Visrepo: A Centralized and Open Source Visualization Repository and Knowledgebase

Pawandeep Kaur* Andreas Ostrowski†

Heinz-Nixdorf Chair for Distributed Information Systems,
Friedrich-Schiller-Universität, Jena, Germany

ABSTRACT

Through this paper, we propose a visualization repository Visrepo. Visrepo is an effort to collect and organize information about different visualization types from variant sources into a one central knowledgebase. This knowledgebase aims to provide an open source and publicly accessible online environment that: 1) provides the visualization community with a curated central database of different visualization types, 2) that can be reused and edited to fulfill the diverse user requirements, 3) fosters visualization literacy by providing knowledge about different visualization types referenced from scholarly articles and 4) enables long term visualization data curation for visualization theoretical research.

Index Terms: [300]Human-centered computing Visualization theory, concepts and paradigms

1 INTRODUCTION

Visualization effectively presents large chunks of data into comprehensible graphics. Thus, it is an integral component for all those domains which are directly or indirectly engaged with data. The creation of a visualization requires a number of nuanced judgments about data, goal, and effective visual encoding to map data values to graphical features such as position, size, shape, and color [20]. Irrespectively, too often the creators of visualizations (application scientists, programmers and non-visualization domain users) do not spend enough time endeavoring to understand the underlying science they are trying to represent. Just as application scientists sometimes create crude visualizations without understanding the algorithms and science of visualization. [21]. Unlike other disciplines (physics or mathematics), visualization knowledge is spread across different venues and is not definite [16] or standardized. The downside of it is that, it leads to a contradictory opinion on different visualization concepts and techniques. While trustworthy visualization sources, e.g. visualization publications or books may be difficult to comprehend for visualization practitioners or domain users, amateur sites or blogs may mislead users [30].

In our previous study [22], we had surveyed the biodiversity community about their challenges concerning the usage of visualization and their possible solutions. We found that due to their limited visualization knowledge, biodiversity domain scientists are often unaware of the various visualization techniques available today. Due to this, they tend to use similar visualization types again and again, even though they feel is not effective in their work. Their suggestion in eliminating this problem was that, a visualization tool should provide a showcase of all the visualizations presented in it. Here, the showcase means a one stop shop of the collection and organization of different types of visualizations, their design requirements and their specific domain relevance.

*e-mail: pawandeep.kaur@uni-jena.de
†e-mail: andreas.ostrowski@uni-jena.de

There exist numerous platforms that inform about different types of visualizations. These platforms are in the form of websites, blog, scientific sources (publications or books) which give information about different visualization in the form of examples, discussion forums etc. For instance, at the time of this writing, more than 300 different visualizations are listed on the D3.js [10] website. D3.js is a javascript library for creating visualizations and their showcase is a collection of different visualizations (with examples and program codes) that can be constructed through this library. In order to build a visualization knowledgebase, one needs to bring all different visualizations types and its related information (why it is used, when it is used, what type of data it represents etc.) at one place. Moreover, to make it easily understandable for different communities one needs to include its domain relevance and usage; for example, a dendrogram is a type of tree diagram which in biological sciences is used for the purpose of representing clusters and phylogenies. All this information needs to be properly organized, maintained and preserved. For every other domain, making such a platform from scratch includes a huge amount of work, cost, resources etc. There are currently many privately owned websites available (see Section 2.1, point 2), that do such a visualization showcase. However, one major problem is that they are not open source. Therefore, they cannot be edited or updated to make them suitable for a specific community or domain (e.g. biodiversity as in our case). Moreover, as they are privately owned, they lack visualization community engagement which can enrich such platforms with expert advice.

In this paper, we propose to build a centrally available, open source and community curated visualization repository. Being an open source environment would allow diverse users to reuse and rework on their own copy as per their domain requirements. Moreover, it would be an eminent source of educational material for visualization courses. As, it has been expressed in this paper [13] as well, such a collected knowledge could serve as an important source of information for future theoretical visualization research activities e.g. taxonomies, ontologies, design gallery and spaces.

In the following sections, we provide a brief overview on the similar work in Section 2, its features and services in Section 3, overview of its components in Section 4, some challenges in Section 5 and conclusion and future directions in Section 6.

