

Collaborative Visualizations for Wikipedia Critique and Activism

**Stefano De Sabbata, Kathryn Eccles,
Scott Hale, Ralph Straumann**

Oxford Internet Institute,
University of Oxford
1 St Giles', Oxford, England
{stefano.desabbata, kathryn.eccles, scott.hale, ralph.straumann}@oii.ox.ac.uk

Arzu Çöltekin

Department of Geography
University of Zurich
Winterthurerstr. 190, Zurich, Switzerland
arzu.coltekin@geo.uzh.ch

Abstract

Wikipedia is one of the largest platforms based on the concept of asynchronous, distributed, collaborative work. A systematic collaborative exploration and assessment of Wikipedia content and coverage is however still largely missing. On the one hand editors routinely perform quality and coverage control of individual articles, while on the other hand academic research on Wikipedia is mostly focused on global issues, and only sporadically on local assessment. In this paper, we argue that collaborative visualizations have the potential to fill this gap, affording editors to collaboratively explore and analyse patterns in Wikipedia content, at different scales. We illustrate how a collaborative visualization service can be an effective tool for editors to create, edit, and discuss public visualizations of Wikipedia data. Combined with the large Wikipedia user-base, and its diverse local knowledge, this could result in a large-scale collection of evidence for critique and activism, and the potential to enhance the quantity and quality of Wikipedia content.

Introduction

Wikipedia articles are a prime example of asynchronous, distributed collaboration on an internet scale. Editors from all the connected part of the world can gather to collaborate on a single shared document, without being in the same physical place or working on that document at the same time. For example, at the time of this writing, the term “city of dreaming spires” used by poet Matthew Arnold to describe Oxford is still part of Oxford’s Wikipedia article¹, as it was in the first version of the article, created in May 2001 by the editor Mjausson². This has given every Wikipedia user the opportunity to reflect on that first descrip-

tion of Oxford, to discuss it in the Talk page, edit, and move snippets of text in the different sections of the article. This is the essence of Wikipedia as a tool for asynchronous, distributed, collaborative sensemaking.

While this asynchronous, distributed collaboration is at the very heart of Wikipedia, the analysis of its content and coverage is still largely composed of many separate, individual efforts. Several research projects have focused on the analysis of Wikipedia content, from cross-language comparison (e.g., Hale, 2015; Hecht and Gergle 2010b, Pfeil et al., 2006), to geographic analysis (e.g., Hecht and Gergle 2009; Graham et al., 2014), to the analysis of controversial topics (e.g., Yasseri et al., 2014). Since systematic collaborative exploration and assessment of Wikipedia content and coverage is still largely missing, the platforms mostly relies on ad-hoc assessments by users for decisions about new content creation; i.e., users compare and analyse article contents individually and then may decide to contribute additional content or amend existing content.

What if the editors (and readers) were able to visualize the content (e.g., word frequencies), structure (e.g., which articles are linked to which other articles), or statistics (e.g., how many people write about a particular topic, where are these people, how many visitors were on this page and when)? What if these visualizations were also collaborative, so that other editors could also edit them? We contend that a tool allowing the broad Wikipedia community to collaboratively explore and analyse Wikipedia at different scales and collect evidence for critique and activism has large potential to enhance the quantity and quality of Wikipedia content. In other words; in this this paper, we argue that collaborative visualizations (Pea, 1993; Isenberg et al., 2011) can afford this function by giving groups of people the opportunity to reflect, discuss, and edit a common visual representation of Wikipedia content (see e.g., Figures 1 and 3, discussed below), in the

¹ en.wikipedia.org/w/index.php?oldid=648076589

² en.wikipedia.org/w/index.php?oldid=271629

same way that editors can discuss and collaboratively edit the content of individual Wikipedia articles today. Therefore, collaborative visualizations would support the process of asynchronous, distributed, collaborative sensemaking of entire parts of Wikipedia, in addition to the sensemaking that already occurs on the level of single articles.

