Originally published in: Lindén, Krister/Niemi, Jyrki/Kontino, Thalassia (Eds.): Proceedings of the CLARIN Annual Conference 2023. 16 – 18 October 2023, Leuven, Belgium. – Utrecht: Clarin, 2023. Pp. 53-56.



# Libraries as Data Infrastructures

**Martin Wynne**
University of Oxford, United Kingdom
martin.wynne@ling-phil.ox.ac.uk

**Andreas Witt**
Leibniz Institute for the German Language, Germany
witt@ids-mannheim.de

**Peter Leinen**
German National Library, Germany
P.Leinen@dnb.de

**Sally Chambers**
DARIAH-EU, Ghent Centre for Digital Humanities, Ghent University and KBR, Royal Library of Belgium
sally.chambers@dariah.eu

**Abstract**

The CLARIN and DARIAH European research infrastructures have a long history of collaboration and cooperation. One recent joint initiative has been to strengthen and deepen collaboration with national and major research libraries, with a particular focus on ways to facilitate the wider use of the extensive and culturally important digital datasets curated by libraries as research data. In order to further this goal, a series of workshops has been initiated, and a Conference of European National Librarians (CENL) Dialogue Forum has been established. Ongoing collaborative work includes a survey of existing collaborations between libraries and research infrastructures, an investigation of the potential for the creation of unique language models from digital library collections and an exploration of emerging initiatives such as the common European Data Space for Cultural Heritage.

## 1 Introduction

National Libraries have not only been pioneers in the development of data infrastructures, but they also play an essential role in facilitating research in the arts and humanities. Likewise, the continual growth of digital (digitised and born-digital) cultural heritage is crucial for arts and humanities researchers, especially for analysis and interpretation using digital methods (Tasovac et al, 2020). The digital data infrastructure landscape is currently in considerable flux, both nationally[1] and internationally[2]. Existing Research Infrastructures, such as DARIAH and CLARIN, are joining forces to contribute to the European Open Science Cloud (EOSC), for example through the establishment of the Social Sciences and Humanities (SSH) Open Marketplace[3]. In the cultural heritage space, emerging initiatives such as the common European Data Space for Cultural Heritage[4] and the European Collaborative Cloud for Cultural Heritage[5] are set to disrupt this landscape further, providing both challenges, as well as unprecedented opportunities for both libraries and research infrastructures alike. It is within this evolving context that the idea of a CENL Dialogue Forum on Libraries as Data Infrastructures was born.

[1] ESFRI National Roadmaps [https://www.esfri.eu/national-roadmaps]
[2] Strategy Report on Research Infrastructures Roadmap 2021 [https://roadmap2021.esfri.eu/]
[3] Social Sciences and Humanities Open Marketplace [https://marketplace.sshopencloud.eu/]
[4] Common European Data Space for Cultural Heritage [https://digital-strategy.ec.europa.eu/en/news/deployment-common-european-data-space-cultural-heritage]
[5] Collaborative Cloud for Cultural Heritage [https://ec.europa.eu/commission/presscorner/detail/en/IP_22_3855]

## 2 Conference of European National Librarians

CENL, the Conference of European National Librarians[6], brings together the National Libraries of Europe. It is a network of 46 national libraries in 45 European countries in the Council of Europe. Founded in 1987, the mission of CENL is to advance the cause of Europe's national libraries through collaboration to preserve the continent's cultural heritage and make it accessible to all, with a specific focus on skills and knowledge exchange. Collaboration between libraries and research infrastructures such as DARIAH and CLARIN is not new. As well as an active CLARIN and Libraries community, which holds regular workshops, DARIAH has been exploring the inter-relationship between digital collections and digital scholarship together with library organisations such as LIBER, Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries[7] and IFLA, International Federation of Library Associations and Institutions[8], and is an active participant in the International GLAM Labs Community[9].

To facilitate structural and strategic collaboration between Europe's National Libraries and Research Infrastructures, the idea of a CENL Dialogue Forum was born. It provides an ideal opportunity to assess the landscape; identify and prioritise specific challenges and opportunities, and understand how (national) libraries could benefit from structural collaboration with, and active participation in Research Infrastructures such as DARIAH and CLARIN. A key issue for debate is the international accessibility of FAIR (Findable, Accessible, Interoperable and Reusable) datasets and related challenges in implementation. Furthermore, the Collections as Data initiative is gaining traction internationally[10]. With the increasing emergence of 'data labs' throughout the library community, such labs could be an ideal point of intersection between the libraries, research infrastructures and digital humanities research communities. Not only could the Dialogue Forum be the voice of libraries in this data space, at the same time, it would raise awareness of this crucial topic throughout the (national) library community.

