# CENL Dialogue Forum:

## National Libraries as Data Infrastructures: an activities update

**Peter Leinen, German National Library**
**Andreas Witt, Leibniz Institute for the German Language & University of Mannheim, representing CLARIN**
**Sally Chambers, DARIAH-EU, The British Library and formally KBR, Royal Library of Belgium**

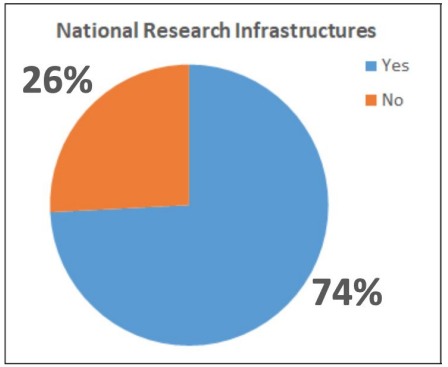
CENL

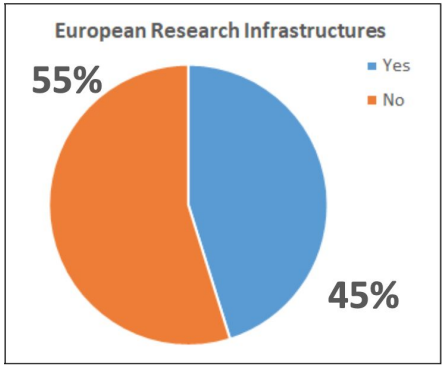
DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

CLARIN
Common Language Resources and Technology Infrastructure

# CENL Dialogue Forum:
# *National Libraries as Data Infrastructures*

- To facilitate structural and strategic collaboration between Europe's National Libraries and Research Infrastructures
- To identify and prioritise specific challenges and opportunities
- A survey of CENL member's institutions was chosen as the starting point for the activities




CENL

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

CLARIN
Common Language Resources and Technology Infrastructure

# Survey results: prioritising National Library needs

| | Ongoing Activity | Area in Development | Area of Interest | Not applicable |
|---|---|---|---|---|
| Fair Data | 31% | 24% | 38% | 14% |
| Open Science | 45% | 24% | 17% | 21% |
| Collection as Data | **62%** | 28% | 10% | 7% |
| Digital Humanities | **66%** | 10% | 24% | 7% |
| Data Labs | 41% | 24% | 28% | 14% |
| Data Access | 45% | 28% | 21% | 14% |
| Data Literacy | 21% | **38%** | **41%** | 7% |
| Data Strategy | 28% | **34%** | 38% | 7% |
| **Artificial Intelligence** | 31% | 24% | **41%** | 10% |
| Data Science | 24% | 24% | **41%** | 17% |

CENL

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

CLARIN
Common Language Resources and Technology Infrastructure

# CLARIN and Libraries 2023: Large Language Models and Libraries

Tuesday, 5 December 2023 , 12:00 - Wednesday, 6 December 2023 , 13:00

## About

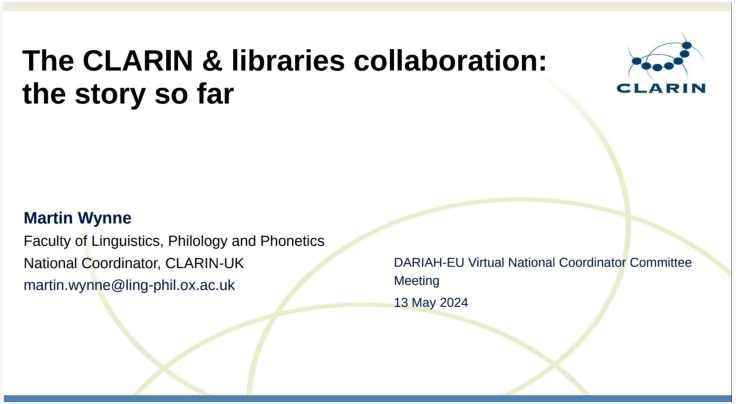
The workshop builds upon the first CLARIN and Libraries workshop held in the Hague in May 2022 (see here).


