Structuring and Exploring User Behavioral Patterns in Social Media Traces

Eelco Herder
Radboud Universiteit Nijmegen
Institute for Computing and Information Sciences
Nijmegen, The Netherlands
eelcoherder@acm.org

Daniel Roßner
Hof University
Institute of Information Systems
Hof, Germany
daniel.rossner@iisys.de

Claus Atzenbeck
Hof University
Institute of Information Systems
Hof, Germany
claus.atzenbeck@iisys.de

ABSTRACT

User behavior and the resulting behavioral data forms the basis of personalized feeds, recommendations and advertisements in social networks such as Facebook. These platforms are now required to provide users with their personal data. However, these dumps with chronological data in different files do not provide users insight in overarching themes and connections in their online behavior. In this paper, we discuss the development and preliminary evaluation of an exploratory interface for visual data exploration. First insights include that the less obvious, more associative and obscure connections are more interesting and relevant to the user than very close semantic or temporal connections.

CCS CONCEPTS

• **Human-centered computing** → **Hypertext / hypermedia**; **Graphical user interfaces**; **Empirical studies in HCI**.

KEYWORDS

spatial hypertext; social networks; user profiles; GDPR; data exploration; transparency

1 INTRODUCTION

Social networks and social media are widely spread. According to the *Digital 2020* report[1] there are 2.5 billion active Facebook users, which makes Facebook the world's most-used social platform, followed by YouTube with 2.0 billion users. The social interaction and communication of these social media users create significant opportunities for the platform provider. For example, Facebook can analyze people's posts, behavior, or habits, trace users' click histories on public websites or identify patterns on various other user data. Information about its users enables social media platforms to issue ads specific to the individual user.

The extent of a company's benefits become clear when looking at Facebook's annual financial report[2], which states that 98 % of its 2019 revenue is based on advertising, a total of 69.5 billion USD. This is an increase of 27 % compared to 2018 or 74 % compared to 2017.

There is a conflict of interest between collecting information about users (possibly by combining various sources such as Facebook and WhatsApp, both platforms owned by Facebook, Inc.) and the users' right of privacy. The same is true for third-party companies crawling social media data for the same purpose [5]. Unexpected results in personalized advertisements have been noticed, which suggests that information of various sources have been used by the social media providers [12].

Increasingly, governments introduce regulations that protect citizens against the company's market power or lobbyists. The European Union introduced the General Data Protection Regulation (EU) 2016/679 (GDPR) [9]. This opens, among others, the right to EU citizens to request their data from platform providers.

Even though users may receive their data in human readable format, it is still difficult to explore it in a meaningful manner and to reach a high awareness of the personal data that is stored and computed by the platform provider's algorithms. In most cases, the *connections* between data snippets are more of interest than the data itself. Thus, a tool is required that can be used for exploring the various connections and associations between data.

In this paper, we discuss the development and preliminary evaluation of our system Mother as a hypertext tool for exploring social media data [11]. The goal is to increase the users' awareness of what data and, more importantly, which associations between information units the platform provider stores, and which insights it may algorithmically deduct from ongoing, reoccurring or co-occuring terms and topics as well as from chronological relations. We will use Facebook data dumps as an example for social media platforms.

The paper is structured as follows. In Section 2 we briefly discuss relevant related work on personal online data and present Mother's spatial hypertext tool as an explorative UI for such data. Section 3 describes the steps from a Facebook dump to its visual exploration in greater detail. Section 4 includes an exploratory evaluation, including the used methodology and our findings. Finally, Section 5 concludes this paper, raises open research questions, and refers to future work.

[1] See https://datareportal.com/reports/digital-2020-global-digital-overview, data updated to 25 January 2020

[2] FORM 10-K for Facebook, Inc., *Annual Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934 for the fiscal year ended December 31, 2019,* http://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/45290cc0-656d-4a88-a2f3-147c8de86506.pdf

MuC'20 Workshops, Magdeburg, Deutschland

https://doi.org/10.18420/muc2020-ws111-343

2 BACKGROUND AND ANALYSIS

2.1 Personal data, the GDPR and interpretability

As argued in the introduction, personal data is used by web platforms for a variety of purposes, varying from personalization and recommendation to monetization, for example via advertisements and nudges to continue visiting the platform [10]. Both real and perceived discrepancies between the use of personal data for the benefit of the end user on the one hand, and the use of the same data for monetization has led to several privacy concerns. A particular concern is the interpretation of data into, among others, user interest profiles, beliefs and demographics, consumer behavior or even health status [1].