A survey of national libraries was carried out from May to July 2023 with the aim of obtaining a deeper understanding of existing and planned collaboration with research infrastructures, and to elicit information about activities in the areas of digital scholarship and digital data curation. Questions covered areas including compliance with the FAIR principles and Open Science, collections as data, data labs, data access, digital literacy, artificial intelligence and data science.

Thirty-one National Libraries responded to the survey, of which 20 (64%) institutions indicated that "Participating in Research Infrastructure" is a strategic priority of the National Library. Overall, the responses indicate that the percentage of institutions actively engaged in research infrastructure, or intending to do so, is 82%. There are 23 (74%) responding institutions already active in National Research Infrastructure initiatives.

Of the participating institutions, 14 (45%) state that they are participating in a European initiative, and a further 6 institutions are planning to do so. Thus, at the European level, Research Infrastructure appears to be a current topic for more than half of the participating institutions. DARIAH (10) and CLARIN (9) are the most frequently mentioned initiatives. When asked about the different stages of development of individual topics, the institutions responded as shown in Figure 1.

[6] Conference of European National Librarians (CENL) [https://www.cenl.org/]
[7] Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries (LIBER) [https://libereurope.eu/]
[8] International Federation of Library Associations and Institutions (IFLA) [https://www.ifla.org/]
[9] International GLAM Labs Community [https://glamlabs.io/]
[10] Collections as Data: State of the Field [https://collectionsasdata.github.io/part2whole/iac/]

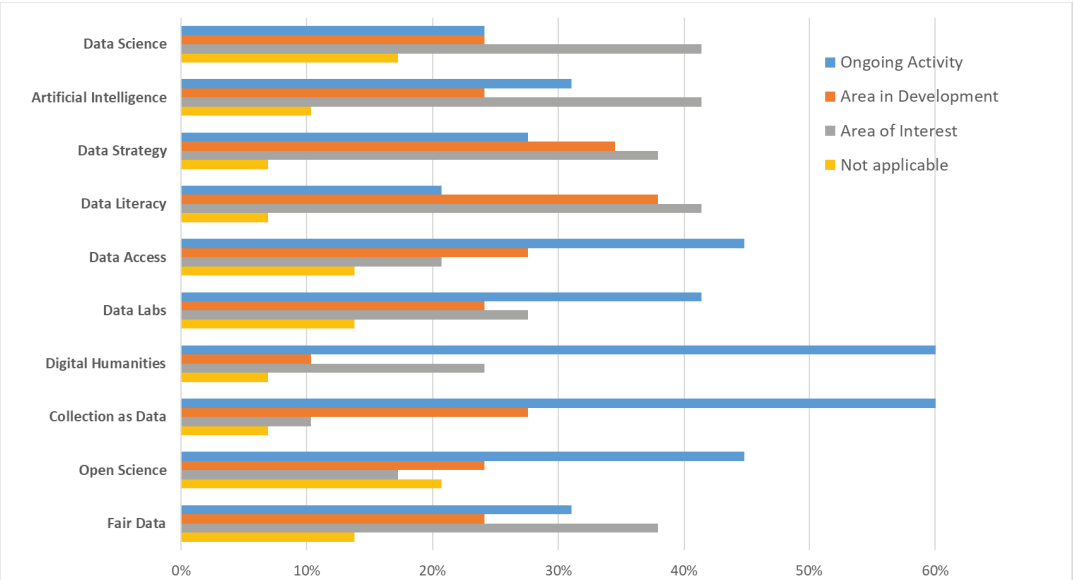


**Figure 1: Results of the CENL Survey**

## 3 Alignment of infrastructure projects

Together, the CENL Dialogue Forum, the ongoing series of CLARIN workshops and DARIAH's contribution to the common European Data Space for Cultural Heritage, will provide a platform for cooperation and collaboration, but will not directly achieve results in achieving interoperability and enhanced services for researchers. However, achieving interoperability and enhanced services for researchers will take place in libraries and national infrastructures, such as within specific projects:.