2 RELATED WORK

2.1 Visualization showcase platform

In recent years plenty of platforms, website, blogs etc. have emerged showcasing different visualizations. In the following we have categorized some of them:

1. Software visualization showcase: Visualization showcases that are provided by individual visualization software projects to showcase all those visualizations that can be build by their software. For example : Vega-lite [34], D3.js [10], RAWGraphs [6] and DC.js [5].

2. Visualization catalogue: Privately owned visualization showcasing websites. Some of them are: data visualization cat-

alogue [2], data visualization project [3], visualization universe [8] and visualization periodic table [24].

3. Forums: Online available privately owned blogs related to visualization topics: adioma [7], reddit [4], information is beautiful [26], flowing data [35] and atlas [1].

4. Surveys: Visualization surveys as books [19], publication or online portal [31]. Their focus is mostly on teaching and learning purposes than the one we described in the section 3.

2.2 Visualization Databases

Data curation can be described as a process of collecting data from diverse sources and integrating it into repositories that are many more times valuable than the independent part. In the curation process, data is organized, described, cleaned, enhanced, and preserved for public use [28]. Here we have listed some of the example data repositories from different disciplines: [9,15,18,23,25,29]. Though the value of having such a repository or showcase for visualization has been advocated years before [11], still such projects are very limited in the visualization domain. Visguides [16] and Milestone Project [17] are two such projects that we are aware of. Visguides is a visualization community driven collaborative effort to collect and discuss various visualization guidelines. Milestone project is an attempt to collect and organize information about the history of different data visualizations. Our proposition of a visualization repository (Visrepo) is an attempt to collect diverse information about different visualization types at one central platform so that it can be easily reused, re-edit and referenced by different communities.

3 Features and Services

In the following, we have highlighted some of the features (1-3) and services (4-7) that should be provided by the Visrepo:

1. Centralized Management: There is no definitive central database for referencing different visualization types. As visualization is perceived as an art of data representation, plethora of visualizations has been developed for efficiently visualizing the data. Visualization knowledge got distributed among different publications, books, websites and blogs. The side effect of this distribution of knowledge is conflicting viewpoints on different visualization guidelines. Thus, to make visualization knowledge more definitive such a visualization repository or database is needed.

2. Community curated: To make it a reliable source, such a knowledgebase is proposed to be created by the members from the visualization community who understand the science of this domain well. Moreover, this information should be supported by references from scholarly articles. Such an environment should be open for the other members of the community where they can question and discuss. Community engagement will also assist in proper curation and preservation of the knowledge.

3. Open source and reusable: To provide the benefit of such knowledgebase to the public, it should be kept as an open source platform. This knowledgebase can only be beneficial to other communities, if they are allowed to reuse, edit and update it for their own specific purposes. It will spare them to do this work from the scratch and they can add up their own work on top of the copy of this database.

4. Version management: Different versions of this knowledgebase referencing to the central and original copy will have their own benefits, as users will remain aware about their work being used and referenced. There are so many such projects going on that use the same phenomena where one main copy is retained and others are forked by different users. For example: github and bioportal [33]. Github is a source code management system wherein the software code can be reused and referenced for other software projects. Bioportal is a biological ontology management system where one can develop new ontologies by building on top of existing ones. The ontology page then shows all the versioned ontologies based on the main ontology.

5. Pedagogical benefits: This knowledgebase could be an eminent source of visualization literacy to the public. Improved visualization literacy is one of the important challenges mentioned in [14]. Visualization information from this knowledgebase will serve to train practitioners, novices and would help teachers in their curriculum. Moreover, with domain specific database versions, one would also be aware of the significance of different visualizations in different domains.

6. Knowledge-based systems: Knowledge-based visualization systems are highly dependent on similar knowledge sources [12]. Thus, such a knowledgebase can be embedded in the development of different visualization systems.

7. Future visualization research innovations: This curation of different visualizations and its long term preservation in the form of structured knowledgebase would be a valuable asset in future visualization theoretical researches, for e.g. in the formalization of ontologies and taxonomies [13].

