Camilla Lindelöw, Swedish School of Library and Information Science, University of Borås

Representing data in Library and information Science

Librarians have represented data (books and other documents) in the library catalogues for thousands of years. This was primarily done to find the books on the shelves. Eventually, the catalogues also became statistical sources, and during the 20th century scholarly indexes were used not only to find literature, but also to evaluate researchers, their productivity and impact (counting citations). With computer science, metadata became a term to describe data – data about data, and this was also the time when things started to be born digitally – a book born digitally is data with metadata describing it. Around the turn of the millennium, librarian and information scholars started to study library classification from a critical perspective and found various biases. Although synthetic data is not discussed that much in Library and Information Science (LIS), as far as I know, fake data is discussed. When metadata has gone from descriptive to evaluative in the context of scholarly production, the rise of fake publications, citations and references have been noted (Biagioli and Lippman, 2020).

Today LIS is a multidisciplinary field which is reflected in the discussion on representation. The information science branch has grown to become close to computer science, and more recently data science. Building on the Mathematical Theory of Communication (MTC) (Shannon and Weaver, 1998), information scientists have taken an objectivist approach to data – as data existing in the real world. This is apparent in the General Definition of Information (GDI) developed by philosopher Luciano Floridi. Here information is defined in terms of the data it is built from (Floridi, 2003). This data is further defined as a distinction, for example ”lacks of uniformity in the real world” (Floridi, 2010, p. 23), e.g. red light on a dark background. This is in line with Data Theory – “Data Theory examines how real world observations are transformed into something to be analysed – that is, data” (Jacoby in Lindgren, 2020, p. 21).

Opposed to this idea of the ”real world”, other branches of LIS emphasize the social construction of data. Jonathan Furner describes the tension in a speech: “…information science would be a whole lot better if people stopped thinking about it as a means to the end of understanding the relationship between people, information, and technology, and started thinking about it as a means to the end of understanding the various ways in which people interact with reality by creating and using representations of that reality” (Furner, 2020). Hope A. Olson’s influential book ”The Power to Name” represents this direction well (Olson, 2002): Olson examines two influential library profiles’ work, Dewey’s classification system (Dewey Decimal Classification, DDC) and Cutter’s rules cataloguing, which are a progenitor of the Library of Congress Subject Headings (LCSH). A feminist perspective is applied to investigate how women are represented in these systems. The tension between universality and diversity is clear and this results in a consistent marginalization and exclusion of *Other* women in these classifications built on a Western, male, heterosexual standpoint. Ronald E. Day writes about representing persons/researchers in citation indexes like the ones found in the Web of Science: ”Technological indexes are, then, not simply convenient tools for searching, but they are technical extensions of the self that then also reinforce the development of selves according to the contents of those indexes” (Day, 2014, p. 55). In this setting the self is

the researcher becoming an index through authorship, which then guides the personal recommendations given by the citation indexes.

This discussion on reality/representation played out nicely in my recent PhD mid-seminar where one of the senior scholars asked me about my quotes from Jacoby about translating reality into data – the senior scholar took an idealist standpoint: are we translating data into reality? I'm still pondering this. Turning now to something closer to synthetic data: I'm currently doing a study on researchers guessing the gender of other researchers. Gender is commonly a piece of metadata not included in the citation indexes. However, it is wanted in order to be able to do studies on gender differences in scholarly production and impact. Researchers thus use the today common technique of algorithms using the names to assign gender, since the names of researchers are usually present in citation indexes. But they also use more detailed methods such as assigning gender by photo and pronouns found on the web. These practices are discussed within the cataloguing community, where the idea of representing gender is problematised, not just because the methods used may be considered invasive, but also because of the current political climate concerning gender (Billey et al., 2014). A complete opposite view can be found in computer science, where I found statements such as the one about a user's refusal to reveal their gender making the inferring of it an interesting task for researchers. Thus, my paper finally discusses the ethical question of the rights to assign gender. Although I'm not sure gender guessing would be considered synthetic data, as it is still referring to individuals, I still believe my work provides some points for the further discussion on synthetic data, e.g. privacy.

References

Biagioli, M., Lippman, A. (Eds.), 2020. Gaming the Metrics, Infrastructures series. MIT Press.

Billey, A., Drabinski, E., Roberto, K.R., 2014. What's Gender Got to Do with It? A Critique of RDA 9.7. Cataloging & Classification Quarterly 52, 412–421. https://doi.org/10.1080/01639374.2014.882465

Day, R.E., 2014. Indexing it all : the subject in the age of documentation, information, and data. The MIT Press, Cambridge.

Floridi, L., 2010. Information: A Very Short Introduction. Oxford University Press.

Floridi, L., 2003. Information, in: The Blackwell Guide to the Philosophy of Computing and Information.

Furner, Jonathan. "Fundamental Research Questions in Information Science." In Information Studies and Other Provocations: Selected Talks 2000-2019, 2020.

Lindgren, S., 2020. Data theory: interpretive sociology and computational methods. Polity, Cambridge ; Medford, MA.

Olson, H.A., 2002. Power to name : locating the subject representation in libraries. Kluwer, Dordrecht ; Boston, Mass.

Shannon, C.E., Weaver, W., 1998. The mathematical theory of communication. University of Illinois Press, Urbana.