Diversity and bias in DBpedia and Wikidata as a challenge for text-analysis tools

Bettina Berendt, TU Berlin / Weizenbaum Institute, Berlin / KU Leuven
Özgür Karadeniz, KU Leuven
Sercan Kıyak, KU Leuven
Stefan Mertens, KU Leuven
Leen d'Haenens, KU Leuven

Summary

Diversity Searcher is a tool originally developed to help analyse diversity in news media texts. It relies on automated content analysis and thus rests on prior assumptions and depends on certain design choices related to diversity. One such design choice is the external knowledge source(s) used. In this article, we discuss implications that these sources can have on the results of content analysis. We compare two data sources that Diversity Searcher has worked with – DBpedia and Wikidata – with respect to their ontological coverage and diversity, and describe implications for the resulting analyses of text corpora. We describe a case study of the relative over- or underrepresentation of Belgian political parties between 1990 and 2020. In particular, we found a staggering overrepresentation of the political right in the English-language DBpedia.

Zusammenfassung

Diversity Searcher ist ein Tool, das ursprünglich entwickelt wurde, um bei der Analyse von Diversität in Nachrichtentexten zu helfen. Es beruht auf einer automatisierten Inhaltsanalyse und stützt sich daher auf Annahmen und hängt von Designentscheidungen in Bezug auf Diversität ab. In diesem Artikel untersuchen wir die Auswirkungen davon, dass Ergebnisse einer automatisierten Inhaltsanalyse in der Regel von externen Wissensquellen abhängig sind. Wir vergleichen zwei Datenquellen, mit denen der Diversity Searcher arbeitet – DBpedia und Wikidata – im Hinblick auf ihre ontologische Abdeckung und Diversität und beschreiben die Auswirkungen auf die daraus resultierenden Analysen von Textkorpora. Wir beschreiben eine Fallstudie zur relativen Über- bzw. Unterrepräsentation belgischer politischer Parteien zwischen 1990 und 2020. Insbesondere stießen wir auf eine erstaunlich starke Überrepräsentation der politischen Rechten in der englischsprachigen DBpedia.

1. Introduction

Originally developed as a tool to help analyse diversity in news media texts, Diversity Searcher is now ready to be used by the general public and in the context of public libraries. The Diversity Searcher, presented in section 2, relies on automated content analysis and thus rests on prior assumptions and design choices.[1] To investigate these, in Section 3 we describe a case study that examines representational biases in the underlying data sources and what they could imply for the extent of diversity that the tool can identify. The case study investigates the relative over- or underrepresentation of Belgian political parties between 1990 and 2020 in the English- and Dutch-language DBpedias and in Wikidata. We found biases including a staggering overrepresentation of the political right in the English-language DBpedia. Section 4 concludes with an outlook on future work.

2. Diversity Searcher as a text analysis tool

Diversity Searcher (DS) is a semi-automated text analysis and knowledge enrichment tool designed to present information to the user about the degree of diversity in news media texts. The tool was developed in the interdisciplinary project DIAMOND (Diversity and Information Media: New Tools for a Multifaceted Public Debate)[2] to be used by media professionals and media consumers, and it is to be integrated with iCandid,[3] a research infrastructure project offering integrated access to several (social) media resources to researchers.

Over the course of the project, DS has been developed further to also be used in libraries. Public libraries in Europe are trusted community spaces with a long tradition of access to information. They connect citizens to critical issues such as public health, climate, societal challenges in education, social and digital inclusion and democracy. Public libraries also make an important contribution to digital skills development.[4] DS can be integrated as a service of public libraries to complement conversations with citizens, including hard-to-reach groups with whom libraries are connected, with online participation. DS contributes to critical news awareness among citizens by helping them to examine how diverse or biased the news is. Librarians can act as guides in promoting and learning to use DS as part of adult education and lifelong learning.