Visual analytics

The term *visual analytics* was coined a decade ago by Thomas and Cook (2005) to refer to the “science of human analytical reasoning facilitated by interactive visualizations” (ibid: p.28). Visual analytics can be considered a direct descendant of the concept of *exploratory data analysis* proposed by Tukey (1977). The fundamental idea is to combine the computer capabilities in automatic analysis and the human capabilities in visual pattern recognition. The aim, therefore, is to address a particular class of problems, which are both too complex or ill-defined to be fully automatized (i.e., too hard for a computer), and involve datasets too large and diverse to be presented in a static visualization for humans to analyse (Keim et al, 2008; 2010). It comes as no surprise that several visual analytics software programs are being developed in the recent wave of ‘big data’ (Zhang et al, 2012), since they offer data exploration functionality and dashboards for making sense of large datasets. Within the domain of visual analytics, the field of geographic information science is devoting particular attention to the development of geo-visual analytics methods, that can account for the spatial and temporal components of data, and the inherent challenges that those dimensions pose in terms of both analysis and visualization methods (Andrienko et al., 2010).

This paper contends that the analysis of Wikipedia content falls into the category of problems that visual analytics has been developed to tackle. The adequacy, correctness, completeness, and currency of Wikipedia articles and categories is a complex and ill-defined problem that could hardly be fully automatized. Moreover, information visualization methods have long been used by researchers to analyse and investigate Wikipedia contents, edits, editors and their geographies, as well as the differences between different editions. Methods employed range from pie charts (Bao et al., 2012) to maps (Yasseri et al., 2014) and from density plots to network diagrams (Hale, 2014).

Nonetheless, while ad-hoc processes and tools have so far been successfully used by researchers, such methods might not be suitable for Wikipedia contributors, who may lack the tools, time, or skills to perform the technical processes needed to create such visualizations. These factors serve as barriers limiting the number of people who have access to such analyses. In turn, not only the scope but especially the scale of such analyses is diminished. “Local” scale analyses might be of great interest and relevance to particular communities, groups, or individuals but might

not be chosen as a research direction by professional scientists with a global audience in mind, or simply lacking local knowledge to do these subjects justice.

This paper further contends that collaborative visualizations, including collaborative geo-visualizations, can be a useful means to enable the analysis of Wikipedia content at scale. That is, a collaborative visualization service would provide the Wikipedia community with a tool to perform analyses of Wikipedia content, in a manner which would be consistent with the principles and practices of Wikipedia. Users would be able to collaborate in investigating the structures and content of the platform, propose hypotheses, collect evidence, formulate critiques, and promote actions, such as new content creation and revision.

Collaborative visualization

One definition of collaborative visualization is “the shared use of computer-supported, (interactive,) visual representations of data by more than one person with the common goal of contribution to joint information processing activities” (Isenberg et al., 2011, p.312) – which covers its most important aspects. The key distinction between collaborative visualization and other visualization environments is the possibility of different users asynchronously accessing, commenting, and editing visualizations created by other users. When specifically applied to visual analytics services, this approach is also referred to as collaborative visual analytics (Heer and Agrawala, 2008).

Collaborative visualization services (e.g., Heer et al., 2007; Viegas et al., 2007) are founded on the same principles as user-generated content websites like Wikipedia. Thus, both offer very similar functionalities. A user of a collaborative visualization service is able to create a new visualization, which is visible and editable by any other user of the same service. Users can edit visualizations, leave comments, and graphically annotate them, while the system records a changelog of each stage in the evolution of the visualization thereby ensuring complete lineage information. Heer et al. (2007) discuss how each of these functionalities has been used in a pilot study of the *sense.us* website. They clearly illustrate how the comment section is key to the ongoing process of sensemaking, as different users observe and point out patterns, ask questions, and suggest interpretations of the visualized data – an analogous role is performed by *Talk* pages in Wikipedia.