The introduction of the General Data Protection Regulation, the GDPR [9], in Europe has led to several restrictions in which data can be used by industry and researchers alike and provides end-users with means for requesting transparency. Following the principles of responsibility, explainability, accuracy, auditability and fairness, several initiatives for responsible (HCI) research have been proposed [20]. A further opportunity that the GDPR offers to the scientific community is the result of Article 20, the "Right to data portability"[3], which states that:

> "The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format [...]"

Recently, researchers investigated user expectations and practices regarding the GDPR Right to Data Portability by asking users to request their data from the German loyalty program Payback. It turned out to be "unexpectedly simple and uncomplicated" to request the data, but participants believed that the data did not "paint the complete picture". Particularly, Payback did not provide any derived data, such as profiling or classification [2].

2.2 Spatial hypertext as a visual data exploration UI

Dumps of personal data or posts from social media platforms may be voluminous. Therefore, they generally need to be teared apart into smaller coherent informational units and associated in a semantically meaningful way. The result is a weighted undirected graph in which the edges represent the associations between information units. We will describe the full process in Section 3.

A naive way for displaying this data would be graph visualization, for example, using frameworks like Graphviz[4]. However, there may be too much information with many relevant associations in between to be displayed. Considering the full graph would result in a visualization that is hard to read and, thus, of low relevance for the users who wants to explore their own data. For this purpose, an iterative process is required that also can be found in *visual analytics*: "The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data." [14] The underlying mantra reads: "Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand" [13].

From its very beginning [6, 8, 16], hypertext always has been been considered a medium for representing human associations, a medium for humans to express their (interconnected) ideas. Beside navigation hypertext, i.e., those based on a node–link paradigm, there have been other types discussed in the past, including taxonomic hypertext [17] or argumentation supporting structures [7].

A special type is *spatial hypertext* [15], which follows a *desk-on-table* metaphor: similar to physical paper notes on a desk, objects can be moved on a 2D canvas. Associations between nodes are encoded by their position, arrangement, distance, size, color, orientation, or other visual cues. The associations are implicit; they appear by interpreting the space. In order to make the system "understand" the user created structure, so-called parsers compute associations based on spatial distance ("spatial parser"), visual appearance ("visual parser"), temporal user interaction ("temporal parser"), or content similarities ("content parser") [18].

The big advantage of using spatial hypertext is that structures can be created and modified at ease and at a low cognitive load [19]. As such, spatial hypertext helps in exploring unknown knowledge by supporting creating, modifying, or destroying contexts of information to which the system reacts to.

Our system *Mother* is a component-based open hypermedia system (CB-OHS) [3]. It provides the infrastructure of multiple knowledge bases, includes structure services (in particular the spatial structure service including its various parsers), and some UI that runs on various devices.

Mother consists of three basic layers [4]: (i) *Hel,* which includes all *knowledge-based components*; (ii) *Asgard,* which hosts Mother's *structure components*; and (iii) *Midgard,* which includes all *user interfaces* or components that have similar functionalities or purposes.

With respect to the application domain of exploring social media data, Mother provides: (i) a graph-based structure with information units taken from the Facebook dump and connected with weighted edges; (ii) a spatial hypertext UI that allows users to create contexts as a result of their exploration of data; and (iii) the parsers that compute a weighted graph from the implicit information given by the user which is used for querying the knowledge graph.

3 FROM RAW DATA TO VISUAL EXPLORATION

3.1 Description of the Facebook data dump

Following the GDPR regulations, Facebook allows its users to obtain a copy of their personal data as a simple download[5], either in (human-readable) HTML format or in (machine-readable) JSON format. In both cases, the user receives an archive with files that contain, among others, the user's own posts, comments, likes and reactions to posts by others (both friends and Facebook pages), search history, lists of friends, subscribed groups and pages, and interaction with advertisements[6].