DS analyses an uploaded text or collection of texts. It extracts actors and offers additional information and interaction functionalities regarding the actors and their occurrence(s) in the text(s). These range

---

1   See for example Kitchin, Rob: The data revolution. Big data, open data, data infrastructures & their consequences, London 2014.
2   KU Leuven, Institut for Media Studies: Diamond, <https://soc.kuleuven.be/ims/diamond/>, last accessed 11.03.2023.
3   KU Leuven Libraries: A snapshot of LIBIS research infrastructures, <https://bib.kuleuven.be/english/libis/projects#icandid>, last accessed 11.03.2023.
4   Kish, Ilona; Thominet, Hannah; Zignani, Tiana: Libraries on the European agenda. How can the EU leverage the potential of public libraries to tackle European challenges? Berlin 2021. Online: <https://futurium.ec.europa.eu/sites/default/files/2021-05/ACTION%205%20-%20Libraries%20on%20the%20European%20Agenda%20-%20PL%202030_Final_0.pdf>, last accessed 11.03.2023.

---

https://doi.org/10.5282/o-bib/5894

from base statistics to a numerical evaluation of the complex notion of "diversity". In this section, we give an overview of the processing and user interface of DS.[5]

2.1 A quantitative measure of (actor) diversity in a text

One of our core assumptions is that diversity in media is a key precondition of a well-functioning democracy and a well-informed public opinion. Diversity is considered a crucial part of information quality, for which we draw on Stirling's (2007) influential proposal of a quantitative measure that draws on a meta-analysis of the literature on diversity from various disciplines.[6] The measure computes a score from three components: in our case, for a news story to truly cover a diversity of actors and viewpoints, it should contain several/many of them ("variety"), they should represent sufficiently different viewpoints ("disparity"), and their coverage should be balanced ("balance"). This score provides for a simple comparison of different texts that the user can supplant by inspecting and/or correcting the tool's information about the entities, their distribution, and their differences.

The components "variety" and "balance" are straightforward to compute once relevant entities have been identified (by counting them and considering their frequencies). However, "disparity" is often a hurdle when using this formula, as it is measured based on strong assumptions.[7] DS therefore focuses on a specific aspect to address diversity in media content: actor diversity. The choice of the concept and its operationalisation are sociologically inspired.[8] We study the diversity of socio-political actors, operationalised in terms of three types of actors: persons, organisations and geopolitical entities.

To operationalise an actor pair's disparity, we draw on external knowledge sources to (a) identify authors in the media texts and (b) retrieve relevant properties and relationships between them from the external source. Step (a) draws on the external tools DBpedia Spotlight and spaCy,[9] step (b) originally worked with the DBpedia ontology. Based on queries to the source, DS creates a contextual feature-based representation: features directly linked to this actor and enriched by features of their associated actors. For example, the tool attempts not only to retrieve a politician's personal features and their party, but also the party's properties such as ideology, as different parties (variety) may have intersecting ideologies, thus affecting the disparity between two politicians. Similarly, finding

5 More details can be found in Berendt, Bettina; Karadeniz, Özgür; Mertens, Stefan et al.: Fairness beyond "equal". The Diversity Searcher as a tool to detect and enhance the representation of socio-political actors in news media, in: Leskovec, Jure; Grobelnik, Marko; Najrok, Marc et al. (eds.): WWW '21: Companion proceedings of the Web Conference 2021, Ljubljana, Slovenia, April 19–23, 2021, New York 2021, pp. 202–212. Online: <https://doi.org/10.1145/3442442.3452303>.

6 Stirling, Andy: A general framework for analysing diversity in science, technology and society, in: Interface. Journal of The Royal Society 4 (15), 2007, pp. 707–719. Online: <https://doi.org/10.1098/rsif.2007.0213>.

7 Ranaivoson, Heritiana: Measuring cultural diversity with the Stirling model, in: New techniques and technologies for statistics, 2013, p. 10. Online: <https://ec.europa.eu/eurostat/cros/content/measuring-cultural-diversity-stirling-model-heritiana-ranaivoson_en>, last accessed 11.03.2023.