Text+[11] is a major new National Research Data infrastructure in Germany, involving a range of partners from academia, including CLARIN and DARIAH, and national and state libraries. The aim is to build a research data infrastructure focused on language and textual data, for a wide range of disciplines in the humanities and social sciences. The data which Text+ aims to deliver includes not only collections of historical texts, but also contemporary language corpora, lexical resources, and digital editions. Text+ will offer a comprehensive support infrastructure for all issues regarding collections, including interfaces, standards, authority data, and long-term preservation, etc.

DATA-KBR-BE[12] is a project at KBR, Royal Library of Belgium, which is developing an open data platform to offer data-level access to KBR's digitised and born-digital collections for digital humanities research. The project collaborates closely with the DARIAH and CLARIN consortia in Belgium, and builds on much recent and ongoing work in the area of 'collections as data'.

SSHOC-NL[13] is the latest in a series of joint CLARIN-DARIAH projects in the Netherlands, which will include the national library and national research institutes and which will build enhanced services for researchers, partly built on important past initiatives and collections such as Nederlab, Delpher and KB Lab Datasets.

Unlocking Digital Texts[14] is a collaboration between the Universities of Oxford, Cambridge and Notre Dame, with links to Text+ and Nederlab, which aims to make it easier to use a variety of textual

[11] Text+ [https://www.text-plus.org/en/home/]
[12] DATA-KBR-BE [https://www.kbr.be/en/projects/data-kbr-be/]
[13] SSHOC-NL Infrastructure awarded [https://www.huygens.knaw.nl/en/sshoc-nl-infrastructure-is-awarded-15-2-million-euros/]
[14] Unlocking Digital Texts [https://www.cdh.cam.ac.uk/research/projects/unlocking-digital-texts/]

formats as data in research. It will develop prototypes, and proofs-of-concept, building on existing standards (e.g. IIIF) and technologies rather than creating new formats or specific code dependencies.

The text digitization programme at the National Library of Norway has already created one of the largest text collections in the world and operates a DH-LAB that offers corpus services via a REST API and are also experimenting with the creation of language models based on the collections. Similarly, the National Library of France (BnF) Data Lab[15] is a service for researchers who wish to work with the BnF's digital collections.

Furthermore, there is the opportunity to align ongoing development in CLARIN online interfaces such as Korp, KonText, Corpuscle, NoSketchEngine etc. to more easily include library texts as datasets. Future development of the Virtual Language Language Observatory, Language Resources Switchboard and Federated Content Search could be optimised to work with more library collections and APIs.

## 4 Relevance to CLARIN

While research infrastructures for the arts humanities such as CLARIN and DARIAH have emerged in recent decades, for many centuries libraries have been the most important resource for researchers, and remain so in the digital age. For virtual, digital, distributed research infrastructures to be effective, they need to work closely with libraries, which play key roles as creators and curators of digital data, and as intermediaries between researchers and digital data, tools and expertise.

Creating language models from trusted and high-quality datasets is becoming an important area. Libraries not only offer access to large amounts of published material of known provenance and quality, but also unique opportunities to work with historical datasets, and thereby to create language models for a wider range of historical language varieties than is usually the case with existing research in the artificial intelligence, machine learning and natural language processing domains.

The ongoing collaborations will also provide an opportunity for libraries and research infrastructures to share knowledge and expertise, and potentially to share technical development work when it comes to user interfaces and APIs for sharing and using large text collections. Such collaborations should help to work against the inherent tendencies to waste effort, reinvent the wheel and create digital silos, and could become an important driver in the development and adoption of standards, common technological solutions and the interoperability of data collections and tools.

## 5 Conclusion

The initiative presented in this paper is intended to be an ongoing strategic collaboration, rather than a time-limited project, and will therefore inevitably be a work in progress rather than a completed discrete research activity. By the end of the summer of 2023, a number of activities currently underway, such as the CENL survey, will be completed, and others, such as the plan for the next CLARIN workshop, will be more firmly developed. However, given that a key aim is to include more participants and to promote openness and interoperability, it is important to disseminate the ongoing outputs as they happen, such as . promoting fruitful dialogue at the CLARIN Annual Conference.

## References

Tasovac, T. A., Chambers, S., Tóth-Czifra, S. (2020). *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*. ⟨hal-02961317⟩

[15] BnF Data Lab [https://www.bnf.fr/fr/bnf-datalab]