Upon first inspection, it becomes apparent that the provided user data is strictly limited to the data provided by this specific user:

[3] https://gdpr-info.eu/art-20-gdpr/
[4] https://www.graphviz.org
[5] https://www.facebook.com/settings?tab=your_facebook_information
[6] For a full overview, see https://www.facebook.com/help/930396167085762?helpref=uf_permalink

users do *not* receive other users' comments or likes on their posts, nor do they receive the content of their friends' posts that they liked or commented on. As the social interactions between posts, likes and comments are an important ingredient of Facebook's algorithm[7], this implies that the data dump cannot be used for better understanding the inner workings of Facebook.

Still, the textual contents of the posts as well as post frequency statistics would provide rich material for users to obtain insight in and to reflect on their Facebook usage, including the reporting of life events, work-related announcements, discussions with friends, shared silly pictures or memes, and interaction with advertisements. However, particularly for active Facebook users, the lengthy, chronologically ordered lists of posts is not directly useful, as it does not allow users to recognize overarching themes and their content-related, associative and chronological connections.

3.2 Knowledge extraction

As a first step towards a visualization, we created a script to process the JSON Facebook post data of a user into a graph-based format, with posts, keywords, months and years as vertices, connected by edges with various weights. The keywords are extracted from the Facebook posts, converted into lowercase, lemmatized, stopwords removed and only keywords that appear in at least 5 posts are stored, in order to keep the number within limits. Edges between posts, years, months and keyword were created and weighted based on tf-idf and/or co-occurrence.

These vertices and edges served as a basis for experimenting with several configurations of word and post visualizations and their connections.

3.3 Exploratory UI

The principle of user interaction with Mother is that one selects a single entity (in our case, a Facebook post), which is then displayed along with *related entities* (other posts and keywords) as recommendations that can be followed in order to create a narrative. An example is shown in Figure 1.

The first developed application area of Mother concerned the movie domain, where movies are connected with one another through actors, genres and other entities [3]. These tight relations allowed users to discover and explore their own areas of interest. Similarly, our first visualization of the Facebook domain – making use of the authors' own Facebook profiles – recommended and displayed the posts that were content-wise closest to the selected post. Content-wise this approach made sense, but resulted in relations that were too obvious for the user (e.g., birthday wishes were related to other birthday wishes). After all, there is a difference between exploring an unknown domain, and *introspection*, the examination or observation of one's own mental and emotional processes; in this process, the most meaningful connections are the ones that are still meaningful, but not entirely obvious.

For this reason, we decided to only recommend posts with a content similarity of $0.6 \pm x$, with the remaining weight calculated by a sum of temporal similarity (post in the same hour of day, day of week, month), manually tuned and evaluated by the authors in several sessions. Furthermore, we also added the keywords as vertices and related them to the posts. Given the large differences in posting behavior, even between the authors, in terms of frequency and content, it was concluded that no optimal a-priori values could be found. Instead, we opted for a configuration that led to the first author's observation that he typically posted his reflections on a day in the early evening, along with some typical themes of these reflections; in addition, some randomness was added in order to prevent users to get locked in a small number of favorite themes.

[7] See, e.g., https://blog.hootsuite.com/facebook-algorithm/

4 EXPLORATORY EVALUATION

The development and iterative refinement of the Facebook post visualization already provided several insights in the type of relations that one would consider meaningful, interesting and relevant for introspection. As we are designing a solution for a range of foreseen user needs or wishes, good design science practice [21] is to have several iterations of design/development and (preliminary) evaluation.

In this exploratory evaluation, we are interested in finding out in which type of posts and terms users were interested and for which reasons: Would users typically try to confirm their most common patterns or themes, or would they try to discover new, surprising relations in order to better understand past events or past behavior? Would this behavior mainly be motivated by introspection and/or would users also aim to investigate possible privacy threats?