8 Bourdieu, Pierre: The social space and the genesis of groups, in: Theory and society 14 (6), 1985, pp. 723–744.

9 DBpedia Spotlight, <https://github.com/dbpedia-spotlight/dbpedia-spotlight-model>; spaCy, <https://spacy.io/>, both last accessed 11.03.2023. DBpedia Spotlight is a tool for annotating DBpedia resources found in text files. This enables named entity recognition (NER) and other information extraction (IE) tasks to be performed on target text files using the tool's API and the large semantic information contained in DBpedia. spaCy is a natural language processing tool; it can be used together with DBpedia Spotlight for better NER and IE results.

the country of an actor will result in a further query into properties of the country such as EU membership or type of government.

Table 1: Some socio-political features of Donald Trump, Hillary Clinton and Recep Tayyip Erdoğan

	Type	Subtype	Gender	Party	Party Ideology	Country
Donald Trump	Person	Politician	Male	Republican Party	Conservative	US
Hillary Clinton	Person	Politician	Female	Democratic Party	Liberal	US
Recep Tayyip Erdoğan	Person	Politician	Male	AKP	Conservative	Turkey

For example, Donald Trump and Hillary Clinton can be considered similar in that they are both American politicians. They are different in that they are associated with a conservative and a liberal party, respectively. When we add other actors to the context, the issue of disparity becomes more complex. Recep Tayyip Erdoğan is similar to Trump and Clinton as he is also a politician. He is more similar to Trump than to Clinton because of his gender, the ideology of his party, and the fact of holding an office as a president. However, he differs from both of them as he is based in a developing country, making Trump and Clinton more similar in this aspect. This example can be extended by adding other features, as well as other types of actors such as institutions, political parties and NGOs.

The disparity value between two actors is computed from the number of differences in these profiles, by dividing the number of common features by the number of all features. In the simplified example of table 1, this amounts to a disparity of (14 / 19 = 0.74) between Hillary Clinton and Donald Trump. The disparity information is then combined with the variety (the set {Hillary Clinton, Donald Trump, Recep Tayyip Erdoğan}) and balance (the number of mentions of these actors in the text) to compute the diversity of news text content.

2.2 User interface and interactivity

The DS interface gives feedback to the user about the socio-political actors in the text or corpus (figure 1), their features (in hover-info boxes with information drawn from DBpedia/Wikipedia), their disparities, and diversity (for example via histograms that visualise variety and balance, figure 2). It also allows users to correct mistakes resulting from the automated extraction and to add missing information (figures 3 and 4).

Figure 1: Colour-coded annotations of the three actor types in an excerpt of an article[10]
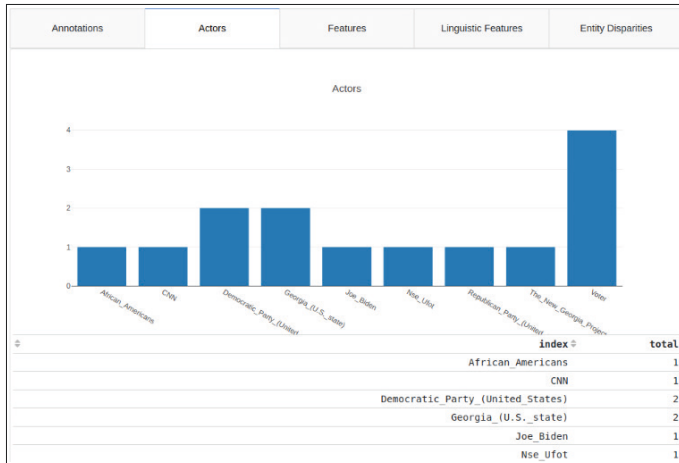
Figure 2: Actors found in the text, displayed on a histogram in the "Actors" tab

Figure 3: The interface tab for editing actor recognition patterns

10 Kenya, Evelyn: How Black voters lifted Georgia Democrats to Senate runoff victories, Guardian, 07.02.2021, <https://www.theguardian.com/us-news/2021/jan/07/georgia-senate-runoff-black-voters-stacey-abrams>, last accessed 11.03.2023.