This year's workshop will investigate further areas of collaboration between CLARIN-related initiatives and libraries with a special emphasis on building (large) language models in and in cooperation with libraries. The workshop will bring together for the second time a group of people associated with both CLARIN (or other research infrastructures) and libraries. Whereas the first CLARIN and Libraries workshop was particularly concerned with digital content delivery for researchers, the main theme of the second workshop will be large language models and library collections, e.g. technical challenges in building such models and legal implications of model training and use.

The host, the National Library of Norway (NLN), has since 2005 digitised its entire text collections, amounting at present to a large corpus of 160 billion words for Norwegian and has built large language models for text (BERT, GPT-2, T5) and speech (wav2vec, Whisper) on these collections. There will be keynotes from the National Libraries of Norway and Germany on the technical aspects of building such models in a library setting, as well as a keynote on the legal aspects of building large language models from the Swedish National Library.

https://www.clarin.eu/blog/clarin-and-libraries-2023-recap

National Library of Norway




KBR


europeana

Belgian Science Policy Office
belspo

COLLECTIONS AS DATA

Народна библиотека Србије Београд

nuk

L N B
NATIONAL LIBRARY OF LATVIA

KB national library of the netherlands

DEUTSCHE NATIONAL BIBLIOTHEK

# Collections as Data: Collaborating across Data Spaces for Cultural Heritage and Open Science

**KBR | Royal Library of Belgium, 19th - 20th February 2024**

Österreichische Nationalbibliothek

BRITISH LIBRARY

Bibliothèque nationale du Luxembourg

16 40
KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

DET KGL. BIBLIOTEK
Royal Danish Library

BnF

## Collections as Data as a bridge between collections and infrastructures

The aim of this session is to better understand what the **key challenges to data sharing are** and **how initiatives such** as the common European Data Space for Cultural Heritage, Research Infrastructures such as DARIAH and CLARIN and the European Open Science Cloud (EOSC) **can facilitate this process.**

To do this, we asked a **number of national libraries to share their experiences of sharing data with us** …

# Diving Deeper into the Collections as Data Workflow

Home / Workflows / A workflow to publish Collections as Data: the case of

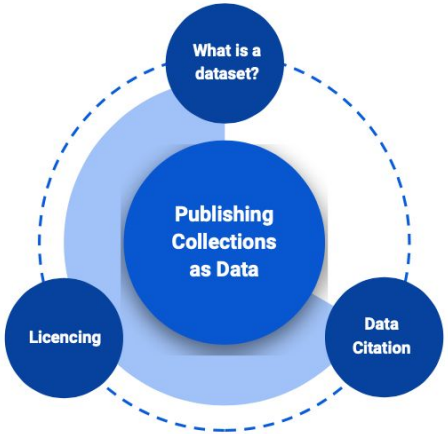
SSH Open Marketplace
Social Sciences and Humanities Open Marketplace

