Collections as Data: interdisciplinary experiments with KBR's digitised historical newspapers: a Belgian case study

Sally Chambers[1&2], Frédéric Lemmers[1], Thuy-An Pham[1],
Julie Birkholz [1&2], Vincent Ducatteeuw[2], Antoine Jacquet [1&3], Wout Dillen[4],
Dilawar Ali[5], Kenzo Milleville[5], and Steven Verstockt[5]

[1]KBR, Royal Library of Belgium
[2]Ghent Centre for Digital Humanities (GhentCDH), Ghent University
[3]Université libre de Bruxelles, Sciences de l'information et de la communication Department
[4]Antwerp Centre for Digital Humanities and Literary Criticism (ACDC), University of Antwerp
[5]Internet Technology and Data Science Lab (IDLab), Ghent University

Keywords: Collections as Data; Digital Cultural Heritage; Digital Humanities Datasets; Data-level access; FAIR (Findable, Accessible, Interoperable, Reusable) data; Open Science

Digital cultural heritage collections in libraries, archives, and museums are increasingly being used for digital humanities research. However, traditional ways of providing access to such collections, for example through digital library interfaces, are less than ideal for researchers who are looking to build datasets around specific research questions. Inspired by the 'Collections as Data' movement[1] as an approach for cultural heritage institutions to prepare their digital collections for analysis using digital methods, KBR, the Royal Library of Belgium, has embarked on a 24 month project[2] called DATA-KBR-BE (2020-2022) to facilitate data-level access to its digitised and born-digital collections for digital humanities research.

This paper will: a) introduce the concept of 'Collections as Data' while exploring the opportunities and challenges for cultural heritage institutions; b) outline the DATA-KBR-BE project and the methodology for piloting 'Collections as Data' at KBR; and c) present the preliminary results of initial experiments to extract thematic datasets in support of digital humanities research scenarios as a first step towards designing a sustainable data extraction workflow for KBR.

Originating in the United States, the '*Collections as Data*' data movement[3] was established to encourage cultural heritage professionals to start thinking differently about how they provide access to their collections to facilitate analysis using digital tools and methods. In its first phase, 'Always Already Computational: Collections as Data' (2016-2018)[4] focussed on exchanging

[1] See, for example: 'Always Already Computational: Collections as Data' and 'Collections as Data: Part to Whole'.
[2] DATA-KBR-BE is financed by the Belgian Science Policy Office (Belspo) as part of the Belgian Research Action through Interdisciplinary Networks, BRAIN 2.0 programme.
[3] You can request your 'Collections as Data' laptop sticker here: https://collectionsasdata.github.io/part2whole/logo/
[4] Always Already Computational: Collections as Data (2016-2018) was funded by the Institute of Museum and Library Services (ILMS).

experiences and sharing knowledge, as well as documenting this process (Padilla et al., 2019). In its current, second phase, 'Collections as Data: Part to Whole' (2019-2021)[5] supports the implementation and use of 'Collections as Data' through a number of funded collaborative case studies that are jointly led by cultural heritage professionals and researchers. Until now, implementations of 'Collections as Data' have largely occured in the United States (Wittmann, et al., 2019), but they are gradually appearing in Europe as well (Candela et al, 2020; Ames & Lewis, 2020).

DATA-KBR-BE (KBR, 2020) is an interdisciplinary research collaboration between cultural heritage experts, digital humanities researchers, and data scientists, which aims to optimise KBR's existing ICT infrastructure to stimulate sustainable data-level access to KBR's digitised collections for digital humanities research. Data-level access means providing access to the underlying files of digitised cultural heritage collections to facilitate data analysis by means of tools and methods developed in the field of digital humanities.

For the project, research teams in Ghent, Antwerp and Brussels are working closely together with the digitisation, collections and ICT experts at KBR to co-design three interdisciplinary research scenarios focused on KBR's digitised historical newspaper collection: BelgicaPress. These research scenarios are conceived as initial case studies to demonstrate the scientific potential of providing data-level access to KBR's collections, as well to understand how 'Collections as Data' could be implemented at KBR.

The interdisciplinary research scenarios that have been selected for the project are:

- **Collective Action Belgium,** led by GhentCDH, focuses on social history in the Interbellum and World War Two period and aims to trace the dynamics of contention, strikes, demonstrations and other forms of collective action in Belgium as reported in Belgian newspapers;
- **The feuilleton in Belgium**, led by ACDC, focuses on literary studies in the period 1830–1930 and aims to map the publication of literature in Belgian newspapers across the first century of the Belgian nation state; and
- **The History of Belgian Journalism**, led by ULB-KBR, focuses on media history from 1886 until now and aims to trace the history of Belgian journalism through the lens of critical discourses about journalism in Belgian newspapers.

The digital humanities research undertaken in this project collaborates closely with the KBR's Digital Research Lab, which facilitates text and data mining research using KBR's digitised and born-digital collections. Furthermore, the project will harness the expertise of data scientists for the semi-automatic extraction and classification of articles from historical newspapers (Ali & Verstockt, 2021).

In this paper we will present the results of initial experiments carried out during the first year of DATA-KBR-BE. This will include: a) recommendations from an interdisciplinary workshop

[5] Collections as Data: Part to Whole (2019-2021) is funded by the Andrew W. Mellon Foundation.

bringing together experts from across the different departments of the KBR, data scientists, and digital humanities researchers; b) the results from a follow-up workshop to co-design an initial 'test' DATA-KBR-BE dataset to support the interdisciplinary research scenarios; and c) the practical implementation and documentation of the extraction of the initial DATA-KBR-BE dataset.

The aim of DATA-KBR-BE is not only to kick-start the implementation of 'Collections as Data' at the KBR, but also to inspire other institutions in Belgium, the Benelux, and beyond to start experimenting with providing data-level access to their collections for digital humanities research.

Bibliography

Ali, D., & Verstockt, S. (2021). Challenges in extraction and classification of news articles from historical newspapers. Presentation at: What's Past is Prologue: the NewsEye International Conference, 16-17 March 2021.

Ames, S., & Lewis, S. (2020). Disrupting the library: Digital scholarship and Big Data at the National Library of Scotland. *Big Data & Society*. https://doi.org/10.1177/2053951720970576

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. Journal of Information Science. https://doi.org/10.1177/0165551520950246 and http://rua.ua.es/dspace/handle/10045/109460

KBR, Royal Library of Belgium (2020). DATA-KBR-BE: facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research. https://www.kbr.be/en/projects/data-kbr-be/

Oberbichler, S., Boros, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H. and Tolonen, M. (2021) Integrated Interdisciplinary Workflows for Research on Historical Newspapers - Perspectives for Humanities Scholars, Computer Scientists and Librarians. Submitted for publication for: *JASIST, Journal of the Association of Information Science and Technology*

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E. & Varner, S. (2019). *Final Report : Always Already Computational: Collections as Data*. http://doi.org/10.5281/zenodo.3152935 & https://osf.io/mx6uk/wiki/home/

Wittmann, R., Neatrour, A., Cummings, R., & Myntti, J. (2019). From Digital Library to Open Datasets. Information Technology and Libraries, 38(4), 49-61. https://doi.org/10.6017/ital.v38i4.11101