An Open Toolkit for Small-Scale Data Visualizations

Andrea Pronzati
Politecnico di Milano
andrea.pronzati@polimi.it
ORCID 0000-0002-9346-6672

Michele Mauri
Politecnico di Milano
michele.mauri@polimi.it
ORCID 0000-0003-1189-9624

Abstract
Wikipedia represents an open and collaborative encyclopedic environment, and is one of the main sources of information today. The information is encoded, among other formats, in graphics and data visualizations. The source of such pieces of information is various — some are taken from governmental sources, or reports published under Creative Commons licenses, others are created and uploaded by the platform users.

Given the current UI of the Wikipedia platform, however, most of these contents are relegated to small thumbnails, making them difficult to read.

In the research presented here we propose an open toolkit for the design of data visualizations in small spaces — such as thumbnails — conceived for the project's contributors.

By engaging with the design of visualizations for the platform and contributing to the project, it was possible to identify good practices that were then synthesized in a toolkit openly accessible to all contributors.

Keywords
Micro visualizations
Graphic design
Open knowledge
Wikipedia
Wikimedia Commons

Introduction

Data visualizations are often seen as complex systems, interweaving information in digital or printed artifacts. In recent years we have seen a rise in this approach, which has produced well-articulated representations that, in the most successful examples, make complex phenomena understandable.

Despite the increasing density and complexity of visualizations, we are facing the opposite problem today: making visualizations in small spaces, for example on small monitors or portions of the screen. A clear example is the use of visualizations on Wikipedia, the well-known collaborative encyclopedia: even if they are very complex, in most cases they are relegated to a thumbnail about 220 pixels wide, making them almost always difficult to read.

In the research presented here, we have framed the problem of “micro visualizations”. By micro visualization we mean data visualization designed for a small space, taking spatial constraints as a design requirement. The notion of spatial constraint will be investigated more precisely and in-depth in subsequent sections.

The concept of micro visualization, however, does not live alone but fits into a broader and more varied design ecosystem. Indeed, in the field of communication design the medium, and its constraints, have a significant impact on the design process, and not just in relation to data visualizations. While several handbooks are available on the topic of data visualization, on typography or on layout, solutions for small spaces are scattered among sources in an array of methods and design processes and there is little guidance about them (Kim et al., 2021).

Wikimedia Projects as Test Fields

Wikipedia is one of the first results for most of the queries on U.S.-based search engines such as Google or Bing (Vincent & Hecht, 2021). It provides vast access to information, both as texts and images. At the moment of writing, however, the images are presented as a second level of information, relegated to small thumbnails that often make it impossible to properly read the information they contain.

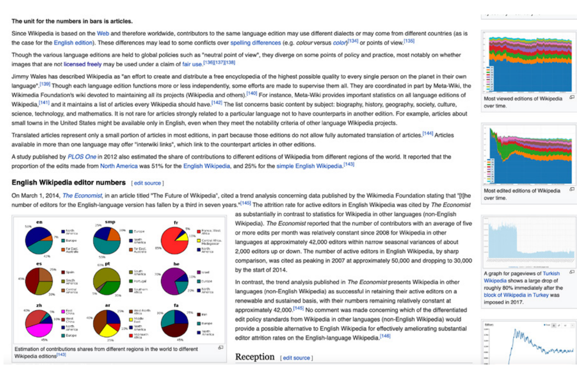

Fig. 1
An example of visualizations in a Wikipedia article (specifically, the article about Wikipedia itself). As one can see, it’s quite difficult to read their content.

To be fair, Wikipedia is just the most visible part of a wider and complex ecosystem known as “Wikimedia projects”, which consists of many projects: Wikipedia for encyclopedic content, Wikisource for book sources, Wikiquote for quotes, and Wikimedia Commons for images,

diid No. 76 — 2022
Doi: 10.30682/diid7622o

to mention just a few. The latter, Wikimedia Commons, is one of the largest media repositories of free-use images, sounds and other media. Like the better-known Wikipedia, anyone can upload images or illustrations free of copyright, edit them, discuss possible improvements, and embed them in the other projects.

Thanks to its inclusive editing policies, it is the perfect place to test, refine and collect feedback on the design of small visualizations. Far from using it as a mere experimentation case, the goal of this research was to inquire into the topic of "micro visualizations", while enriching the platform with well-designed data visualizations.