Similar concepts have been developed within the field of geographic information science (Brewer et al., 2000; Brodlić et al., 2005). These take the forms of *participatory geographic information systems* (GIS) or *public participation GIS* (Abbot et al., 1998; Dunn, 2007) and *volunteered geographic information* (VGI; Goodchild, 2007). These developments are also partially rooted in critical cartography (Crampton and Krygier, 2006), and critical geographic

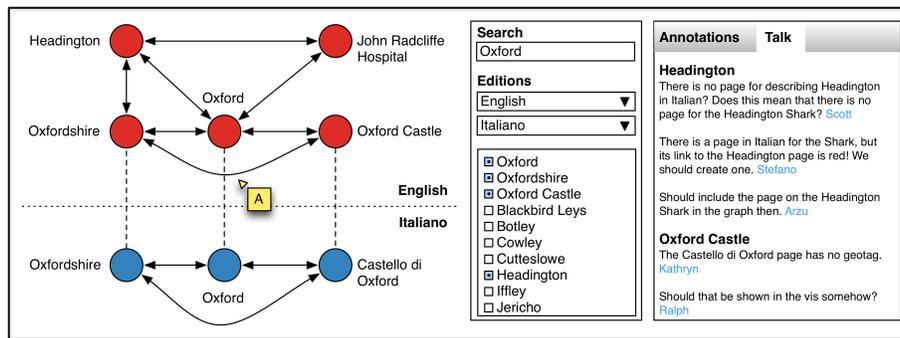


Figure 1. Illustrative example of usage of a network diagram to illustrate interlanguage links on Wikipedia.

information systems (Harvey et al., 2005), and thus in the on-going discussion within geography concerning the concepts of space and place (Sui and Goodchild, 2011).

The following section presents three scenarios that illustrate how collaborative visualizations, visual analytics, and geo-visual analytics methods could be applied to Wikipedia as object of analysis.

Collaborative visualizations for Wikipedia

A long-standing challenge for Wikipedia has been that most of its content (over 74 percent of all concepts) is written in only one language (Hecht and Gergle 2010b). Furthermore, even when users edit multiple language editions of Wikipedia, they are much more likely to edit articles in a second language that have a corresponding article in their first languages (Hale, 2015). So-called interlanguage links are a valuable resource to analyse what articles exist in certain language editions but not others. Interlanguage links connect articles about the same concept in different languages. For example, the article on Oxford in English is linked to the article on *Оксфорд* in Japanese.

Interlanguage links were previously maintained separately in each language edition of Wikipedia through a mix of human and machine processes. They did not necessarily align perfectly between different language editions. In 2013, these separate interlanguage links were replaced with a global, conflict-free, centrally stored and edited repository in WikiData³. WikiData provides a knowledge base that is closely coupled with Wikipedia, making it a good possible source of information for collaborative visualization applications in general.

A collaborative visualization of the interlanguage link data stored in WikiData could allow Wikipedia editors to understand what concepts are covered in other languages beyond the languages they edit in most frequently. This could help both multilingual readers to discover additional content and multilingual editors to write about some of this

content in their primary languages thereby expanding the coverage of each language edition of Wikipedia.

Figure 1 illustrates how a network diagram could be used to explore how different entities related to Oxford are represented in English and Italian Wikipedia.

Besides exploiting data on relative coverage in different languages through interlanguage links to enhance Wikipedia, WikiData could also be used to monitor specific aspects of coverage such as the gender of biography article subjects or the representation of different locations.

The Wikimedia Lab DB offers another crucial source of data for a collaborative visualizations service, as it stores the complete structure of Wikipedia and other wikis in an SQL format (i.e., a standard relational database format). These databases provide a variety of information about single pages as well as their metadata. For instance, from data accessible through Wikimedia Lab DB (or related services, such as Quarry⁴ or the MediaWiki web API⁵), a hierarchical matrix plot (see Figure 2) could be created for comparing the coverage of a category in two different language editions. Each cell in such a plot would show the difference in, e.g., the number of pages or the page lengths contained in a category and its subcategories (the latter two structured using marginal dendrograms in Figure 2).