4 Components of a Knowledgebase

In (Fig. 1), we provide an overview of the proposed framework. Individual components are described below:

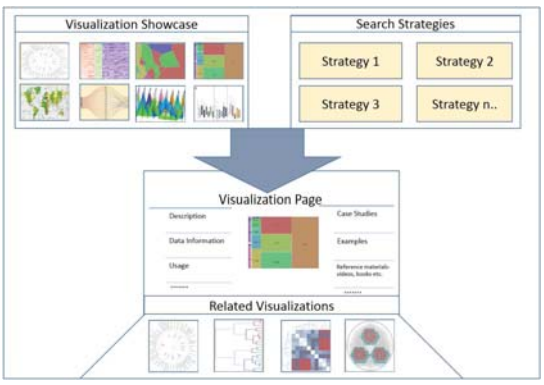

Figure 1: Schema showing the components of visrepo. Visualization showcase and different search strategies enable users to find the desired visualization. Then, from that visualization page one can navigate to other related visualization pages.

1. Visualization showcase: A landing page of this system which showcases all the visualizations in a thumbnail view available in the portal.

2. Visualization search strategies: Having different visualizations at one place can be too cumbersome for the users to browse. Therefore, to reduce the visualization search space, we propose that such system should have different visualization filters or search strategies. Visualization literature provides many different dimensions by which intended visualization can be specified. For example: visualization search by dataset type [27], data attribute type, user goals, visualization

tasks, visualization techniques [32], data domain etc. As currently, we are in the preliminary stages, therefore we are not in the position to provide a definite list of different search strategies that would be used in the system.

A specific 'Domain relevant search' strategy for different domains could be used to filter down the visualizations that are specific to that particular domain (see Fig. 2). Thus different versions of the same copy could be merged into the original version providing a complete overview of information present in the database. Then, through the use of filter/search functions specific copies can be accessed.

3. Visualization page: A dedicated page for each visualization providing detailed information about it. To maintain the authenticity of this information, references of scholarly articles are proposed to be included for every claim. Further component of this page would provide a detail information related to the specific visualization, for example: visualization image, description, general usage, domain specific usage (see Fig. 2), domain specific examples (from authentic sources), visualization goals, data types, guidelines specific to this visualization, related visualization as per (goal, data types, etc.), related publications, related explanatory videos/blogs, etc.

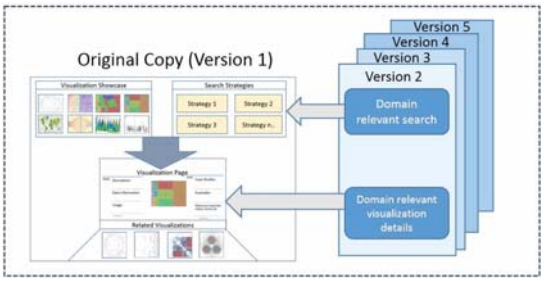

Figure 2: Versioning in Visrepo

5 Challenges

Deploying such an environment on a large scale, comes with the multitudinal challenges. Though we cannot foresee all of them beforehead but we have tried to provide few of them in (see Fig. 3) as a starting point for further planning and discussion.

6 Conclusion and Future Directions

In this paper, we have discussed our proposal about building a visualization repository for the collection and organization of different visualizations on one platform. Our proposal is motivated from the results of our previous user study, previous expert advice and a current state of diverse, scattered and dynamic visualization knowledge. We will like to strongly emphasize that our proposal for Visrepo is not about creating one more visualization collection but also to fulfill the goals that are mentioned in Section 3. Our immediate need is to provide such a repository to our biodiversity domain users. Alongside, we are equally interested in building it on a large scale so that it can fulfill various discussed services. We have also identified few of the challenges that needs to be addressed for the deployment of the Visrepo at the large scale. Through the medium of this paper, we would also like to invite interested visualization community members to come together and participate in the future discussion and planning of this project.

Figure 3: Some of the challenges in the deployment of the Visrepo

Acknowledgments

The work has been funded by the DFG Priority Program 1374 "Infrastructure-Biodiversity-Exploratories" (KO 2209 / 12-2). We would like to thanks our project principal investigator Prof. Dr. Birgitta König-Ries for her valuable feedback.

References

[1] Atlas. https://www.theatlas.com/.

[2] The data visualisation catalogue. https://datavizcatalogue.com/.

[3] Data viz project. http://datavizproject.com/.

[4] dataisbeautiful. https://www.reddit.com/r/dataisbeautiful/.

[5] Index of dc.js examples. http://dc-js.github.io/dc.js/examples/.

[6] Rawgraphs gallery. https://rawgraphs.io/gallery/.

[7] V. A. and V. M. Adioma blog - infographics, design, and content marketing. https://blog.adioma.com/.

