Principles for Designing for Data Exploration for a Non-Expert Audience

Natasha Tylosky*, Antti Knutas*, Annika Wolff*
* LUT University, Department of Software Engineering, Lappeenranta, Finland
Natasha.Tylosky@lut.fi, Antti.Knutas@lut.fi, Annika.Wolff@lut.fi

***Abstract*—A gap currently exists between how data is presented by researchers and how the average person understands data. Thus, although an abundance of good research data is available for free online, it can be difficult to impossible for laypeople to access and make sense of it on their own. The issue of data accessibility and interpretability is consequently preventing open science goals from being reached. Data exploration systems can provide people with limited data expertise an entrance point into complex data and help remedy this problem. This paper presents design principles that can be applied when creating a data exploration software system.**

Keywords*—*Data Exploration, Interactive Data Visualization, Sensemaking, Systematic Mapping Study, Design Principles

I. INTRODUCTION

In today's world, access to open data is readily available to anyone with an internet connection. There is, however, a gap in how the scientific community presents data and how the average person can interpret that data. Consequently, open scientific data sets and open data portals are not always findable or usable for those with less data expertise [1], [2]. Furthermore, there is research potential in inviting those with less data expertise to give their own input and interpretations of data [3], [4]. While there have been recent literature reviews on human data interaction systems and user interfaces in the field of HCI [5]–[7], the focus of these reviews has been on usability as a whole without considering the intended users of such systems and interfaces. Given that traditional data systems often assume that users possess good knowledge of the meaning and contents of a database, and that users are certain of the information they are looking for [8], this lack of consideration given towards non-expert audiences when designing data exploration systems can be considered a potential hurdle for users who are new to such systems and could discourage non-expert users from engaging with open data.

In this work, we conduct a systematic mapping review that gives a perspective on the types of data exploration systems that have been created with a non-expert audience in mind. From this systematic mapping review, we seek to synthesize a set of design principles for designing data visualization systems that encourage data exploration in a non-expert audience. In this paper, we define a "non-expert audience" as anyone who can be considered as not being a domain expert of the information found in a visualization. For example, in the context of medical data visualization, medical doctors would be considered an expert audience, whereas patients would be considered non-expert. If a data visualization system is designed for patients, it would be designed for a non-expert audience. In addition to the above, we also examine the domains in which data visualization research is being applied and the types of visualizations that are being used, as it is possible that there may be interesting links between the domain and visualization model or audience and visualization model.

In the following Section, we review related research on the topic. In Section 3, we present the systematic mapping process and continue with thematic synthesis of mapping outcomes in Section 4. We conclude with discussion of outcomes in Section 4 and conclusions in Section 5.

II. RELATED RESEARCH ON DATA VISUALIZATION AND EXPLORATION

When hearing the term "data visualization", the first thing that springs to mind may be 2D visualizations such as line graphs and bar charts, and while these are indeed common methods of visualization [9], data visualization does not have to be limited to such methods. For example, the Peruvian knot tying technique known as quipu is an indigenous method of data visualization that predates computers [10]. In this paper, we are interested in uncommon methods of data visualization, specifically interactive data visualizations that enable sensemaking of data. Sensemaking can be described as the process of constructing meaning from information and is an iterative process that involves linking different pieces of information into a single conceptual representation [11], [12]. Here, we will be using the inclusion of sensemaking in a data visualization system as a marker for whether the system is exploratory [8], i.e., it is a data exploration system or tool.

III. SYSTEMATIC MAPPING STUDY

As a starting point for our research, we conducted a systematic mapping study of trends in interactive data visualizations in the Human Computer Interaction (HCI) field. The aim of a systematic mapping study is to gain a comprehensive overview of a particular research topic, present unbiased assessment of current literature, identify research gaps and collect evidence for future research directions [13].

This study been funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 872500.

A. Keyword and database selection

We initially did deep search testing of various keywords and search strings in Scopus and Science Direct. The goal of this deep search was to find the keywords that produced the best results; a variety of keywords were needed in this review as there are science communication related papers that may contain pertinent research which may not appear under computer science focused terms. After the deep search, another search was done across 4 databases, namely, Scopus, Science Direct, IEEE, and ACM, using the keywords we had chosen. The final search terms were as follows.

