The Importance of Data Curation for Data Physicalization

Hannes Waldschütz
Bauhaus-Universität Weimar,
99423 Weimar, Germany
hannes.waldschuetz@uni-weimar.de

Eva Hornecker
Bauhaus-Universität Weimar,
99423 Weimar, Germany
eva.hornecker@uni-weimar.de

DIS '20 Companion, July 6–10, 2020, Eindhoven, Netherlands

ACM ISBN 978-1-4503-7987-8/20/07.
https://doi.org/10.1145/3393914.3395892

Abstract
Creating physicalizations requires a high amount of decision taking, in particular in selecting and staging the data to be conveyed, which can be challenging. We give insight into the process of conceptualizing and building physical representations of data. Based on our practical experiences, we discuss preliminary recommendations on how to approach design of physicalizations.

Author Keywords
Data physicalization; Design; Data processing.

CSS Concepts
•Human-centered computing~Visualization~Visualization application domains~Information visualization

Introduction
Physical representations of data allow novel and multimodal approaches of representing data [1,2] that are not bound, unlike traditional information visualization to visual 2D communication, but can be realized in three dimensional, multimodal setups (fig. 1,2). A growing body of projects and literature investigate concepts of conveying data and user research. Regarding perception, accessibility and legibility, the benefits of 'physicalizations' [3,4], as well

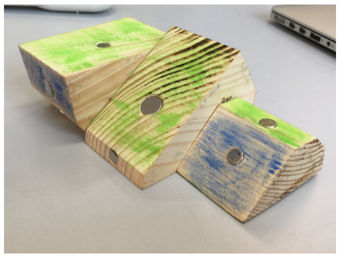

Figure 1: A data physicalization of 'Mean Appropriateness Ratings', based on workshop material from Huron et al [6]

Figure 2: Data physicalization of 'Distribution of Consumed Alcohols' based on workshop material from Huron et al [6]

Acknowledgments.
We thank the students taking part in our courses who created the depicted objects and installations. Work on this article was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 420584617 / HO3901/4-1

as their potential for new ways of interaction [5], have been objects of research. Yet there is little research on the genesis of such representations, although this could be useful to gain a better understanding of the difficulties involved, and the possibilities and limitations encountered in their development and design.

In our research-oriented projects with student groups we identified a recurring difficulty in the process of conceptualizing and designing data physicalizations. It occurs at the step of choosing and processing appropriate data to be translated into a design. This observation is important since it sheds some light on what can be represented with data physicalization and allows to provide guidelines for the design process.

Design

Conceptualizing and designing a data physicalization is a multidimensional process: A topic needs to be defined, data sources identified, technology evaluated towards feasibility of construction, and an artifact worked out. Each step involves a wide range of methodologies. This may include (but is not limited to) methods of data analysis, programming, building of electronic circuitry and mechanics, design and form-giving, and finally, constructing and crafting the prototype.

Similar to others, we have utilized 'finger exercises' as part of workshops on data physicalization to introduce our students to the topic [6]. We found these exercises to provide an excellent entry point for physicalization novices, who develop interesting representations in a matter of hours. But since these tasks are based on predefined and rather small datasets, they do not lead to an understanding of the challenges in designing data physicalizations for more unconstrained, larger data sets.

Throughout our work in developing physicalizations with small project groups, one step in the process required effort and careful choices: the selection and treatment of a dataset so as to fit the technology and materials of choice and to be legible for its audience. From a distance, this process resembles the InfoVis visualization pipeline – consisting of analysis, filtering, mapping and rendering data [7]. For designing graphic (in particular digital) data representations, there are virtually no constraints to convey a full dataset. Digital representations tend to come with support tools for interactive exploration, search, dynamic filtering and highlighting.

But with physical artifacts and constructions, one has to deal with constraints regarding the properties and feasibility of technology and materiality. These constraints are not just a minor side effect, they heavily shape the scope and complexity of the displayed data. As a result, not only processing, but also curation of the data is likely if not necessary, to reduce the number of data points to fit the physical representation (see case studies). This seems even more so if dynamic, e.g. shape changing objects, are planned.

Regarding its impact on process and output, treatment of the data is the most crucial and delicate part of the process. Moreover, this step leads to core decisions in the final representation and thus, is *part of the creative work*: The data needs to be curated, not just to fit the used technology and materials, but even more so for aesthetic or narrative reasons. Taking decisions in curating (selecting, filtering, and highlighting) data,

Figure 3: '*Drum Roll*'. Installation view. An automatically played drumroll, representing real-time radiation readings. [16]

Figure 4: '*Drum Roll*' detailed view of the drumstick mechanics. The data was translated into speed and intensity of the drumroll.