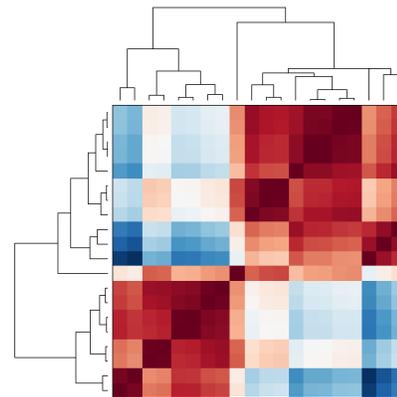


Figure 2. Illustrative example of hierarchical matrix plot (generated using random data).

³ Launched in 2012, www.wikidata.org

⁴ Launched in 2014, quarry.wmflabs.org

⁵ www.mediawiki.org/wiki/API:Main_page

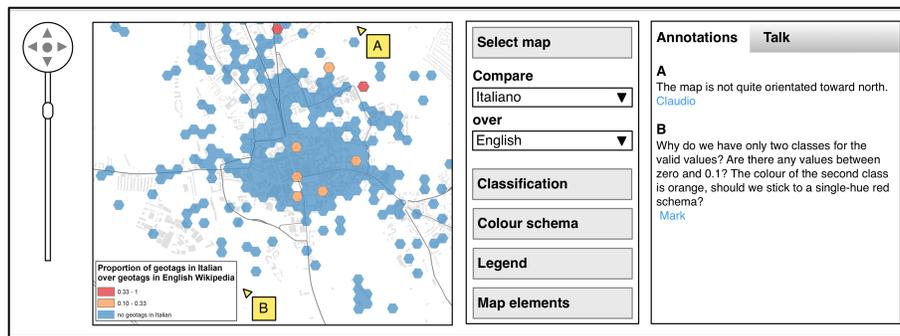


Figure 3. Illustrative example of usage of a density map to illustrate presence and absence of Wikipedia geotags.

As another branch of exploration, collaborative visualization of the data in WikiData and Wikimedia Lab DB could also be used to analyse the geographic biases present in Wikipedia (e.g., Graham et al., 2014). Many Wikipedia articles about places and events have geolocation information attached to them (a.k.a., geo-tags). As such, it is possible to map the coverage of Wikipedia as a whole as well as the coverage of any particular language edition. Figure 3 illustrates how a map could be used to explore the presence and absence of geo-tags in Oxford, comparing English and Italian Wikipedia.

Additionally, one can map the locations of contributors using either IP address, geocoding, or user profile geocoding⁶. Ongoing work has also started to geolocate the third-party sources (e.g., newspaper articles, websites, etc.) cited in each language edition (Sen et al., 2015) enabling a third layer of geographic coverage to be visualized and collaboratively analysed. Plotting any of these three layers of geographic information in the form of a dot map or a density map could reveal interesting patterns, and possibly coverage gaps. Such visualizations are especially meaningful and useful to Wikipedia users holding deep local knowledge of a certain geographical region (e.g., a valley, or a village) and may motivate their future contribution efforts.

More generally, we believe that opening up shortcomings of Wikipedia content and structure to reflection and discussion by rendering them explicit through collaborative visualizations has great potential for alleviating the known biases such as geography, gender or status present in all user-generated content platforms. Similarly, displaying the strengths of Wikipedia may allow inferring which content may be better quality controlled than others, or potentially lead to channelling the content creators' energy and efforts to less attended topics.

Furthermore, we envision that collaborative visualization tools could expand to encompass contributor statistics and user retention metrics in the future. Such data is not currently available in WikiData, but efforts are underway to make this data more easily accessible. Analysis and vis-

ualizations of such data would be potentially very valuable for promoting diversity among contributors and thus another vector for improving the quality of Wikipedia as a community and platform.