[8] V. A. and V. M. Visualization universe. http://visualizationuniverse.com/charts/.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, pp. 675–684. Springer, 2006.

[10] M. Bostock. d3.js. https://github.com/d3/d3/wiki/Gallery.

[11] C. Chen. Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications*, 25(4):12–16, July 2005. doi: 10.1109/MCG.2005.91

[12] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, K. L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *IEEE Computer Graphics and Applications*, 29(1):12–19, Jan 2009. doi: 10.1109/MCG.2009.6

[13] M. Chen, G. Grinstein, C. R. Johnson, J. Kennedy, and M. Tory. Pathways for theoretical advances in visualization. *IEEE Computer Graphics and Applications*, 37(4):103–112, 2017. doi: 10.1109/MCG.2017.3271463

[14] M. Chen, A. Trefethen, R. Banares-Alcantara, M. Jirotka, B. Coecke, T. Ertl, and A. Schmidt. From data analysis and visualization to causal-

ity discovery. *Computer*, 44(10):84–87, Oct 2011. doi: 10.1109/MC.2011.313

[15] C. Cieri and M. Liberman. Issues in corpus creation and distribution: The evolution of the linguistic data consortium, 2000.

[16] A. Diehl, A. Abdul-Rahman, M. El-Assady, B. Bach, D. Keim, and M. Chen. VisGuides: A Forum for Discussing Visualization Guidelines. In J. Johansson, F. Sadlo, and T. Schreck, eds., *EuroVis 2018 - Short Papers*. The Eurographics Association, 2018. doi: 10.2312/eurovisshort.20181079

[17] M. Friendly, M. Sigal, and D. Harnanansingh. The milestones project: a database for the history of data visualization. *Visible Numbers: The History of data Visualization. Ashgate Press, London*, 2012.

[18] W. M. Gelbart, M. Crosby, B. Matthews, W. Rindone, J. Chillemi, S. R. Twombly, D. Emmert, M. Ashburner, R. A. Drysdale, E. Whitfield, et al. Flybase: a drosophila database. the flybase consortium. *Nucleic acids research*, 25(1):63, 1997.

[19] R. Harris. *Information Graphics: A Comprehensive Illustrated Reference : Visual Tools for Analyzing, Managing, and Communicating*. Management Graphics, 1996.

[20] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Commun. ACM*, 53(6):59–67, June 2010. doi: 10.1145/1743546.1743567

[21] C. Johnson. Top scientific visualization research problems. *IEEE Comput. Graph. Appl.*, 24(4):13–17, July 2004. doi: 10.1109/MCG.2004.20

[22] P. Kaur, F. Klan, and B. König-Ries. Issues and Suggestions for the Development of a Biodiversity Data Visualization Support Tool. In J. Johansson, F. Sadlo, and T. Schreck, eds., *EuroVis 2018 - Short Papers*. The Eurographics Association, 2018. doi: 10.2312/eurovisshort.20181081

[23] S. M. Kent. Sloan digital sky survey. *Astrophysics and Space Science*, 217(1-2):27–30, 1994.

[24] R. Lengler and M. J. Eppler. A periodic table of visualization methods. `http://www.visual-literacy.org/periodic_table/periodic_table.html`.

[25] T. Lotz, J. Nieschulze, J. Bendix, M. Dobbermann, and B. König-Ries. Diverse or uniform?intercomparison of two major german project databases for interdisciplinary collaborative functional biodiversity research. *Ecological informatics*, 8:10–19, 2012.

[26] D. McCandless. Information is beautiful. `https://informationisbeautiful.net/`.

[27] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015.

[28] L. Peer, A. Green, and E. Stephenson. Committing to data quality review. *IJDC*, 9(1):263–291, 2014.

[29] R. J. Sandusky. Computational provenance: Dataone and implications for cultural heritage institutions. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3266–3271, 2016. doi: 10.1109/BigData.2016.7840984

[30] E. K. Schofield. Quality of graphs in scientific journals: an exploratory study. *Science Editor*, 25(2):39–41, 2002.

[31] C. Tominski and W. Aigner. A visual survey of visualization techniques for time-oriented data. `http://timeviz.net/`.

[32] M. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Taylor & Francis, 2010.

[33] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl_2):W541–W545, 2011.

[34] K. Wongsuphasawat. vega-lite. `https://vega.github.io/vega-lite/examples/`.

[35] N. Yau. Flowing data. `http://flowingdata.com/`.