4.1 Methodology

Given the qualitative research questions, we chose an exploratory, scenario-based study setup with convenience sampling. Three participants, from the age of 27 to 35, were recruited and asked to provide us with their Facebook posts, which were used for the study and finally deleted, in order to prevent privacy and security issues. The evaluation itself involved the exploration of three themes (i.e., keywords given by the participant, representing hobbies, work-related announcements or life events). We invited them to select a post, and then to further explore related posts or keywords. No explicit time limit was set for the test.

After the evaluation, we asked the participants several open questions regarding their issues with Facebook data in general, to what extent the visualization would help to obtain answers with respect to these issues, which meaningful or surprising relations they discovered, and whether they would have any other question, wishes or ideas for the Facebook visualization.

4.2 Results

Our first participant had a history of 1096 Facebook posts, posted between 2009 and now, with an average of 9 posts per month. Most of these were short announcements of one or more photos (e.g., "wonderful small things"), as reflected by the low average number of 16.3 words per post. Consequently, the recommended related posts were typically based on an overlap of one or two often used phrases. Consequently, she could not relate to many of the displayed relations. She acknowledged that this was probably because most of her posts were visual and not so much textual. As points for improvement, she mentioned the inclusion of related places, persons, photos and links. Furthermore, she noticed – similar as the first author – the many birthday wishes in her log.

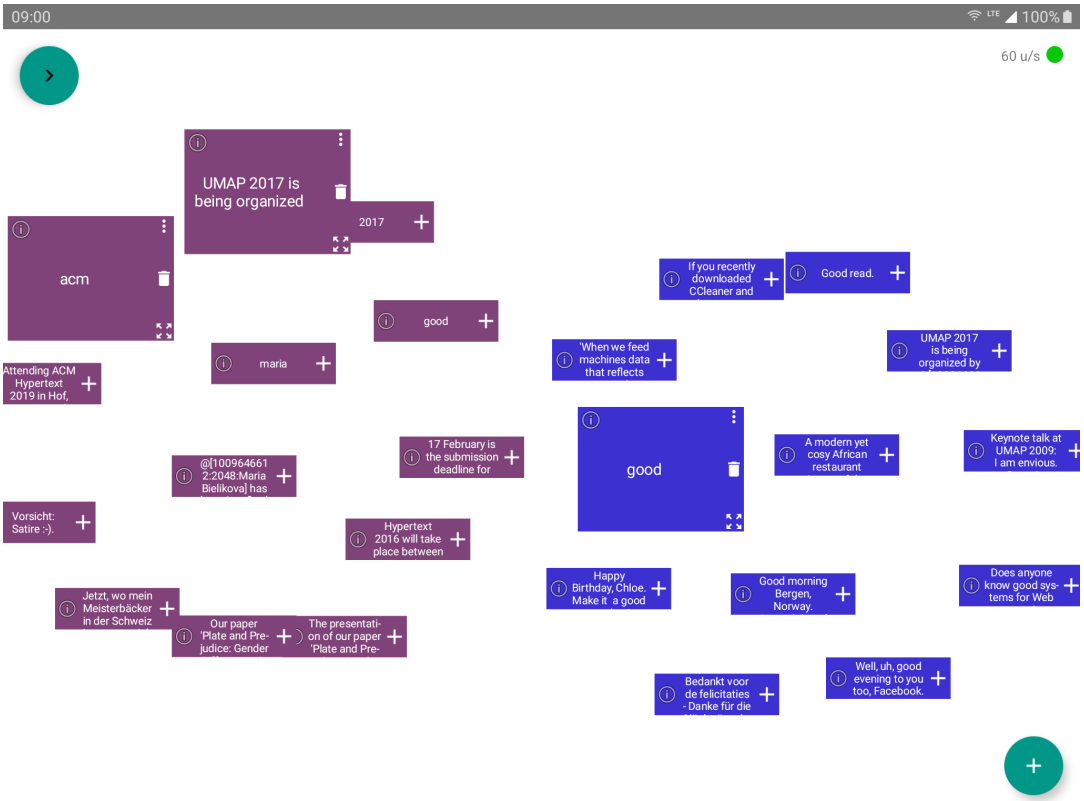

Figure 1: Screenshot of the prototype – one post and two keywords are added

The second participant had a Facebook log of 275 posts between late 2011 and late 2019, with 78 active months out of the 106 months in this period. The average post length was 25.6 words. In his posts, he mainly reported about activities of the youth organization that he is involved in. As these activities follow regular patterns and have regular relations, it was very easy for the participant to recognize the thematic clusters. However, there were also several posts that he couldn't remember having written and he also could not reconstruct when and why the post was written, the context remained entirely unclear.

