Data as Infrastructure – Infrastructuring for Data Analytics

Position paper for the

Infrastructuring in Participatory Design

Workshop at the PDC 2018.

Yvonne Dittrich, Cathrine Seidelin, Erik Grönvall
IT University of Copenhagen

The internet has long been a core infrastructure especially for the western societies. In the last decade, the usage of the internet changed from a mainly communication infrastructure giving mainly access to static information to become a platform to full-scale applications implemented as services. Even more recently we can observe a growing use of the internet to collect and provide data: social media platforms and search engines collect, use and sell data about their usage and users; online shops use data to recommend goods to the buyer; municipalities and other public organisations provide data through open data interfaces so it can be used by companies, e.g. to provide advice for car drivers searching for a parking lot. The data infrastructure making data accessible, though, requires IT expertise.
In this position paper we explore the implication of this development from a participatory design perspective and argue that we need to engage in infrastructuring in order to make the common data a common good.

Data as infrastructure
Google maps does not only provide information of how to get from A to B, but also uses data collected from service users to inform about traffic density and expected travel times so that commuters can select a different route or time for their trip; Amazon and other online shops use past data about shopping patterns to recommend books and goods for purchase; traffic management systems use historical and current data about traffic to regulate speed on highways. These examples have in common that data is not any longer (only) used to store and retrieve information about specific persons, situations or events, but as infrastructure based on which services for the users or the public are provided.
Similar, many municipalities and public agencies today use data to inform decision making: the most cited case might be New York City directing fire inspectors to addresses where fires with casualties are likely to happen (Hofman 2012). However, more and more municipalities make part of their data available as open data. Such open data can be anything from statistic data about socio-economic status of citizens, geospatial data about the physical infrastructures like streets and parks,

or data related to the usage of parking meters in order to enable companies to develop services based on these data sets, e.g. helping drivers to find a parking lot. (See the Copenhagen Open Data web site (2018) as an example.) "Mixing public data with commercial, civil society and citizen input data, and pooling and sharing with those produced by other public agencies and/or cities – i.e. data sharing for developing shared content, services and policies between cities– holds considerable potential for public value creation." (Ubaldi 2013, p.42) Like roads and other physical infrastructures, the data is regarded as an enabler for both municipal agencies, companies, and other societal actors, as part of the infrastructure a municipality provides for its citizens and enterprises.

Infrastructuring for Data Use

In order to make use of data as an infrastructure both social and technical structures need to be in place.

Examples for technical infrastructures are e.g. data portals like the CKAN (2018) allowing an organization to make data available for public use in a controlled and accountable manner. Another example for technical infrastructure development is the Danish DABAI project (2018) that has as one of its goals to provide Machine Learning as a Service with the ambition to enable domain experts to use Machine Learning algorithm to analyse their own data.

The social infrastructures are already mentioned in the above-cited OECD working paper on Open Government Data: "As intermediate consumers of data, infomediaries (e.g. media, developers, civil society) play an essential role in making sense of, and creating value out of, raw data. Media can tell interesting stories based on such data; developers can develop apps using them; civil society organisations can spot the relevance of certain data for specific segments of the population (e.g. charities in remote areas), can play a critical role to build capacities at the community level, and can create a culture that appreciates the relevance of the data." (Ubaldi 2013, p. 28)

An example

We study data-based service design and provisioning in cooperation with Industriens Uddannelser (English: Education secretariat for industry, hereafter the acronym IU is used), an education secretariat based in Copenhagen (Denmark). IU facilitates the collaboration between diverse labor market partners to develop educational programs for vocational training and adult vocational training in the industrial sector in Denmark. Data, both own data collected and maintained by IU in cooperation the involved labor market partners, and public data provided by schools and governmental agencies are at the core of this collaboration. See (Seidelin et al. 2017, Seidelin et al. 2018) for initial results. IU can be regarded as an 'infomediary' proposed in the above-cited OECD white paper (Ubaldi 2013). The research so far resulted both in understanding of the complexity of distributed data management and the challenges when it comes to the design and development of databased services, that is ways to make better use of data towards the objectives of IU.

Infrastructuring challenges for data infrastructures

We all know how to use streets and trains. If data becomes an infrastructure through which and based on which services are provided, we all need to learn how to 'drive on' data. To enable the

public to understand and use data and data-based services in a competent way requires both social and technical arrangements. Below we list the challenges we see so far.

Making data first-class citizen in the design process. Data has so far only been indirect part of the design process: E.g. rich pictures model the problem domain in object-oriented analysis as a basis for the design of the class diagram which describes what data is to be held (Mathiassen et al. 2000); data examples populate the interface mock-ups and prototypes (Blomberg et al. 1996, Lauesen 2005). When data becomes an infrastructure on and through which services are provided, data needs to become a design material in its own right. What are notations that allow domain experts to relate to data as a resource for services?

Making data readable. Data needs to be read and explored in order to understand its value for one-self and for others. We need ways to read, visualize, make sense of and be creative with data. This requires both technical infrastructures that allow to explore open and proprietary data sources without IT development expertise and social infrastructures like trusted infomediaries (Ubaldi 2013) that can help to locate and interpret available data sources.

Understanding and supporting (distributed) data management. One of the cases that motivated the notion of infrastructuring has been the sharing and distributed management of environmental data (Karasti & Baker 2004). To emphasise need to address the management of personal health related data, the term Human Data Interaction has been coined to describe the management (Crabtree & Motier 2015, Motier et al. 2014). Preliminary results of the IU case provide insights of what is needed in terms of data maintenance by an 'infomediary'. Especially the later makes the requirements for accountability of data management processes visible. What are social arrangements that need to be in place and how can they be supported technically?

References

Blomberg, J., Suchman, L., & Trigg, R. H. (1996). Reflections on a work-oriented design project. *Human-Computer Interaction*, *11*(3), 237-265.

CKAN Wikipedia entry (2018), https://en.wikipedia.org/wiki/CKAN, last retrieved May 15th. 2018.

Copenhagen Open Data (2018), https://data.kk.dk/, last retrieved May 15th. 2018.

Crabtree, A., & Mortier, R. (2015), Human data interaction: historical lessons from social studies and CSCW. in *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway*, Springer, 3-21.

DABAI (2018), https://dabai.dk/en/research-domains/case-2, last retrieved May 15th. 2018.

Hoffmann , L. (2012). Data mining meets city hall. Communications of the ACM, 55(6), 19-21.

Karasti, H., Baker, K. S. (2004), Infrastructuring for the Long-Term: Ecological Information Management. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 1 - Volume 1, 10020.3.

Seidelin, C., Dittrich, Y., & Grönvall, E. (2017). Identification of Data Representation Needs in Service Design. *Selected Papers of the IRIS, Issue Nr 8*. 10. http://aisel.aisnet.org/iris2017/10

Seidelin, C., Dittrich, Y., & Grönvall, E. (2018). Data Work in a Knowledge-Broker Organization: How Cross-Organizational Data Maintenance shapes Human Data Interactions. Accepted for the British HCI conference.

Lauesen, S. (2005). *User interface design: a software engineering perspective*. Pearson Education.

Mathiassen, L., Munk-Madsen, A., Nielsen, P. A., & Stage, J. (2000). *Object-oriented analysis & design* (Vol. 25). Aalborg: Forlaget Marko.

Mortier, R., Haddadi, H., Henderson, T., McAuley, D., & Crowcroft, J. (2014), Human-Data Interaction: The Human Face of the Data-Driven Society. *Social Science Research Network*. http://dx.doi.org/10.2139/ssrn.2508051

Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing. http://dx.doi.org/10.1787/5k46bj4f03s7-en