## A workflow to publish Collections as Data: the case of Cultural Heritage data spaces

**This is version 1 of the workflow, currently under community review for version 2.**

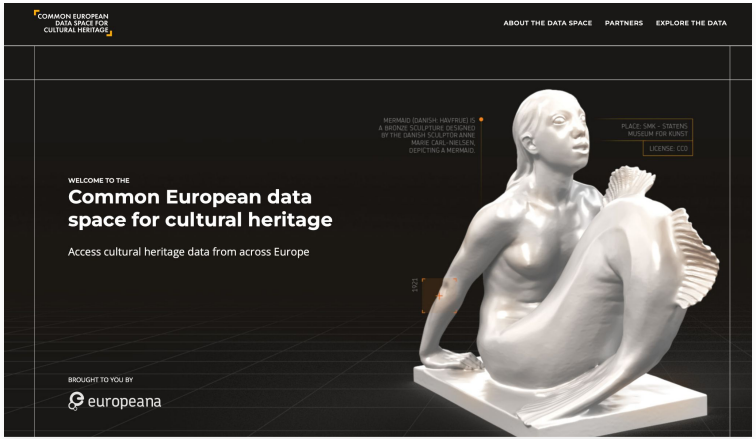
Cultural Heritage institutions have been making their digital collections available for the public for several decades. Advances in technology such as Artificial Intelligence and Machine Learning have provided a new context in which digital collections can be analysed using computational methods. Initiatives such as Collections as data and the FAIR data principles, have emerged to provide best practices and guidelines for publishing digital collections suitable for computational use. These initiatives are complemented with the CARE principles to strengthen ethical considerations in data governance and reuse. In parallel, experimental Labs have been implemented in Galleries, Libraries, Archives and Museums (GLAM) in order to reuse the digital collections.

## Diving Deeper session




**https://marketplace.sshopencloud.eu/workflow/I3JvP6**

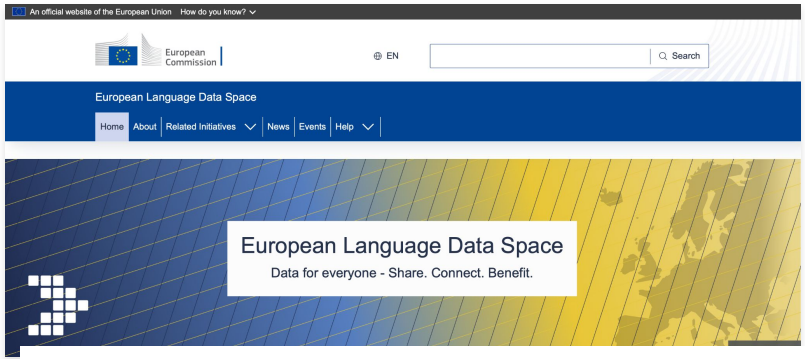
# Dynamic data infrastructure landscape



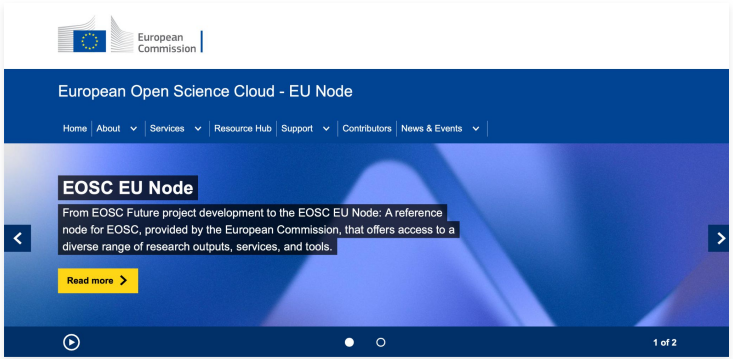

https://www.dataspace-culturalheritage.eu/en


**ECHOES (European Cloud for Heritage OpEn Science)**

https://www.msh-vdl.fr/2024/06/06/lancement-echoes/




https://language-data-space.ec.europa.eu/index_en




https://open-science-cloud.ec.europa.eu/

# Next Steps: ‘Collections as Data’ pilots

- **Develop 2-3 ‘Collections as Data’ pilots using the** workflow to publish Collections as Data: the case of Cultural Heritage data spaces
  - Harvesting cultural heritage datasets e.g. from existing repositories, e.g. BL Research Repository, SODHA, TextPlus Registry etc.?
  - Could Europeana galleries be optimised for research reuse?
  - Adapt the Europeana Data Model (EDM) to include ‘datasets’?
- **Share these ‘collections as data’ on other research platforms such as the** SSH Open Marketplace or the European Open Science Cloud.
- **Experiment with analysis platforms for large scale cultural heritage corpora**

Text+ Text+ Registry

KorAP