Wikimedia Projects and Their Communities

One of the factors that make the Wikimedia projects unsurpassable players in the open knowledge panorama is the kind of communities they are able to attract, and the mechanisms and dynamics that users identify to regulate them. Almost anyone with an Internet connection has the possibility to intervene and contribute to one of the projects.

Each project has a separate user base: some users work across different projects (e.g. on Wikipedia and on Wikimedia Commons), while others are focused on a specific one (Yun et al., 2019). Each community, over time, has defined guidelines and internal hierarchies, creating a delicate but solid system of self-regulation, which can intervene in case of acts of vandalism.

This incredible and ingenious cultural machine has attracted different kinds of studies and critiques over time, on the main project Wikipedia in particular. One of the primary elements that emerged is the speed with which Wikipedia assumed a founding role in contemporary culture and knowledge archetypes (Lanier, 2006). Lanier harshly criticizes the concept of New Online Collectivism, raising awareness of how a piece of knowledge based on absolute collectivism, on the idea that the group has an all-knowing prerogative, is wrong in and of itself, if not treated with the appropriate standards.

It is precisely on the concept of collectivism, however, that the platform is based, and the idea of community lies at the core of the online encyclopedia. The possibility that anyone, registered to the platform or not, can make modifications, makes it possible to collect many and diverse opinions and edits. In spite of its stated goal of guaranteeing a "neutral point of view" (Wikipedia: Neutral Point of View, 2021), this widespread participation allows people to mediate a shared vision of the topic under consideration.

Wikipedia & Images

Over time, a research skeleton has been structured around the Open Knowledge landscape, and in particular on Wikipedia, to analyze its potential and associated criticalities (Giles, 2005; Kittur et al., 2007; Stvilia et al., 2008).

As these studies proliferate, the focus often regards the written content, analyzing its reliability, production model, and community behaviour. The situation for images is different. The last study on the topic (Viegas, 2007) highlighted that only 9% of the most edited articles have actual visual content. At the time of writing, the authors have hypothesized that the main reason for this can be found in the Wiki technology and system, which were created and structured essentially to work with text, making the processing of images more difficult.

Pictures are however powerful containers of information. Among images, visualizations play a key role, because they have the incredible power to remain imprinted longer in the human mind, as Medina points out (Medina, 2008). Images, diagrams, and visualizations are considered such powerful knowledge tools (Mauri & Ciuccarelli, 2016; Ricci, 2010), that disciplines have emerged to study their characteristics and benefits even within the socio-technical landscape (Shum & Okada, 2014).

While there are no studies evaluating the usage of data visualizations uploaded on Wikimedia Commons, a spot check on the most viewed pages shows that while some visualizations are featured in the articles, many of them are difficult to read due to their small size. Typically, images on Wikipedia are rendered in a thumbnail 220px wide, and therefore images meant for greater resolutions are difficult, if not impossible, to read.

There are at least a couple of reasons for this. The first is that images on Wikimedia Commons are not explicitly meant to be used on Wikipedia: the only constraint is that the images must be freely licensed (Commons: First Steps/Uploading Files, 2021). There is no requirement for them to be used on Wikipedia or other Wikimedia projects.

This means that images can also be taken from existing publications. This is the second reason why visualizations are often not meant for a small scale: since many governmental bodies release their reports with open licenses, such as Creative Commons, the visualization in question has been uploaded on Commons and then used on the different Wikipedia editions. If on the one hand this is a great example of the circulation of open knowledge, it makes clear the fact that many visualizations were not designed to be seen through the Wikipedia interface.

Finally, while there is an increasingly wide array of tools and software available today to create data visualizations in open formats, it is still quite difficult for non-designers to use them.

Toward an Open Toolkit

The Wikimedia ecosystem therefore provided a suitable environment to identify and test solutions for the design of micro visualizations. As described in the previous section, it establishes a set of constraints (due to the Wikipedia interface) that would require an ad-hoc design process for visualizations.

Another interesting feature is the fact that the Wikimedia Commons platform does not contain just uploaded media, it also includes service pages used by the community to discuss and define issues to be solved. Some pages are used to collect requests — for example, lists of requested photos, maps, diagrams, or data visualizations.

The entire project is driven by the “be bold” approach (Wikipedia: Be Bold, 2022): if you see something you can fix, don’t worry just do it, in the worst case someone else will reverse your edit. This also applies of course to the production of images, allowing them to be tested in the environment.