TDRM	Cluster	Location	ID	Current	Trend	Max	Average	Min	Std Dev	Temp	state
1	Doel	Kloetinge, NL	TDRM.32	0.087	-0.005	0.089	0.078	0.059	0.007	38.1	ok
2	"	Bergen oZ, NL	TDRM.36	0.111 (1)	-0.014	0.122	0.102	0.081	0.011	46.2	ok
3	"	Meerdonk	TDRM.14	0.15	-0.037	0.333	0.215	0.05	0.078	36.5	ok
4	"	Antw/Berchem	TDRM.12	0.203	-0.018	0.261	0.215	0.161	0.024	39.2	ok
5	"	Schoten	TDRM.2	0.134	0.009	0.149	0.13	0.111	0.011	40.1	ok
6	Brussels	Mechelen	TDRM.22	0.113	-0.004	0.123	0.109	0.093	0.008	31.1	ok
7	"	Center	TDRM.25	---							offline
8	Tihange	Celles	TDRM.7	0.19	0.023	0.219	0.181	0.127	0.026	31.1	ok
9	"	Marneffe	TDRM.16	0.077	-0.001	0.094	0.063	0.035	0.017	31.6	ok
10	"	Fumal	TDRM.13	0.11	-0.003	0.153	0.094	0.051	0.023	42.4	ok
11	"	Wanze	TDRM.10	0.126	-0.015	0.165	0.135	0.116	0.013	41.2	ok
12	"	Amay	TDRM.30	0.111	-0.005	0.127	0.114	0.101	0.006	32.2	ok
13	"	Nandrin	TDRM.34	0.106 (1)	-0.014	0.123	0.108	0.09	0.009	32.7	ok
14	Liege	St.-Walburge	TDRM.15	0.074	0.004	0.109	0.063	0.023	0.027	35.4	ok
15	"	Guillemins	TDRM.4	0.077	-0.015	0.124	0.108	0.077	0.012	27.2	ok
16	"	Tilff	TDRM.18	0.141	0.001	0.147	0.13	0.109	0.01	32.2	ok
17	"	Comblain a/Pt	TDRM.35	0.226 (4)	-0.019	0.257	0.233	0.209	0.012	31.1	ok

Figure 5: Excerpt of the TDRM Dataset. Too many data points to be conveyed by the installation. Curational treatment of the data needed to be done.

without risking a biased or even wrong representation of the data, is a non-trivial task.

Case Studies

Drum Roll is a data physicalization, featuring an automated drum. The drum plays a constant drumroll based on real-time radiation sensor readings of the Tihange Doel Radiation Monitoring network (TDRM) [8], which consists of 25-30 sensor nodes located in the geographic region around Belgian nuclear powerplants (Fig. 3). The TDRM provides data such as current readings in μSv/h, average, standard deviation and trend (Fig. 5). The drum roll was chosen as metaphor for the tension and threat of a possible nuclear accident. The drum itself as an automatically played mechanical instrument offered only two dimensions for conveying data: frequency and intensity (loudness) of the drumroll (Fig. 4). This defined the possible output, so the task was how to reduce, fit and funnel the given data to the installation. After several approaches, we decided on using the real-time input of two sensors at a time, with their actual value as input for the drum; mapping one sensor value to speed (frequency) of the drumroll and the other to its intensity (see side bar).

Bellum omni contra omnes—the war of all against all—uses data from a cybersecurity company website which shows a real-time "threat map" of ongoing cyberattacks [9]. Even though this is a questionable data source, our student group decided to work with this data feed. The data covers 193 nations, with information on their attack state. Since this would be impossible to build, the 12 countries with the highest activity were identified. This resulted in a setup of 12 entities, which interact in a mechanical performance of targeting and attacking each other (Fig. 6, 7, see next page). The data is conveyed by a mechanical change of orientation (who is attacking whom) and a follow-up affirmative movement (the 'shot') of the entities. Besides the pure number of data points, the choice for representing attacks as a 'pointing at' movement required a reduction, since we wanted it to be legible who was 'shooting' at whom.

The sixth Wave is a data physicalization on animal extinction. It shows the decline of selected animals in a 'pixel-based' 4x4 matrix setup, where the height of floating balloons displays the values at a certain timepoint (Fig. 8). The given dataset [10] consisted of >10k animals with their respective numbers of populations over the last 60 years. All balloons start at the same height and then subsequently move according to their population data over a defined period.

In all three projects a radical reduction of data had to be done. Even if the reasons to do so were inherent to each project, they all relate to general constraints, such as budget, feasibility, maintainability and also legibility of the resulting representation. This reduction can be described as curation of data, with a focus on the following questions: To how many data points can we reduce it? Which data points are representative for the entire dataset? Which are the most salient?

Discussion

In all cases, finding the data and appropriate ways of dealing with it, took a big part in ideation and building. The data had to be selected, filtered and reduced so it could be represented. Based on this experience and with this perspective in mind, we revisited other projects from sources like [11]. As many of these show

Figure 6: '*Bellum omni contra omnes*', installation view. Each entity interacts with the others on the field to represent cyber-attacks worldwide. [17]

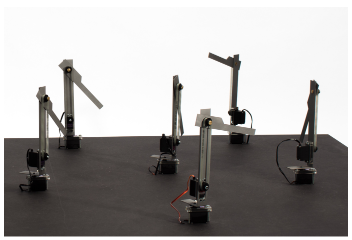

Figure 7: '*Bellum omni contra omnes*', detail on the mechanical states of the representing entities

Figure 8: '*The Sixth Wave*'.
16 balloons represent population data of endangered animals.

complex contexts in a simplified manner [e.g. 14, 15] we believe the creators had to make similar decisions.