Challenges and research agenda

In this paper, we have illustrated how a collaborative visualization service would enable users to analyse Wikipedia content using visual analytics methods to investigate diverse aspects of the platform in a collaborative and asynchronous manner. Such activities would then ideally result in new content creation or in amendments of existing content. Building a collaborative visualization service on top of a user-generated content platform (to expand and improve the platform's coverage through collaborative introspection and discussion) is not restricted to Wikipedia, but could also benefit other crowdsourcing and open data initiatives. However, this new perspective also poses some questions and opens up new challenges in a number of research areas related to technology, design and social sciences.

First, a number of technical challenges need to be addressed in order to implement a service allowing collaborative visualization and analyses as discussed above. In the case of Wikipedia, the WikiData project and the Wikimedia Lab DB currently seem the most promising foundations for such a service, providing the necessary underlying input data. Currently, vector-based interactive visualization tools represent the state-of-the-art for visual analytics (possibly using WebGL for complex visualizations (see e.g. Garaizar et al., 2012)). Custom-made tools could be built for collaborative sensemaking or adapted from existing projects such as RAW⁷ (see Uboldi and Caviglia, 2015).

Second, information visualization design challenges need to be carefully considered to decide which type of graphs and maps should be made available for which kind of data. In order to reach a broad user base among Wikipedia editors, the overarching emphasis in service and visualization development needs to be put on ease of use for con-

⁶ E.g., cii.oii.ox.ac.uk/visualising-the-locality-of-participation-and-voice-on-wikipedia

⁷ github.com/densitydesign/raw

structuring, editing, annotating and discussing visualizations, while the visualization designs should be guided by the cognitive and perceptual principles. The interface design should focus on learnability, consistency with Wikipedia and its modalities of interaction, and support users in their visualization process by offering informed choices and annotations leading them to good design choices.

Furthermore, assuming that a collaborative visualization service for Wikipedia has been developed, deployed, and is being actively used, new opportunities for development and testing of new ideas and methods in the field of *computer-supported collaborative work* will arise. A critical perspective from the digital humanities community could lead to significant improvements of the service, resulting from rich historical understandings of the construction of knowledge, and experience of using such mixed methods (visualizations alongside discussion) for collaborative sensemaking. Such a service would also be a valuable tool for digital humanities research, allowing for multilayered analyses of articles on, for example, historical events, literary texts, and historiography. The open-source approach at the core of Wikipedia will provide researchers in the social sciences with a great source of data on collective behaviour on the internet, and the use of data and visualization for decision-making, critique, and activism.

Finally, a distributed, large-scale analysis of Wikipedia, which developed to one of the pivotal sources of information on the internet, will shed light on the role of digital mediation in content production, reproduction, and its representativeness.

