design guidelines and patterns for visualisation of polyvocal Knowledge as well as reusable tools and methods.

[3] https://newsreader-project.eu.

4 Discussion

Creating a polyvocal and contextualised Semantic Web will be a community effort driven by interdisciplinary teams. Whilst the Cultural AI consortium is embedded in different cultural heritage institutions in the Netherlands that represent different voices, we have to be aware of the danger that this remains a singular (Western/Northern) voice. To create globally relevant and culturally aware semantic web resources, diversity and inclusivity is key, and for this we can for example take cues from Black in AI[4] and Widening NLP.[5] The Semantic Web community is a diverse one, as our conference attendants, organising and programme committees, editorial boards and mailing lists show. Different voices are present, so let us make our resources, models and projects reflect that.

Acknowledgements. This paper builds upon the core concepts and research agenda of the Cultural AI Lab (https://www.cultural-ai.nl/), the authors would like to thank the co-creators of this research agenda: Antal van den Bosch, Laura Hollink, Martijn Kleppe, Johan Oomen, Jacco van Ossenbruggen, Stephan Raaijmakers, Saskia Scheltjens, Rosemarie van der Veen-Oei, and Lotte Wilms. We also want to thank our colleagues of the "Culturally Aware AI" project funded by the Dutch Research Council's Responsible use of AI programme, from the SABIO project, funded by the Dutch Digital Heritage Network, and from "Pressing Matter: Ownership, Value and the Question of Colonial Heritage in Museums", funded through the National Science Agenda of the Dutch Research Council.

References

1. Álvarez, G., Oeberst, A., Cress, U., Ferrari, L.: Linguistic evidence of in-group bias in English and Spanish Wikipedia articles about international conflicts. Discourse Context Media **35**, 100391 (2020)
2. Aroyo, L., Welty, C.: Truth is a lie: crowd truth and the seven myths of human annotation. AI Mag. **36**(1), 15–24 (2015)
3. Batchelor, C., et al.: Scientific lenses to support multiple views over linked Chemistry data. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 98–113. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_7
4. Beytía, P.: The positioning matters: estimating geographical bias in the multilingual record of biographies on Wikipedia. In: Companion Proceedings of the Web Conference 2020, pp. 806–810 (2020)
5. Bonatti, P.A., Decker, S., Polleres, A., Presutti, V.: Knowledge graphs: new directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In: Dagstuhl Reports, vol. 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2019)
6. Farda-Sarbas, M., Mueller-Birn, C.: Wikidata from a research perspective-a systematic mapping study of wikidata. arXiv preprint arXiv:1908.11153 (2019)

[4] https://blackinai2020.vercel.app/.
[5] http://www.winlp.org/.

7. Fokkens, A., Vossen, P., Rospocher, M., Hoekstra, R., van Hage, W.R.: GRaSP: grounded representation and source perspective. In: Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP 2017, pp. 19–25. INCOMA Inc., Varna, September 2017. https://doi.org/10.26615/978-954-452-040-3_003
8. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018)
9. Hendler, J., Berners-Lee, T.: From the semantic web to social machines: a research challenge for AI on the world wide web. Artif. Intell. **174**(2), 156–161 (2010)
10. Isaac, A.: Europeana data model primer (2013)
11. Jo, E.S., Gebru, T.: Lessons from archives: strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 306–316 (2020)
12. Konieczny, P., Klein, M.: Gender gap through time and space: a journey through Wikipedia biographies via the Wikidata human gender indicator. New Media Society **20**(12), 4608–4633 (2018)
13. Meyer, B.: Idolatry Beyond the Second Commandment: Conflicting Figurations and Sensations of the Unseen. Bloomsbury, London (2019)
14. Miles, R.S., Tzialli, A.O.E., Captain, E. (eds.): Inward Outward, Critical Archival Engagements with Sounds and Films of Coloniality - A Publication of the 2020 Inward Outward Symposium (2020). https://doi.org/10.18146/inout2020
15. National Museum for World Cultures: Words matter: An unfinished guide to word choice in the cultural sector. https://www.tropenmuseum.nl/nl/over-het-tropenmuseum/words-matter-publicatie (2018)
16. Ockeloen, N., et al.: BiographyNet: managing provenance at multiple levels and from different perspectives. In: LISC@ ISWC, pp. 59–71. Citeseer (2013)
17. Rospocher, M., et al.: Building event-centric knowledge graphs from news. Journal of Web Semantics **37**, 132–151 (2016)
18. Smith, C.E.: Decolonising the museum: The national museum of the American Indian in Washington, D.C. (2005)
19. Tiddi, I., d'Aquin, M., Motta, E.: Quantifying the bias in data links. In: Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E. (eds.) EKAW 2014. LNCS (LNAI), vol. 8876, pp. 531–546. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13704-9_40
20. Veltman, K.H.: Towards a semantic web for culture. J. Digital Inf. **4**(4) (2004)
21. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
22. Wei, L.: Crossing the ethnic boundaries: diversity promotion effort of museum professionals in the United Kingdom. In: Proceedings of the Papers, CAAS 7th Conference at TUFS, October 2016
23. Wolf, M.J., Miller, K.W., Grodzinsky, F.S.: Why we should have seen that coming: comments on Microsoft's Tay "Experiment", and wider implications. ORBIT J. **1**(2), 1–12 (2017)
24. Zhou, Y., Demidova, E., Cristea, A.I.: Who likes me more? Analysing entity-centric language-specific bias in multilingual Wikipedia. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 750–757 (2016)