Data physicalization tends to have a highly subjective component due to the need to select/transfer/process/translate data. Conceptualizing and building physical data representations entails dealing with constraints for what can be build and represented. Based on our experiences, we compiled a number of recommendations for creating data physicalizations.

Finding Datasets
Finding and choosing a suitable dataset is challenging. There are many accessible (online) sources. From Governments to NGOs to private entities [e.g. 12,13].

Reduction + Aggregation
For meaningful data physicalization, reduction is key. Thus, the most salient data points need to be selected or highlighted, without altering the data too much (so it is still a valid representation). Sometimes data can be aggregated (e.g. interpolation over time) as well.

Focus (on One Message)
Even if the dataset allows multiple results or messages, focusing on one is important. Else it may get difficult to convey meaning and legibility suffers. This also helps to narrow down which data points are important enough to be represented.

Curation Needs Time
Curation of data takes time and effort, potentially even more so than in visualization. The material and technology will introduce constraints to the data filtering/curation process. Therefore, this step tends to be a highly iterative process, instead of following a linear processing pipeline.

Non-Pixel Designs
Many recent examples of data physicalizations are based on a 'pixelated' setup with arrays of hundreds of actuators to realize universal displays [e.g. 5]. The possibilities of such technology are spectacular and tempting. With the effect, that every issue, error or step multiplies by the number of actuators, and so does the impact on production and budget. Can we find non-pixel representational strategies?

Expect Noise
Physicalizations cannot be controlled, nor can its presence be defined and restricted as clearly as with visualizations on a screen. Motors, valves, solenoids all make noise and cause vibrations. Varnish smells. Lighting is never perfect, perspectives vary. There is no clean, truly monomodal conveying of data.

Finger Exercises.
Designers and researchers new to the topic benefit from finger exercises [e.g. 6]. These help to understand physicalization principles, and to envision and explore the design space and material capabilities. But remember that such exercises benefit from data sources that are already well curated.

Finally - Future research
Future research should investigate how data sculpture artists and designers [e.g. 14,15] decide on which data to use and how to select it. Designers and artists have lots of experience in developing data sculptures and physicalizations. How do they select and curate data, how do they find appropriate representation strategies?

References

[1] A. V. Moere, "Beyond the Tyranny of the Pixel: Exploring the Physicality of Information Visualization," 2008 12th International Conference Information Visualisation, London, 2008, pp. 469-474.

[2] Yvonne Jansen, Pierre Dragicevic, Petra Isenberg, Jason Alexander, Abhijit Karnik, et al.. Opportunities and Challenges for Data Physicalization. CHI 2015 - Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Apr 2015, Seoul, South Korea. 2015.

[3] Yvonne Jansen, Pierre Dragicevic, and Jean-Daniel Fekete. 2013. Evaluating the efficiency of physical visualizations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’13). Association for Computing Machinery, New York, NY, USA, 2593–2602.

[4] Yun Wang, Adrien Segal, Roberta Klatzky, Daniel F. Keefem, Petra Isenberg, Jörn Hurtienne, Eva Hornecker, Tim Dwyer, Stephen Barrass. An Emotional Response to the Value of Visualization. In IEEE Computer Graphics and Applications (Volume: 39 , Issue: 5 , Sept.-Oct. 1 2019)

[5] Sean Follmer, Daniel Leithinger, Alex Olwal, Akimitsu Hogge, and Hiroshi Ishii. 2013. InFORM: dynamic physical affordances and constraints through shape and object actuation. In *Proceedings of the 26th annual ACM symposium on User interface software and technology* (*UIST ’13*).

[6] Samuel Huron, Pauline Gourlet, Uta Hinrichs, Trevor Hogan, and Yvonne Jansen. 2017. Let’s Get Physical: Promoting Data Physicalization in Workshop Formats. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (*DIS ’17*).

[7] Kenneth Moreland, 2012. A Survey of Visualization Pipelines. IEEE Transactions On Visualizations And Computer Graphics, Vol. 19, No. 3, March 2013

[8] https://tdrm.fiff.de/ last accessed: February 18, 2020

[9] http://www.livingplanetindex.org/documents/LPR2018data_public.csv Retrieved February 15, 2020

[10] https://threatmap.fortiguard.com/, accessed: March 1, 2020

[11] http://dataphys.org, last accessed: February 18, 2020

[12] https://data.europa.eu/euodp/en/data, last accessed March 1, 2020

[13] https://Unicef.org/reports, last accessed March 1, 2020

[14] http://dataphys.org/list/barrass-physical-data-sonifications/, last accessed February, 25, 2020

[15] https://www.stefanieposavec.com/airtransformed, last accessed February, 25, 2020

[16] Fig. 3 & 4 image credit: Leoni Fischer

[17] Fig. 6 & 7 image credit: Pauline Temme, Anas Alnayef, Sujay Shalawady