1) ("Data Storytelling" OR "Participatory Storytelling" OR "Data Visualization" AND "Storytelling")
2) ("Data Exploration" AND "HCI" OR "Data Exploration" AND "Human Computer Interaction" OR "Science Education" AND "Data Exploration")
3) ("Sensemaking in Interactive Space")

B. Inclusion criteria and selection process

Following the literature search, a set of inclusion criteria was defined that selected papers must meet.

1) Inclusion criteria:

- The paper must research data dissemination aimed at a non-expert audience. This review excludes data dissemination directed at domain experts. [14]
- The methods used in the paper must include a data visualization aimed at a non-expert audience. This is necessary as the goal of this paper is to look at data exploration systems for non-expert audiences.
- The audience must interact with the data in some way. The intended audience must be able to interact with the data in some way. In the case of static visualizations, this could be an activity that does not change the visualization itself.
- The paper needs to be peer reviewed.
- The paper is published in English.
- The paper must be published between 2016 and 2021. This criterion is included to focus on recent research, we are only interested in recent research as data exploration is a new concept.

2) Paper selection process:

1) The title of the paper was read to determine if the paper potentially matched the search.
2) Duplicate studies were removed.
3) During the first pass the abstract of each paper was read to determine if the study fits the inclusion criteria.
4) During the second pass the remaining papers were reviewed more deeply.

C. Search outcomes

A total of 24 papers were ultimately chosen for inclusion in the literature review. The stages of the paper selection process are shown in Fig. 1.

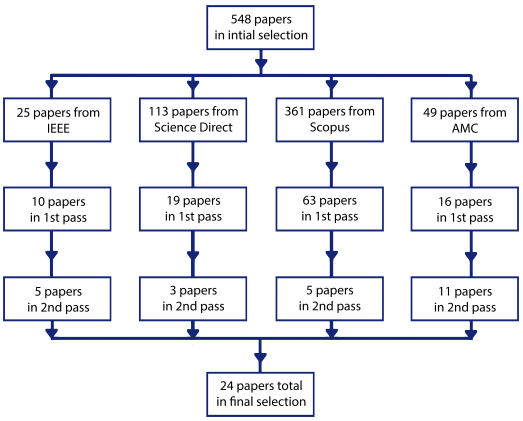

Fig. 1: Flowchart showing the paper selection process for the literature review

Full list of articles included in systematic mapping study are available at the Zenodo open data repository [1].

IV. THEMATIC SYNTHESIS

Data was extracted and summarized using the thematic synthesis method defined by Cruzes and Dyba [15]. The approach comprises a process of steps for extracting, coding, and translating codes into themes. The themes are then distilled into higher order themes and subsequently synthesized to shed light on the research question. The categories selected based on thematic synthesis were: 1. Visualization Model, 2. [9] Intended Audience, and 3. Application Domain [16]. The category of visualization model referred to words and phrases related directly to the visualization itself, such as 3D, interactive, or time series; intended audience included any words or phrases that mentioned participants, users, and such synonyms; and application domain included words and phrases related to the purpose of a visualization. The initial coding process produced 16 distinct codes for visualization models, 14 distinct codes for intended audience, and 14 distinct codes for application domains. These codes were then grouped into further codes based on similarity. 3 final codes related to the visualization model category, 4 codes related to the intended audience, and 4 codes related to the application domain. Diagrams of each thematic synthesis can be seen in Fig. 2.

From the thematic synthesis, we compared the application domains and intended audience in relation to the visualization models to examine if there were any trends in how these concepts were used. The results can be seen in the Table 1.

V. PRINCIPLES BASED ON THE SYSTEMATIC MAPPING STUDY

Next, we searched for common themes in the findings of the 24 selected papers, and from these common themes, we then synthesized three higher-order themes related to

[1] https://doi.org/10.5281/zenodo.7575802.