The third participant found the visualization "exciting" and interacted with it for a long period. This participant had a history of 155 posts between July 2011 and now, with an average of 2.6 posts in each active month. With an average of 33.9 words per post, her posts were relatively long. This participant was interested to find out how her personal interests and writing style developed over time. Sometimes this led to interesting observations and explorations; she recognized a forgotten event in which she sold her study books. Some other relations remained unclear, which led to some frustration and the wish to be able to decide herself on the factors and weights for the post and word relations.

Apart from exploring one's post history, the tool also prompted participants to think about privacy-related questions, such as: what happens with a Facebook profile after one's death and where exactly are the personal data stored? Monthly statistics revealed that each participant's post behavior in terms of post frequency and average post length has remained stable and similar in the past decade.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented the adaptation of a spatial hypertext tool, Mother, in order to allow users to explore their Facebook post history. The post data can be downloaded by the users themselves, thanks to the data portability requirements of the European GDPR. Even though the users can directly scroll through the posts in chronological order, it does not allow for recognizing overarching themes or relations. Mother aims to fill this gap by providing navigation via recommended related posts, based on a combination of content similarity and temporal relations.

An important lesson learned during the design of the system and exploration of the author's own profiles is that there is a difference between exploring an unknown domain (e.g., movies) and introspective exploration of one's own activities and posts: for unknown domains, close semantic relations (such as actor X plays in movie Y) are meaningful and useful, but when exploring one's own activities, these relations turn out to be to obvious to be useful. Despite large differences in Facebook use in terms of post frequency, post topics and post types (e.g., short vs long, text-based or photo-based), a combination of semantic closeness (excluding the closest relations) and temporal similarity (including seemingly less obvious relations,

such as hour of day) delivered a rich domain that generally led to thematically meaningful clusters.

As future work, we aim to improve the interface, so that the posts are presented along with photos and links, if present. Furthermore, we will use similar approaches for exploring different parts of a user's Facebook profile: we expect that users' remembrance of their own comments to friends' posts and page contents is far lower than of their own posts, and that consequently introspection of commenting behavior will lead to more introspection on one's social media behavior and peripheral interests. Similarly, we think that a better understanding of one's interaction (comments, likes) with advertisements (or "suggested posts") will lead to more insight in which triggers one is sensitive to.

To summarize, even though the personal data provided by Facebook does not provide insight in the inner workings of profiling or recommendation algorithms, or ads practices, it does provide sufficient data for introspection and reflection on how individual users respond to the contents, suggestions and advertisements that Facebook provides, which is beneficial for one's own social media competencies and self-reflection.

REFERENCES

[1] Esma Aïmeur and Alexis Tremblay. 2018. Me, Myself and I: Are Looking for a Balance between Personalization and Privacy. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 115–119.

[2] Fatemeh Alizadeh, Timo Jakobi, Jens Boldt, and Gunnar Stevens. 2019. GDPR-Reality Check on the Right to Access Data: Claiming and Investigating Personally Identifiable Data from Companies. In *Proceedings of Mensch und Computer 2019*. 811–814.