Figure 4: The interface tab for pairwise actor disparities

The application includes a corpus upload function. Using this function, users can upload XML files exported from Belga Press[11] and batch process them, and export the results as an Excel file for further analysis.

DS allows media users to evaluate the news text and journalists to choose an alternative perspective when writing a news story. The question is no longer only whether enough diversity of content is present, but also whether the audience pays attention to the diversity of content present. The computerised measurement and display of diversity content thus help stimulate critical thinking around diversity.

3. The role of the underlying ontology for bias and diversity: Comparing DBpedia and Wikidata

Despite enriching textual content by further contextualising the actors, the integration of information from linked entities has the disadvantage of including errors and biases contained in these resources. For example, the data quality of the entry for "US" will also affect the data about "Hillary Clinton", "Joe Biden" and "Donald Trump" in the local knowledge base. Similarly, any systematic political bias related to the representation of political parties of certain ideologies will affect the tool's representation of the root politician, even if they are detected correctly.

---

11 Belga Press (belgapress.be) is an online press database covering a large number of international and Belgian news sources, where journalists and media researchers can search and export corpora. The specific source and therefore its XML format are relevant for our target users, but the schema could easily be made configurable.

We therefore began to question what consequences the choice of the English-language DBpedia versus (if applicable) the language of the country or region under investigation would have (in general, local DBpedias have better coverage of local actors), and to suspect that Wikidata, which operates with more stringent quality controls,[12] could be a better basis for DS.[13] Prior research indicates that the English DBpedia may offer fewer entities (especially in non-English language contexts) and have bias and quality issues. First inspections of actors in this ontology suggested an overrepresentation of the political right. Our observations comply with the existing comparative literature concerning these two information sources.[14] To better understand how this might be related to possible representational bias, we asked how "overpresentation in a source ontology" should be defined in the context of diversity search in the first place.

3.1 Method

We concentrated on the attributes "political party affiliation" and "political alignment of a party" and focused our analysis on "politicians from Belgium" to ensure sufficient domain knowledge for steering and interpreting the analysis. As a baseline, we used parties' share of the vote or number of seats at times T in the national parliament, and also looked at the Flemish parliaments.[15] We studied this in five-year intervals starting with 1990.[16]

Interpreting the ontologies as cultural memory, we asked what image the ontologies (in their current form) give of the representation of parties at these time points in the past. We queried the ontologies with SPARQL[17] and postprocessed the data to obtain lists of all represented politicians active at the studied time points T. We regard these as giving visibility to the party or parties they belonged to. Since individuals with multiple party affiliations across their career may be perceived differently (giving visibility to all these parties, or to only one of them, or re-centring attention on themselves and thus not giving visibility to any party), we derived a lower bound on the visibility of any given party (politicians who only ever belonged to that party) and an upper bound (politicians who belonged to that party and possibly also others).

---

12 Piscopo, Alessandro; Simperl, Elena: What we talk about when we talk about Wikidata quality: a literature survey, in: OpenSym '19: Proceedings of the 15th International Symposium on Open Collaboration, Skövde, Sweden, August 20–22, 2019, New York 2019, pp. 1–11. Online: <https://doi.org/10.1145/3306446.3340822>.

13 More details of the study can be found in Karadeniz, Özgür; Berendt, Bettina; Kıyak, Sercan; d'Haenens, Leen; Mertens, Stefan: Political representation bias in DBpedia and Wikidata as a challenge for downstream processing, in: arxiv.org, 2022 (CoRR abs/2301.00671). Online: <https://doi.org/10.48550/arXiv.2301.00671>.