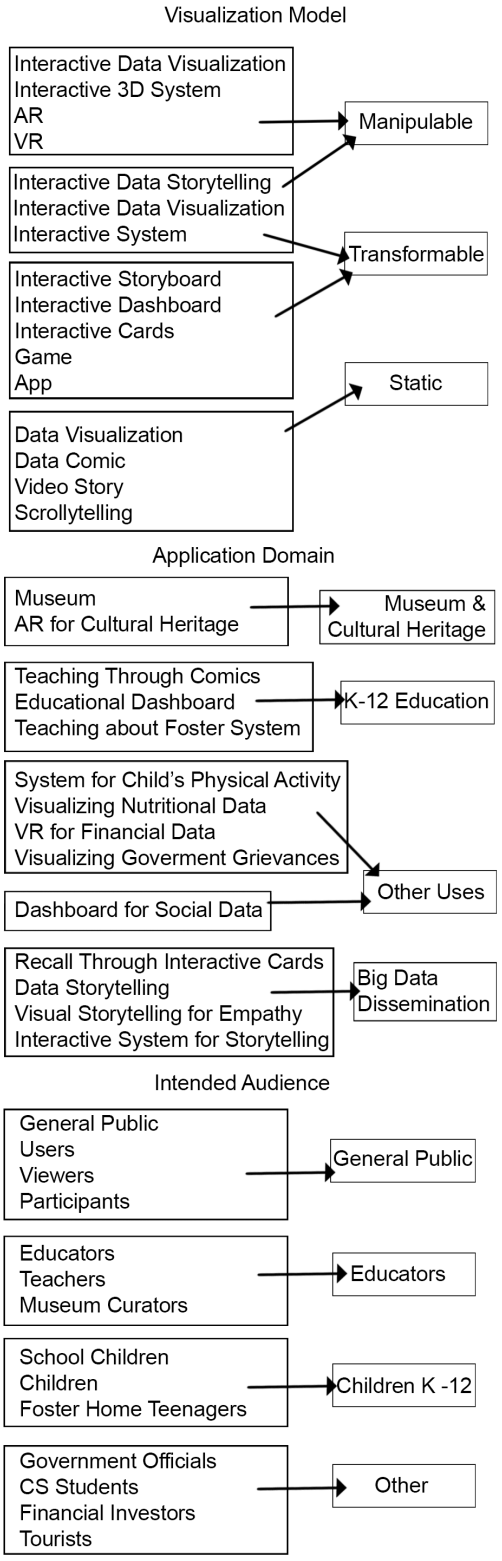

Fig. 2: Thematic synthesis process

TABLE I: Relationships between thematic synthesis categories

	Visualization Models		
	Manipulable	**Transformable**	**Static**
Intended Audience			
General Public	3	3	4
Children	2	1	2
Educators	0	2	1
Other	3	2	1
Application Domain			
Museum & Cultural Heritage	3	2	0
K-12 Education	0	3	1
Big Data Dissemination	3	2	5
Other	2	2	1

sensemaking of data with a non-expert audience. From this we ascertained three themes which provided the principles for designing data exploration software systems.

1) Principle 1: Allow users to find their own patterns in the data. Data exploration requires that a user not only be able to interact with the data but come to their own conclusions about what may be going on in the said data by allowing users to find their own patterns within the data [8], [17]–[20]. The same data often allows for different interpretations depending on the user [21], which can be seen as an advantage in the context of data exploration.

2) Principle 2: Visualize relationships between data sets. It is important to provide a way to visualize the potential connections between different data sets, as data exploration is about finding connections in data [17]. A data exploration system should allow users to sort information via a set of concepts or ideas [22]–[24] rather than having to investigate each data set in isolation.