Finally, there is a space devoted to research meant to improve the platform: Meta-Wiki (Meta-Wiki: Main Page, 2022). Here anyone can notify users of new research projects, handle discussions, and publish results.

Doi: 10.30682/diid7622o

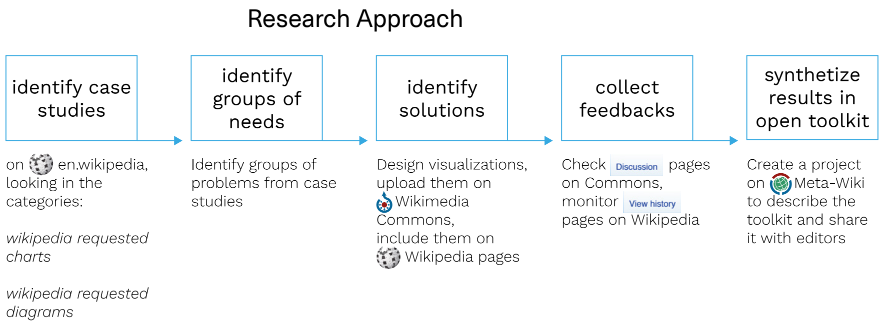

Fig. 2
A diagram representing the process that was followed, the Wikimedia projects involved and the applied functions.

Given the overall goal of providing guidelines to non-designers for the production and improvement of data visualizations on Wikipedia and Wikimedia Commons, we defined an approach based on the following steps: identify case studies, identify groups of needs, identify suitable solutions, collect eventual feedback, and synthesize the results in an open toolkit.

- Identify case studies. Starting from two pages (Category: Wikipedia Requested Charts, 2021; Category: Wikipedia Requested Diagram Images, 2022) we were able to identify several cases in which there was a clear need for data visualizations.
- Identify groups of needs. By analyzing the requested images, it was possible to identify some common needs: for example, the representation of values over time, or the comparison of a couple of continuous dimensions. The goal of this step was to understand if groups of problems could be defined, and solutions identified to solve them.
- Identify suitable solutions. By designing visualizations and uploading them to the platform, it is possible to test them in the environment.
- Collect eventual feedback. Since Wikimedia Commons provides a discussion page for every uploaded image, it can be used to test the community reaction to it. Furthermore, when an image is added to a Wikipedia page, all the users who modified it are notified so they can evaluate the edit quality.
- Synthesize the results in an open toolkit. All the results gathered were then described on a page on the Media-Wiki platform (Optimization of Data Visualizations on Wikipedia in Thumbnail Format, 2021). The page is freely accessible to all interested users and is meant to be a “basecamp” for the discussion related to the topic.

A Taxonomy of Visual Elements

To guarantee a uniformity of results, before starting the design process, a taxonomy of visual elements was defined. From Bertin’s work we adopted the concepts of invariants and components:

> “The INVARIANT is the complete and invariable notion common to all the data. [...] The COMPONENTS are the variational concepts.” (Bertin & Barbut, 1967, p. 16)

Among the invariants, therefore, it is possible to identify external elements such as title, axes, mode of representation, while in the components we can find the elements that change in the graph in relation to the data, for example shape, size, color, or position.

On a parallel path, we leveraged the concept of “visual models”, the strategies identified in the literature to visually map data fields to visual variables (Heer et al., 2010). They can be seen as “blueprints” which can be reused. While there are different ways to categorize them, we followed the approach identified by the Carbon Design System (*Chart Types – Carbon Design System*, 2021), IBM’s open-source library for the design and development of digital products and experiences. Within the Data Visualization section of the CDS, it is possible to trace the categories under which the visual models have been classified, the same used in the project presented here:

- comparisons
- trends
- part-to-whole
- correlations
- connections
- geospatial: overlays.

Visualizations are essentially divided into categories, based on the kind of reading they enable. This classification is not strict and univocal, as explained above; visual models can be versatile, used in different ways, with new communication goals.

Designing Visualizations in a Small Space

Following the steps described in the previous section, sixteen visualizations were designed and uploaded to the platform (Fig. 02). All the uploaded images were added to Wikipedia articles in English, to see how they would be received by other users. Three months after their upload, only two of them have been removed from the Wikipedia page. Three of them have been adopted in other language versions of the same article, or for other articles.

