Data Stories in CLARIAH – Developing a Research Infrastructure for Storytelling with Heritage and Culture Data

Roeland Ordelman Willemien Sanders Richard L. Zijdeman

Rana Klein Julia Noordegraaf Jasmijn Van Gorp

Mari Wigham Menzo Windhouwer

ABSTRACT

Online stories, from blog posts to journalistic articles to scientific publications, are commonly illustrated with media (e.g. images, audio clips) or statistical summaries (e.g. tables and graphs). Such "illustrations" are the result of a process of acquiring, parsing, filtering, mining, representing, refining and interacting with data [3]. Unfortunately, such processes are typically taken for granted and seldom mentioned in the story itself. Although recently a wide variety of interactive data visualisation techniques have been developed (see e.g., [6]), in many cases the illustrations in such publications are static; this prevents different audiences from engaging with the data and analyses as they desire. In this paper, we share our experiences with the concept of "data stories" that tackles both issues, enhancing opportunities for outreach, reporting on scientific inquiry, and FAIR data representation [9].

In journalism data stories are becoming widely accepted as the output of a process that is in many aspects similar to that of a computational scholar: gaining insights by analyzing data sets using (semi-)automatized methods and presenting these insights using (interactive) visualizations and other textual outputs based on data [4] [7] [5] [6].

In the context of scientific output, data stories can be regarded as digital "publications enriched with or linking to related research results, such as research data, workflows, software, and possibly connections among them" [1]. However, as infrastructure for (peer-reviewed) enhanced publications is in an early stage of development (see e.g., [2]), scholarly data stories are currently often produced as blog posts, discussing a relevant topic. These may be accompanied by illustrations not limited to a single graph or image but characterized by different forms of interactivity: readers can, for instance, change the perspective or zoom level of graphs, or cycle through images or audio clips.

Having experimented successfully with various types and uses of data stories[1] in the CLARIAH[2] project, we are working towards a more generic, stable and sustainable infrastructure to create, publish, and archive data stories. This includes providing environments for reproduction of data stories and verification of data via "close reading". From an infrastructure perspective, this involves the provisioning of services for persistent storage of data (e.g. triple stores), data registration and search (registries), data publication (SPARQL end-points, search-APIs), data visualization, and (versioned) query creation. These services can be used by environments to develop data stories, either or not facilitating additional data analysis steps.

For data stories that make use of data analysis, for example via Jupyter Notebooks [8], the infrastructure also needs to take computational requirements (load balancing) and restrictions (security) into account. Also, when data sets are restricted for copyright or privacy reasons, authentication and authorization infrastructure (AAI) is required.

The large and rich data sets in (European) heritage archives that are increasingly made interoperable using FAIR principles, are eminently qualified as fertile ground for data stories. We therefore hope to be able to present our experiences with data stories, share our strategy for a more generic solution and receive feedback on shared challenges.

Acknowledgement. This work was made possible by the CLARIAH-PLUS project funded by NWO (Grant 184.034.023), and the PDI-SSH project SANE (https://pdi-ssh.nl/en/2021/11/funded-projects-2021-call/)

KEYWORDS

data stories, data sets, storytelling, digital journalism, digital humanities, enhanced publications

REFERENCES

[1] Alessia Bardi and Paolo Manghi. 2015. A framework supporting the shift from traditional digital publications to enhanced publications. *D-Lib Magazine* 21, 1 (2015), 1–1. https://doi.org/10.1045/january2015-bardi

[2] Andreas Fickers and Frédéric Clavert. 2021. On pyramids, prisms, and scalable reading. *Journal of Digital History* 1, 1 (2021).

[3] Ben Fry. 2008. *Visualizing data: Exploring and explaining data with the processing environment.* " O'Reilly Media, Inc.".

[4] Jonathan Gray and Liliana Bounegru. 2021. *Introduction.* Amsterdam University Press, 11–24. https://doi.org/10.1017/9789048542079.001

[5] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 1139–1148.

[6] Charles D. Stolper, Bongshin Lee, Nathalie Henry Riche, and John Stasko. [n. d.]. Emerging and Recurring Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Recent Stories. ([n. d.]). https://www.microsoft.com/en-us/research/publication/emerging-and-recurring-data-driven-storytelling-techniques-analysis-of-a-curated-collection-of-recent-stories/

[7] Wibke Weber, Martin Engebretsen, and Helen Kennedy. 2018. Data stories: Rethinking journalistic storytelling in the context of data journalism. *Studies in Communication Sciences* 2018, 1 (2018), 191–206.

[8] Mari Wigham, Liliana Melgar, and Roeland Ordelman. 2018. Jupyter Notebooks for generous archive interfaces. In *2018 IEEE International Conference on Big Data (Big Data).* IEEE, 2766–2774.

[9] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (2016).

[1] see e.g., https://stories.datalegend.net/ and https://mediasuitedatastories.clariah.nl/

[2] clariah.nl