[3] Claus Atzenbeck, Daniel Roßner, and Manolis Tzagarakis. 2018. Mother – An Integrated Approach to Hypertext Domains. In *Proceedings of the 29th ACM Conference on Hypertext and Social Media*. ACM Press, 145–149. https://doi.org/10.1145/3209542.3209570

[4] Claus Atzenbeck, Thomas Schedel, Manolis Tzagarakis, Daniel Roßner, and Lucas Mages. 2017. Revisiting Hypertext Infrastructure. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media 28th ACM Conference on Hypertext and Social Media*. ACM Press, 35–44. https://doi.org/10.1145/3078714.3078718

[5] Joseph Bonneau, Jonathan Anderson, and George Danezis. 2009. Prying Data out of a Social Network. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*. 249–254. https://doi.org/10.1109/ASONAM.2009.45

[6] Vannevar Bush. 1945. As we may think. *The Atlantic Monthly* 176, 1 (7 1945), 101–108. http://www.theatlantic.com/doc/194507/bush

[7] Jeff Conklin and Michael L. Begeman. 1987. gIBIS: a hypertext tool for team design deliberation. In *Proceedings of the ACM Conference on Hypertext*. ACM Press, 247–251. http://doi.acm.org/10.1145/317426.317444

[8] Douglas C. Engelbart. 1962. *Augmenting Human Intellect: A Conceptual Framework*. Summary Report AFOSR-3233. Standford Research Institute. http://dougengelbart.org/content/view/138

[9] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* 59, L 119 (5 2016), 1–149. http://data.europa.eu/eli/reg/2016/679/oj

[10] Eelco Herder. 2019. The Need for Identifying Ways to Monetize Personalization and Recommendation. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 291–294.

[11] Eelco Herder, Claus Atzenbeck, and Daniel Roßner. 2020. Hypertext as a Tool for Exploring Personal Data on Social Media. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*.

[12] Eelco Herder and Boping Zhang. 2019. Unexpected and Unpredictable: Factors That Make Personalized Advertisements Creepy. In *Proceedings of the 23rd International Workshop on Personalization and Recommendation on the Web and Beyond (ABIS '19)*. ACM, 1–6. https://doi.org/10.1145/3345002.3349285

[13] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual Analytics: Definition, Process, and Challenges. In *Information Visualization*, Andreas Kerren, JohnT. Stasko, Jean-Daniel Fekete, and Chris North (Eds.). Lecture Notes in Computer Science, Vol. 4950. Springer Berlin Heidelberg, 154–175. http://dx.doi.org/10.1007/978-3-540-70956-5_7

[14] Daniel Keim and Leishi Zhang. 2011. Solving Problems with Visual Analytics: Challenges and Applications. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM, 1:1–1:4. http://doi.acm.org/10.1145/2024288.2024290

[15] Catherine C. Marshall, Frank G. Halasz, Russell A. Rogers, and William C. Janssen. 1991. Aquanet: a hypertext tool to hold your knowledge in place. In *Proceedings of the 3rd ACM Conference on Hypertext*. ACM, 261–275. http://doi.acm.org/10.1145/122974.123000

[16] Theodor Holm Nelson. 1965. Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the ACM 20th National Conference*. ACM Press, 84–100. https://doi.org/10.1145/800197.806036

[17] H. Van Dyke Parunak. 1991. Don't link me in: Set based hypermedia for taxonomic reasoning. In *Proceedings of the 3rd Annual ACM Conference on Hypertext*. ACM Press, 233–242. http://doi.acm.org/10.1145/122974.122998

[18] Thomas Schedel and Claus Atzenbeck. 2016. Spatio-Temporal Parsing in Spatial Hypermedia. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. ACM, 149–157. http://dx.doi.org/10.1145/2914586.2914596

[19] Frank M. Shipman, J. Michael Moore, Preetam Maloor, Haowei Hsieh, and Raghu Akkapeddi. 2002. Semantics happen: Knowledge building in spatial hypertext. In *Proceedings of the 13th Conference on Hypertext and Hypermedia*. ACM Press, 25–34. http://doi.acm.org/10.1145/513338.513350

[20] Eva Thelisson, Kshitij Sharma, Hanan Salam, and Virginia Dignum. 2018. The General Data Protection Regulation: An Opportunity for the HCI Community?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–8.

[21] Roel Wieringa. 2009. Design science as nested problem solving. In *Proceedings of the 4th international conference on design science research in information systems and technology*. 1–12.