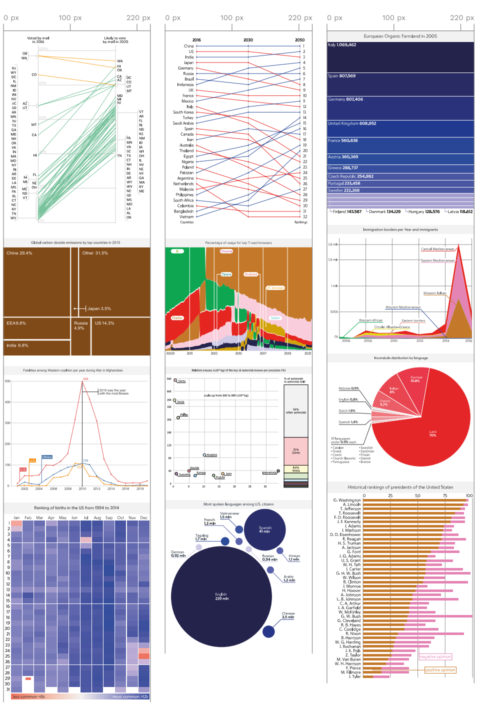

Fig. 3
A selection of the visualizations created for the project, printed simulating their appearance on a standard monitor with a pixel density of 110 PPI (about 44 points per cm).

Notably, there have been few discussions about the uploaded works on the Wikimedia Commons platform. Other users interacted with the uploaded visualizations but mainly on their description and metadata. The impression is that unlike the Wikipedia project, the interaction on Wikimedia Commons takes shape in a more practical way (by adding metadata or fixing errors in the images) rather than in discussions.

The design of micro visualizations for the platform led to a reflection on the design issues that emerged. The main critical areas were the typography, the handling of axes and grids, and the position of the legend.

The issue of typography was expected: in such a small space, it is difficult to identify proper solutions while maintaining readability. Furthermore, Wikimedia Commons is able to render a specific selection of open fonts (SVG Fonts, 2021).

After an analysis of available solutions, two fonts were empirically identified that are highly legible at a small scale: *Poppins* and the *Dm Sans*. These fonts proved to be useful at very small points (5.5pt). Three main uses of typography were identified: handling item labels (e.g. to identify a pie slice, or a point on the scatterplot), annotations (e.g. highlighting a specific data point), text related to axes. Item labels were articulated in two parts: the name of the record and its value.

Due to the lack of space, and the fixed proportion imposed by the interface, the handling of legend boxes proved to be particularly complex. Legends usually contain all the information required to decode a visualization: for example, the meaning of colours or sizes. The solution we identified is to place key values near the encoded values, to make them easier to decode and to save space for the visualization. Part of the solution is therefore using typography to include the legend in the visualization: for example, using a visual style to express the meaning of colours or to connect colours to represented categories.

Axes also proved to be a critical point: they are often irreducible, meaning that they cannot be aggregated to reduce the space they occupy. For this reason, solutions have been found to minimize their impact, reducing them to single values with no visual elements (e.g. maintaining only the graduated grid and avoiding the use of vertical lines).

Results and Discussion

The Open Toolkit

By designing the visualizations, it was possible to evaluate the relevance of single invariants or components across visual models: some of them were more frequent, others had value only on some visual models, others were selected before the design of the individual visualization, being necessary to the overall design. It became evident that for the creation of an open toolkit, the concepts of invariants and components were too abstract, and more pragmatic categories would be more efficient for a lay public. The toolkit was therefore divided into levels: elements, components, and visual models.

Elements constitute the first level of the compositional hierarchy of the proposed toolkit: they represent the features common to all the visualizations. Elements are defined a priori, before the design of the individual visualization, and act as a binder between all the proposed designs. The toolkit defined three elements :

- Typography. Typographical choices and uses necessary to convey the project in the best possible way, building on the experience described in the previous section.
- Chromatics. The chromatic choices that regulate the hierarchies, the reading, and the cognitive processes necessary to decode a visualization.
- Shapes. Visual elements used across visualizations: callouts, highlights, and non-data related symbols.