14 Abián, D.; Guerra, F.; Martínez-Romanos, J. et al.: Wikidata and DBpedia. A comparative study, in: Szymański, Julian; Velegrakis, Yannis (eds.): Semantic keyword-based search on structured data sources, Cham 2018 (Lecture notes in computer science 10546), pp. 142–154. Online: <https://doi.org/10.1007/978-3-319-74497-1_14>.

15 Kamer van volksvertegenwoordigers, <https://nl.wikipedia.org/w/index.php?title=Kamer_van_volksvertegenwoordigers&oldid=61849646>; Vlaams Parlement, <https://nl.wikipedia.org/w/index.php?title=Vlaams_Parlement&oldid=61662474>, both last accessed 11.03.2023.

16 T = 1 January of 1990, 1996, 2000, 2005, 2011, 2015, 2020. The exceptions from the 5-year spacing were done to capture the effects of the general elections held in 1995 and 2010.

17 SPARQL 1.1 Query Language, <https://www.w3.org/TR/sparql11-query/>, last accessed 11.03.2023. SPARQL is the query language for RDF, a representation format used in graph databases, DBpedia, Wikidata, and other Linked data resources.

3.2 Results and interpretation

Figures 5 and 6 show examples of over- and underrepresentation: the 2015 data for the English-language DBpedia (figure 5 a), the Dutch-language DBpedia (figure 6 a) and Wikidata (figure 5 b). In each diagram, the parties are ordered from left to right according to their ideology (and then according to their acronym). For example, the figure shows that the 52 politicians associated with the right-wing N-VA constituted more than 60 % of the 2015 politicians in the English DBpedia (bold line), while the party had obtained only 22 % in the 2014 elections (thin line). The results confirm our informal observation of overrepresentation of right-wing parties (especially the N-VA) in the English-language DBpedia. Different biases seem to occur in the Dutch-language DBpedia: although on the whole comparatively similar to the baseline, this ontology seems to overrepresent the main centrist party (CD&V with close to 20 % of 355 represented politicians compared to 12 % of the national vote). Wikidata, in contrast, gives a rather accurate picture of party shares in the national parliament. Similar overrepresentation in media coverage have been identified in earlier international research, such as a centrist bias in media coverage of the UK elections of 2017[18] and the right-wing overrepresentation in social media, despite the cries of censorship in the United States.[19]

---
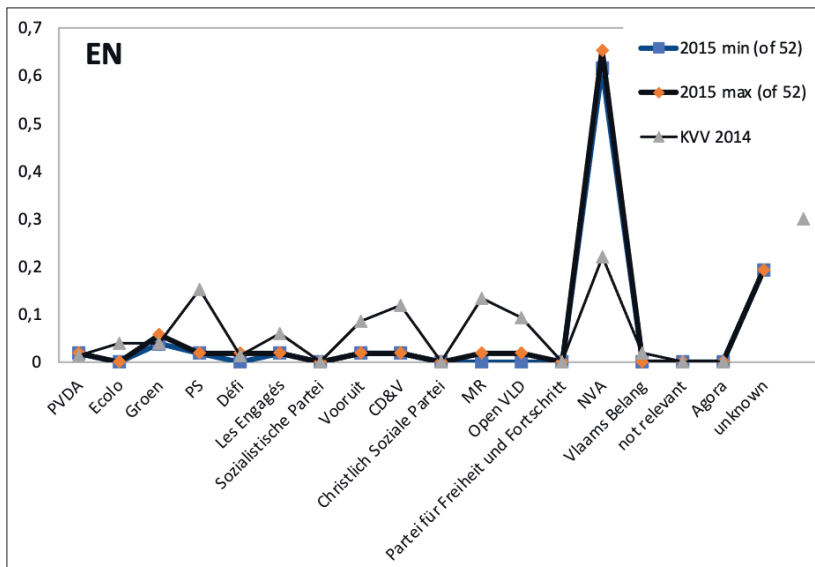
18 Deacon, David; Downey, John; Smith, David, Stanyer et al.: National news media coverage of the 2017 election. Centre for Research in Communication and Culture, Loughborough University, report 4: 5 May–7 June 2017. Online: <https://blog.lboro.ac.uk/crcc/wp-content/uploads/sites/23/2017/06/media-coverage-of-the-2017-general-election-campaign-report-4.pdf>, last accessed 11.03.2023.
19 Scott, Mark: Despite cries of censorship, conservatives dominate social media, Politico, 26.10.2020, <https://www.politico.com/news/2020/10/26/censorship-conservatives-social-media-432643>, last accessed 11.03.2023.