3) Principle 3: Use artistic expression to engage non-expert users. One of the three principles is that a data visualization should use artistic expression [25], [26] and drive the audience's attention by using visual and narrative elements [27], [28]. An example of this could be data comics which make use of storytelling applied to sequential art or using info-graphics. [29]–[31]

VI. DISCUSSION

In this study, we synthesized literature on data exploration into a set of design principles for creating new data exploration systems for non-expert users. Data exploration is a fairly new sub-field and currently no prescriptive guidelines exist for the process of creating data exploration systems. The design principles introduced in this paper may fill that gap when designing such systems for a non-expert audience and extend the state of the art in prescriptive design knowledge.

Our recommendations, grounded in literature from the field, underline the importance of: 1) empowering the user and allowing the freedom to explore, 2) providing visual ways of interacting with the field, and 3) using artistic expression to engage non-expert users.

Related literature studies exist for example on information systems usability and prescriptive knowledge in the form of design patterns [5]–[7]. In comparison to earlier

studies, we have three novel contributions to the start of the art: 1) We consider the specific intended audience of a visualization or a system, 2) provide prescriptive knowledge specifically to designers of data exploration systems for general audiences, and 3) provide an updated descriptive review to academic audiences.

The main limitation of the study is that while the principles are strongly grounded in literature, they need further testing in practice together with designers and system users. The second limitation of this study is that the systematic mapping review only incorporated papers published between the years of 2016 and 2021, meaning that some seminal papers may not have been included in the review.

VII. CONCLUSION

In this paper, we sought to define a set of design principles by conducting a systematic mapping review, followed by a thematic synthesis. This gave us a perspective on the types of data exploration tools that have been made, and we then derived a set of principles for designing a data exploration system for a non-expert audience, based on the findings and perspectives in the 24 papers selected in the review. These design principles were: 1. Enable users to find their own patterns in the data 2. Visualize the relationships between different data 3. Use artistic expression to attract the attention of a user. In future study we will further evaluate these principles via user testing of a prototype data exploration tool that makes use of these principles in its design.

REFERENCES

[1] E. Osagie, M. Waqar, S. Adebayo, A. Stasiewicz, L. Porwol, and A. Ojo, "Usability evaluation of an open data platform," 06 2017, pp. 495–504.

[2] V. Weerakkody, Z. Irani, K. Kapoor, U. Sivarajah, and Y. Dwivedi, "Open data and its usability: an empirical view from the citizen's perspective," *Information Systems Frontiers*, vol. 19, 04 2017.

[3] L. Garcia, R. Riedy, J. Polman, R. Swanson, T. Nuessle, and N. Garneau, "Expanding public participation in science practices beyond data collection," *Citizen Science: Theory and Practice*, vol. 6, p. 11, 04 2021.

[4] J. Shirk, H. Ballard, C. Wilderman, T. Phillips, A. Wiggins, R. Jordan, E. McCallie, M. Minarchek, B. Lewenstein, M. Krasny, and R. Bonney, "Public participation in scientific research: A framework for deliberate design," *Ecology and Society*, vol. 17, pp. 29–48, 06 2012.

[5] E. Victorelli, J. dos Reis, H. Hornung, and A. Prado, "Understanding human-data interaction: Literature review and recommendations for design," *International Journal of Human-Computer Studies*, vol. 134, 09 2019.

[6] T. Zaki and M. N. Islam, "Neurological and physiological measures to evaluate the usability and user-experience (ux) of information systems: A systematic literature review," *Computer Science Review*, vol. 40, p. 100375, 05 2021.

[7] L. Punchoojit and N. Hongwarittorrn, "Usability studies on mobile user interface design patterns: A systematic literature review," *Advances in Human-Computer Interaction*, vol. 2017, pp. 1–22, 11 2017.

[8] S. Idreos, O. Papaemmanouil, and S. Chaudhuri, "Overview of data exploration techniques," ser. SIGMOD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 277–281. [Online]. Available: https://doi.org/10.1145/2723372.2731084

[9] R. Mazza, *Introduction to Information Visualization*, 1st ed. Springer Publishing Company, Incorporated, 01 2009.

[10] A. Hosne, "Assessing indigenous forms of writing," *Journal of Jesuit Studies*, vol. 1, pp. 177–191, 03 2014.

[11] A. Blandford and S. Attfield, *Interacting with Information*, 01 2010, vol. 3.