Components, the second level of the toolkit, are everything that involves the encoding of information, and can be reused in multiple visualizations. Unlike the elements analyzed above, the components are closely related to the bound data, and are used both to encode and to enable decoding. The components include:

- Graphic primitives used to encode data (points, lines, and areas). Good practices are described for each primitive in the toolkit, for example the suggested padding, the minimum and maximum size of single items, and how to handle their appearance, for example the fills and strokes.
- Axes and grids. To optimize the usage of space, suggestions are provided to minimize the size of the axes, with examples and practical cases. The same is true for the grid, providing guidance on when it's useful to keep it and when it's better to discard it to avoid cluttering.
- Legends. As described in the previous section, legends are pivotal for decoding visualizations. They are also space-consuming. Therefore, the toolkit provides good practices to break them down by putting keys near the represented items, avoiding the need for a legend box when possible.

The third and final level of the hierarchy is dedicated to the visual models. Since visual models are categorized based on their function, the context of usage (a small thumbnail rather than other kinds of support) is irrelevant. Four macro-categories of function were structured, namely *comparison*, *time trends*, *part-to-whole*, and *correlations*.

All the contents of the open toolkit are accessible on a dedicated page (Optimization of Data Visualizations on Wikipedia in Thumbnail Format, 2021) on Meta-Wiki, providing a reference for the design of visualizations for all the users involved in the various Wikimedia projects .

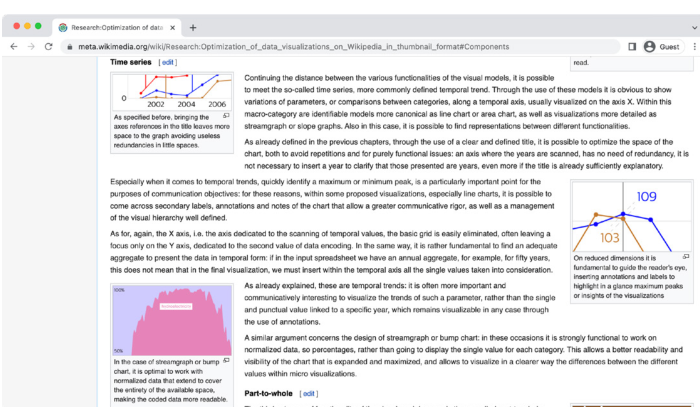

Fig. 4
A screenshot of the page created on Meta-Wiki to illustrate the toolkit. https://w.wiki/_y6m2 (shortened URL)

Discussion

In this research study, we have described the design and development of an open toolkit meant to help non-designers create visualizations in a small space within the Wikimedia projects ecosystem. The toolkit is the result of a practice-based research in which, based on user-defined needs and medium constraints, it has been possible to define macro-groups of problems and propose possible visual solutions to solve them.

The goal from the beginning has therefore been to directly offer solutions to people interested in the subject, to respond to the growing demand for micro design, in this case, micro visualizations. Taking advantage of places such as Wikimedia Commons or Wikipedia, where we could use and test them, it became possible to support and validate the concepts that emerged in the final open toolkit. Underlying the project is a consistent concept of openness, starting from the choice of the publication platform, open to changes and possible implementations.

Given the absence of resources explicitly focused on micro visualizations, in the research presented here we provide a first step for a structured and openly implementable pool of practices and guidelines for designing micro visualizations.

In designing the open toolkit, it became possible to identify plausible future work. Many opportunities have been left open, starting from the compositional hierarchy defined in the toolkit. Some elements could be further explored, such as the use of textures or shapes for encoding data in small visualizations. Likewise, visualizations related to relational data, such as networks, and the vast field related to geospatial representations, were left out of the research. Beyond the possible implementations in the toolkit, there are future prospects for work in areas of design, with a wide range of opportunities in the world of smartwatches, and mobile in general.

Finally, in this research we highlighted the potential of Wikimedia Commons (and more in general, the Wikimedia ecosystem) as an interesting environment for testing and developing visual solutions. As described in the first section of this article, its open policies, which stimulate users to contribute and engage with the work of others, make it a suitable place for collecting feedback. It is important to note that in any case, because the mission of the platform is to provide encyclopedic information, such tests should aim to improve the platform itself.

Acknowledgement of Authors' roles

Andrea Pronzati developed the research presented here as part of his master's thesis in Communication Design, under the supervision of Michele Mauri, and provided a first draft of this article.

Michele Mauri identified the area of research and supervised the work, curated the structure of the article, reworking the content, and was responsible for the editorial process.