References

- Abbot, J.; Chambers, R.; Dunn, C.; Harris, T.; Merode, E. D.; Porter, G.; Townsend, J.; and Weiner, D. 1998. Participatory GIS: opportunity or oxymoron. *PLA notes.*, 33: 27-33.
- Andrienko, G.; Andrienko, N.; Demsar, U.; Dransch, D.; Dykes, J.; Fabrikant, S. I.; Jern, M.; Kraak, M.J.; Schumann, H.; and Tominski, C. 2010. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10): 1577-1600.
- Bao, P.; Hecht, B.; Carton, S.; Quaderi, M.; Horn, M.; and Gergle, D. 2012. Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1075-1084). ACM.
- Brewer, I.; MacEachren, A. M.; Abdo, H.; Gundrum, J.; and Otto, G. 2000. Collaborative geographic visualization: Enabling shared understanding of environmental processes. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on* (pp. 137-141). IEEE.
- Brodlie, K.; Fairbairn, D.; Kemp, Z.; and Schroeder, M. 2005. Connecting people, data and resources—distributed geovisualization. *Exploring Geovisualisation*: 425-443.
- Crampton, J. W., and Krygier, J. 2006. An introduction to critical cartography. *ACME: An International e-Journal for Critical Geographies*, 4(1): 11-33.
- Dunn, C. E. 2007. Participatory GIS—a people's GIS?. *Progress in Human Geography*, 31(5): 616-637.
- Garaizar P.; Vadillo, M.A.; and Lopez-de-Ipina, D. 2012. Benefits and Pitfalls of Using HTML5 APIs for Online Experiments and Simulations. *International Journal of Online Engineering*, 8: 20-25.
- Goodchild, M. F. 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4): 211-221.
- Graham, M.; Hogan, B.; Straumann, R. K.; and Medhat, A. 2014. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4): 746-764.
- Hale, S. A. 2015. Cross-language Wikipedia editing of Okinawa, Japan. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '15. ACM.
- Hale, S. A. 2014. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM Conference on Web Science, WebSci '14*, (pp. 99–108). ACM.
- Harvey, F.; Kwan, M. P.; and Pavlovskaya, M. 2005. Introduction: critical GIS. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 40(4): 1-4.
- Hecht, B., and Gergle, D. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies, CandT '09*, (pp. 11–20). ACM.
- Hecht, B., and Gergle, D. 2010a. On the “localness” of user-generated content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW '10*, (pp. 229–232). ACM.
- Hecht, B., and Gergle, D. 2010b. The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI '10*, (pp. 291–300). ACM.
- Heer, J., and Agrawala, M. 2008. Design considerations for collaborative visual analytics. *Information visualization*, 7(1): 49-62.
- Heer, J.; Viégas, F. B.; and Wattenberg, M. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1029-1038). ACM.
- Isenberg, P.; Elmqvist, N.; Scholtz, J.; Cernea, D.; Ma, K. L.; and Hagen, H. 2011. Collaborative visualization: definition, challenges, and research agenda. *Information Visualization*, 10(4): 310-326.
- Keim, D.; Andrienko, G.; Fekete, J. D.; Görg, C.; Kohlhammer, J.; and Melançon, G. 2008. Visual Analytics: Definition, Process, and Challenges. In *Information Visualization* (pp. 154-175). Springer-Verlag.
- Keim, D. A.; Mansmann, F.; and Thomas, J. 2010. Visual analytics: how much visualization and how much analytics?. *ACM SIGKDD Explorations Newsletter*, 11(2): 5-8.
- Pea, R. D. 1993. The collaborative visualization project. *Communications of the ACM*, 36(5): 60-63.
- Pfeil, U.; Zaphiris, P.; and Ang, C. S. 2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1): 88–113.
- Sen, S.; Ford, H.; Musicant, D.; Graham, M.; Keyes, O. S.; and Hecht, B. 2015. Barriers to the localness of volunteered geo-

graphic information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2015*. ACM.

Sui, D., and Goodchild, M. 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11): 1737-1748.

Thomas, J. J., and Cook, K. A. 2005. *Illuminating the path: the research and development agenda for visual analytics*. IEEE Computer Society.

Tukey, J. W. 1977. Exploratory data analysis. *Reading, Ma*, 231: 32.

Uboldi, G., and Caviglia, G. 2015. Information Visualizations and Interfaces in the Humanities. In *New Challenges for Data Design* (pp. 207-218). Springer London.

Viegas, F. B.; Wattenberg, M.; Van Ham, F.; Kriss, J.; and McKeon, M. 2007. Manyeyes: a site for visualization at internet scale. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6): 1121-1128.

Yasseri, T.; Spoerri, A.; Graham, M.; and Kertész, J. 2014. The most controversial topics in Wikipedia: A multilingual and geographical analysis. In *Global Wikipedia: International and cross-cultural issues in online collaboration*, Scarecrow Press.

Stoffel, A.; Behrisch, M.; Mittelstadt, S.; Schreck, T.; Pompl, R.; Weber, S.; Last, H.; Keim, D. 2012. Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (pp. 173-182). IEEE.