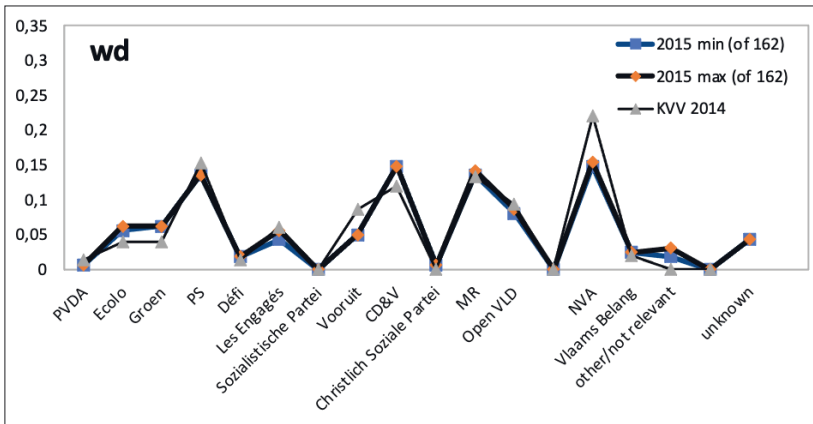*Figure 5: Visibility of parties in the English-language DBpedia (a), the Dutch-language DBpedia (middle) and Wikidata (b) for 2015*
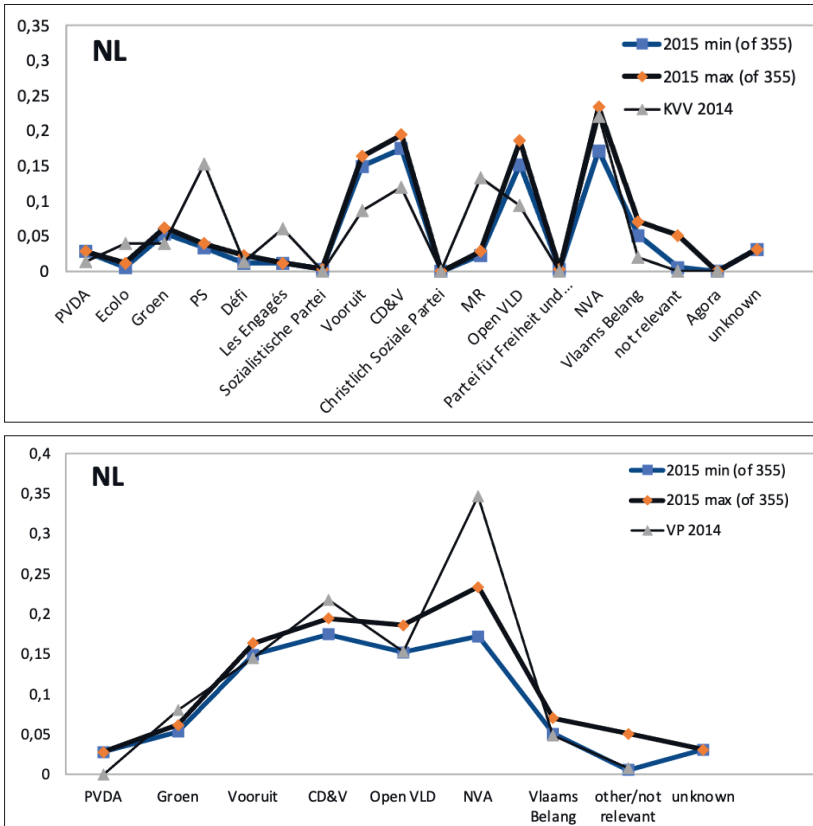