Andrea Pronzati
He earned his Bachelor's and Master's degrees in Communication Design from the Politecnico di Milano, with a thesis that explored good practices for small-scale data visualizations. He is currently an art director and information designer. His interest is in information and graphic design, with a focus on typography.

Michele Mauri
Scientific director of DensityDesign Lab, he is a researcher in the Design Department of the Politecnico di Milano. Within the laboratory, he coordinates the research and the design and development of projects related to data and information visualization. He is one of the authors of RAWGraphs, an open-source platform for the creation of data visualizations.

References

Bertin, J., & Barbut, M. (1967). *Sémiologie graphique. Les diagrammes, les réseaux, les cartes*. Mouton; Gauthier-Villars.

Category: Wikipedia requested charts. (2021). *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Category:Wikipedia_requested_charts&oldid=1056827450

Category: Wikipedia requested diagram images. (2022). *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Category:Wikipedia_requested_diagram_images&oldid=779371512

Chart types – Carbon Design System. (2021). https://www.carbondesignsystem.com/data-visualization/chart-types/

Commons: First steps/Uploading files. (2021). *Wikimedia Commons*. https://commons.wikimedia.org/w/index.php?title=Commons:First_steps/Uploading_files&oldid=549651909

Giles, J. (2005). Internet encyclopedias go head to head. *Nature, 438*(7070), 900-901. https://doi.org/10.1038/438900a

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM*. https://doi.org/10.1145/1743546.1743567

Kim, H., Moritz, D., & Hullman, J. (2021). Design Patterns and Trade-Offs in Responsive Visualization for Communication. *Computer Graphics Forum, 40*(3), 459-470. https://doi.org/10.1111/cgf.14321

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He Says, She Says: Conflict and Coordination in Wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453-462. https://doi.org/http://dx.doi.org/10.1145/1240624.1240698

Lanier, J. (2006). *DIGITAL MAOISM: The Hazards of the New Online Collectivism*. Edge.Org. https://www.edge.org/conversation/jaron_lanier-digital-maoism-the-hazards-of-the-new-online-collectivism

Mauri, M., & Ciuccarelli, P. (2016). Designing diagrams for social issues. *Proceedings of DRS2016: Design + Research + Society - Future-Focused Thinking*, *3*, 1–17. https://doi.org/10.21606/drs.2016.185

Medina, J. (2008). *Brain Rules*. Pear Press.

Meta-Wiki: main page. (2022). *Meta-Wiki*. https://meta.wikimedia.org/wiki/Main_Page

Optimization of data visualizations on Wikipedia in thumbnail format. (2021). *Wikimedia Meta-Wiki.* https://meta.wikimedia.org/w/index.php?title=Research:Optimization_of_data_visualizations_on_Wikipedia_in_thumbnail_format&oldid=21367648

Ricci, D. (2010). Seeing what they are saying: Diagrams for socio-technical controversies. In D. Durling, R. Bousbaci, L.-L. Chen, P. Gauthier, T. Poldma, S. Rowoth-Stokes, & E. Stolterman (Eds.), *DRS2010 - Design & Complexity.*

Shum, S. J. B., & Okada, A. (2014). Knowledge cartography for controversies: The Iraq debate. *Advanced Information and Knowledge Processing*, 9781447164692, 293-308. https://doi.org/10.1007/978-1-4471-6470-8_13

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983-1001. https://doi.org/10.1002/asi.20813

SVG fonts. (2021).In *Wikimedia Meta-Wiki*. https://meta.wikimedia.org/w/index.php?title=SVG_fonts&oldid=22030365

Viegas, F. (2007). The Visual Side of Wikipedia. *2007 40th Annual Hawaii International Conference on System Sciences* (HICSS'07), 1–10. https://doi.org/10.1109/HICSS.2007.559

Vincent, N., & Hecht, B. (2021). A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1-15. https://doi.org/10.1145/3449078

Wikipedia: Be bold. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Be_bold&oldid=1074051448

Wikipedia: Neutral point of view. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:-Neutral_point_of_view&oldid=1030312226

Yun, J., Lee, S. H., & Jeong, H. (2019). Early onset of structural inequality in the formation of collaborative knowledge in all Wikimedia projects. *Nature Human Behaviour, 3*(2), 155-163. https://doi.org/10.1038/s41562-018-0488-z