[12] L. Koesten, K. Gregory, P. Groth, and E. Simperl, "Talking datasets – understanding data sensemaking behaviours," *International Journal of Human Computer Studies*, vol. 146, 11 2020.

[13] M. Salama, R. Bahsoon, and N. Bencomo, *Managing Trade-offs in Self-Adaptive Architectures: A Systematic Mapping Study*, 08 2016, pp. 249–297.

[14] D. Cvetkovic, *Introductory Chapter: Multimedia and Interaction*, 04 2019.

[15] D. S. Cruzes and T. Dyba, "Recommended steps for thematic synthesis in software engineering," in *2011 International Symposium on Empirical Software Engineering and Measurement*, 2011, pp. 275–284.

[16] R. Glass and I. Vessey, "Focusing on the application domain: everyone agrees it's vital, but who's doing anything about it?" vol. 3, 02 1998, pp. 187 – 196 vol.3.

[17] Y. Loukissas, *All Data Are Local: Thinking Critically in a Data-Driven Society*, 01 2019.

[18] W. Zeng, A. Dong, X. Chen, and Z.-l. Cheng, "Vistory: Interactive storyboard for exploring visual information in scientific publications," *J. Vis.*, vol. 24, no. 1, p. 69–84, feb 2021. [Online]. Available: https://doi.org/10.1007/s12650-020-00688-1

[19] F. Bosello, C. Colli, P. Clini, I. Dražić, D. Dražina, A. Pugliese, R. Quattrini, and G. Vettorel, "Remember: an adriatic network of virtual museums to highlight ports culture, traditions and history," *IOP Conference Series: Materials Science and Engineering*, vol. 949, p. 012082, 11 2020.

[20] D. Blezinger and E. van den Hoven, "Storytelling with objects to explore digital archives," ser. ECCE '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2970930.2970944

[21] J. Liem, C. Perin, and J. Wood, "Structure and empathy in visual data storytelling: Evaluating their influence on attitude," *Computer Graphics Forum*, vol. 39, 06 2020.

[22] S. Podworny, Y. Fleischer, S. Hüsing, R. Biehler, D. Frischemeier, L. Höper, and C. Schulte, "Using data cards for teaching data based decision trees in middle school," 10 2021.

[23] Z. Koukopoulos and D. Koukopoulos, "A participatory digital platform for cultural heritage within smart city environments," 12th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), 2016, pp. 412–419.

[24] A. Beheshti, A. Tabebordbar, and B. Benatallah, "istory: Intelligent storytelling with social data," 04 2020, pp. 253–256.

[25] S. Ketchell, W. Chinthammit, and U. Engelke, "Situated storytelling with slam enabled augmented reality," in *Proceedings of the 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry*, ser. VRCAI '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3359997.3365681

[26] Z. Wang, H. Dingwall, and B. Bach, "Teaching data visualization and storytelling with data comic workshops," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–9. [Online]. Available: https://doi.org/10.1145/3290607.3299043

[27] V. Echeverria, R. Martinez-Maldonado, R. Granda, K. Chiluiza, C. Conati, and S. Buckingham Shum, "Driving data storytelling from learning design," 03 2018, pp. 131–140.

[28] V. Echeverria, R. Martinez-Maldonado, and S. Buckingham Shum, "Towards data storytelling to support teaching and learning," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, ser. OZCHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 347–351. [Online]. Available: https://doi.org/10.1145/3152771.3156134

[29] B. Bach, N. Henry Riche, S. Carpendale, and H. Pfister, "The emerging genre of data comics," *IEEE Computer Graphics and Applications*, vol. 38, pp. 6–13, 05 2017.

[30] D. Seyser and M. Zeiller, "Scrollytelling – an analysis of visual storytelling in online journalism," 07 2018, pp. 401–406.

[31] Z. Wang, S. Wang, M. Farinella, D. Murray-Rust, N. Henry Riche, and B. Bach, "Comparing effectiveness and engagement of data comics and infographics," 02 2019.