Figure 6: Visibility of (a) all parties and (b) the Flemish parties in the Dutch-language DBpedia for 2015

The seeming overrepresentation of the political centre in the Dutch-language DBpedia may be an artefact of the language. Figure 6 (a) also shows that the French-language Walloon parties (esp. Ecolo, PS, Les Engagés) are underrepresented. Figure 6 (b) maps the shares of seats in the Flemish parliament (which was first elected directly in 1995) as the baseline and is otherwise analogous to the other figures. The set of parties is a subset, since only the Flemish parties can be elected to the Flemish parliament (while the national parliament also contains representatives from the Walloon, Brussels, and German-Community parties). The contrast between figure 6 (a) and (b) suggests that the representation of the parties between extreme left and centre-right in the Dutch-language DBpedia mirrors these parties' shares in the regional parliament rather closely; while the right and extreme right tend to be underrepresented especially in the later years, when they were highly successful especially in the Flemish elections.

Figures and interactive figures for all years 1990 to 2020 are available;[20] these also illustrate that the overrepresentation of the political right in the English-language DBpedia increased over time.

3.3   Implications for the Diversity Searcher

Such biases matter in the context of DS because the tool can only compute diversity between entities that it recognises. Imagine (a) a well-balanced text with, say, several left-wing actors and several actors from one right-wing party, and (b) a text that only contains actors from the right-wing and extreme-right parties. Assuming that all else is equal between these texts, the "ground truth" would be that (a) is more diverse than (b). However, when sourced by an ontology with a bias to the right, the tool will only (or mostly) recognise actors from the one right-wing party in (a) and most actors from the two parties in (b), which will lead to a higher computed diversity score for text (b) and a lower score – maybe even zero – for text (a).

4. Conclusions and outlook

The shifts towards digital economies and smart cities resulted in a change in the societal role of public libraries.[21] The latter are now increasingly considered and expected to function as "hubs" or "platforms" that "link people to information, services and to each other".[22] Moreover, public libraries also have the task of bridging digital divides by facilitating access to digital resources and training the public in digital literacy. This renewed societal role of the public libraries within the digital economy makes their increasing usage of automated content analysis more problematic. Automated content analysis software has the potential to provide insight into large digital corpora quickly, making them valuable additions to research and critical thinking in a library setting. However, as illustrated by the case study in this article, such tools are prone to reproducing bias in the upstream components, even when they are designed to alleviate biases. Working with data requires knowledge and recognition

---

20 They can be found in Karadeniz et al.: Political representation bias, 2022, and at <http://www.berendt.de/DIAMOND/>, last accessed 11.03.2023.

21 Leorke, Dale; Wyatt, Danielle; McQuire, Scott: "More than just a library". Public libraries in the 'smart city', in: City, culture and society 15 (12), 2018, pp. 37–44. Online: <https://doi.org/10.1016/j.ccs.2018.05.002>.

22 Ibid., p. 40.

that data are not neutral and that they can be used to maintain an unequal status quo. Thus, data are part of the problem but can also be part of the solution if we maintain a number of principles: one is to situate the data in a historical and social context. Library scientists can help create this awareness and a critical and informed attitude at the user end.

These numerical considerations can be exacerbated by the inability of automated tools to recognise and understand context. This is another reason to treat the numerical and categorical results as a starting point for deeper text analysis and involve users – whether they be individual citizens, researchers, organisations representing a particular target group or journalists looking for a new angle for a news story – in sense-making by means of the interactive DS interface. For future work, a two-pronged strategy is recommended: (a) identifying and using the best-suited ontology for a given task, and at the same time (b) making its properties and shortcomings transparent to users so as to keep users aware of challenges associated with the (and any) dataset.

Acknowledgements

References

– Abián, D.; Guerra, F.; Martínez-Romanos, J. et al.: Wikidata and DBpedia. A comparative study, in: Szymański, Julian; Velegrakis, Yannis (eds.): Semantic keyword-based search on structured data sources, Cham 2018 (Lecture notes in computer science 10546), pp. 142–154. Online: <https://doi.org/10.1007/978-3-319-74497-1_14>.

– Berendt, Bettina; Karadeniz, Özgür; Mertens, Stefan et al.: Fairness beyond "equal". The Diversity Searcher as a tool to detect and enhance the representation of socio-political actors in news media, in: Leskovec, Jure; Grobelnik, Marko; Najrok, Marc et al. (eds.): WWW '21: Companion proceedings of the Web Conference 2021, Ljubljana, Slovenia, April 19–23, 2021, New York 2021, pp. 202–212. Online: <https://doi.org/10.1145/3442442.3452303>.

– Bourdieu, Pierre: The social space and the genesis of groups, in: Theory and society 14 (6), 1985, pp. 723–744.

– Deacon, David; Downey, John; Smith, David, Stanyer et al.: National news media coverage of the 2017 election. Centre for Research in Communication and Culture, Loughborough University, report 4: 5 May – 7 June 2017. Online: <https://blog.lboro.ac.uk/crcc/wp-content/uploads/sites/23/2017/06/media-coverage-of-the-2017-general-election-campaign-report-4.pdf>, last accessed 11.03.2023.

– Karadeniz, Özgür; Berendt, Bettina; Kıyak, Sercan; d'Haenens, Leen; Mertens, Stefan: Political representation bias in DBpedia and Wikidata as a challenge for downstream processing, in: arxiv.org, 2022 (CoRR abs/2301.00671). Online: <https://doi.org/10.48550/arXiv.2301.00671>.

– Kish, Ilona; Thominet, Hannah; Zignani, Tiana: Libraries on the European agenda. How can the EU leverage the potential of public libraries to tackle European challenges? Berlin 2021. Online: <https://futurium.ec.europa.eu/sites/default/files/2021-05/ACTION%205%20-%20Libraries%20on%20the%20European%20Agenda%20-%20PL%202030_Final_0.pdf>, last accessed 11.03.2023.

– Kitchin, Rob: The data revolution. Big data, open data, data infrastructures & their consequences, London 2014.

– Leorke, Dale; Wyatt, Danielle; McQuire, Scott: "More than just a library". Public libraries in the 'smart city', in: City, culture and society 15 (12), 2018, pp. 37–44. Online: <https://doi.org/10.1016/j.ccs.2018.05.002>.

– Piscopo, Alessandro; Simperl, Elena: What we talk about when we talk about Wikidata quality: a literature survey, in: OpenSym '19: Proceedings of the 15th International Symposium on Open Collaboration, Skövde, Sweden, August 20–22, 2019, New York 2019, pp. 1–11. Online: <https://doi.org/10.1145/3306446.3340822>.

– Ranaivoson, Heritiana: Measuring cultural diversity with the Stirling model, in: New techniques and technologies for statistics, 2013. Online: <https://ec.europa.eu/eurostat/cros/content/measuring-cultural-diversity-stirling-model-heritiana-ranaivoson_en>, last accessed 11.03.2023.

– Scott, Mark: Despite cries of censorship, conservatives dominate social media, Politico, 26.10.2020, <https://www.politico.com/news/2020/10/26/censorship-conservatives-social-media-432643>, last accessed 11.03.2023.

– Stirling, Andy: A general framework for analysing diversity in science, technology and society, in: Interface. Journal of The Royal Society 4 (15), 2007, pp. 707–719. Online: <https://doi.org/10.1098/rsif.2007.0213>.