This is a peer-reviewed, accepted author manuscript of the following research article: Candela G, Gabriëls N, Chambers S, Dobreva M, Ames S, Ferriter M et al. A checklist to publish collections as data in GLAM institutions. Global Knowledge, Memory and Communication. 2023 Nov 9. Epub 2023 Nov 9. doi: 10.1108/GKMC-06-2023-0195

A Checklist to Publish Collections as Data in GLAM Institutions

Gustavo Candela[1], Nele Gabriëls[2], Sally Chambers[3], Milena Dobreva[4], Sarah Ames[5], Meghan Ferriter[6], Neil Fitzgerald[7], Victor Harbo[8], Katrine Hofmann[8], Olga Holownia[9], Alba Irollo[10], Mahendra Mahey[11], Eileen Manchester[6], Thuy-An Pham[3], Abigail Potter[6], and Ellen Van Keer[12]

[1]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain
[2]KU Leuven Libraries, Leuven, Belgium
[3]KBR, Royal Library of Belgium, Brussels, Belgium
[4]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
[5]National Library of Scotland, Edinburgh, Scotland
[6]Library of Congress, Washington, United States of America
[7]British Library, London, United Kingdom
[8]Royal Danish Library, Aarhus, Denmark
[9]International Internet Preservation Consortium, United States of America
[10]Europeana Foundation, The Hague, Netherlands
[11]Tallinn University, Tallinn, Estonia
[12]Meemoo, Ghent, Belgium

June 22, 2023

Abstract

Purpose. The purpose of this study is to offer a checklist that can be used for both creating and evaluating digital collections, which are also sometimes referred to as datasets as part of the Collections as data movement, suitable for computational use.

Design/methodology/approach. The checklist was built by synthesising and analysing the results of relevant research literature, articles and studies and the issues and needs obtained in an observational study. The checklist was tested and applied both as a tool for assessing a selection of digital collections made available by GLAM institutions as proof of concept and as a supporting tool for creating Collections as data.

Findings. Over the past few years, there has been a growing interest in making available digital collections published by GLAM organisations for computational use. Based on previous work, we defined a methodology to build a checklist for the publication of Collections as data. Our evaluation showed several examples of applications that can be useful to encourage other institutions to publish their digital collections for computational use.

Originality. While some work on making available digital collections suitable for computational use exists, giving particular attention to data quality, planning and experimentation, none of the work to date provides an easy-to-follow and robust checklist to publish collection datasets in GLAM institutions. This checklist intends to encourage small and medium-sized institutions to adopt the Collection as data principles in daily workflows following best practices and guidelines.

Keywords— Digital Libraries, Collections as data, Digital Collections, Metadata, Data Spaces, GLAM

A Checklist to Publish Collections as Data in GLAM Institutions

June 21, 2023

1 Introduction

During the past few decades Galleries, Libraries, Archives and Museums (GLAM) have provided access to some of their collections and materials in digital format either online or onsite and often including related metadata. Organisations have been exploring the benefits of adopting the concept and setting

up of experimental *Labs* to publish under open licenses where possible in order that digital collections may be reused in innovative and creative ways (Mahey *et al.* 2019). Advances in technology have paved the way to publish digital collections suitable for computational use through the initiative known as Collections as data (Padilla *et al.* 2019b). Furthermore, with the emergence of other initiatives such as the common European data space for cultural heritage (Europeana 2022) and the European Cultural Heritage Cloud (European Commission 2022), there is even a more urgent need to incorporate Collections as data activities into the day-to-day operations of cultural heritage institutions in combination with building the necessary services, capacities and changes in mindsets to proactively contribute to such initiatives.

Many GLAM organisations provide digital collections for computational use in several ways. For instance, the Data Foundry at the National Library of Scotland provides metadata and digitised collections using a CC0 license (National Library of Scotland 2021). The Library of Congress provides access to information about historic newspapers and selected digitised newspaper pages as JavaScript Object Notation (JSON), Linked Data and bulk data (Library of Congress n.d.[b]). The Bibliothèque nationale du Luxembourg provides access to a newspapers dataset with rich metadata using international XML standards such as Metadata Encoding and Transmission Standard (METS) and Analyzed Layout and Text Object (ALTO) (Bibliothèque nationale du Luxembourg 2021). These initiatives can encourage other GLAM organisations to publish their collections following specific guidelines and principles so that they are suitable for computational use when resources allow. However, as there is a wide diversity of approaches for publishing digital collections, organisations need consistent and quality guidance in selecting the best approach suited to their goals and, at the same time, to researchers' and other reusers' needs. Several aspects must be considered in terms of how datasets are made available, including metadata formats (e.g., MARCXML, Dublin Core, JSON, etc.), data and metadata cleaning, appropriate ways to provide access to files online or onsite and documentation about the datasets.

This paper defines a checklist that can be used for both creating and evaluating digital collections suitable for computational use that are published or made available by institutions in the GLAM sectors. This approach provides an easy-to-apply method to encourage small and medium-size organisations to publish their digital collections for computational use. The main contributions of this paper are: i) a checklist to create datasets and to assess their suitability for use with computational methods; ii) the application of the checklist; and iii) the results of the applications.

The paper is organised as follows: after a brief description of the state of the art in Section 2, Section 3 describes the methodology used to build the checklist. The application of the methodology and results are shown in Section 4. The paper concludes with an overview of the methodology and future work.

2 Related Work

The use of Artificial Intelligence and Machine Learning in the GLAM sectors has become an important topic aiming at applying new methods to the rich digital collections made available by the GLAM organisations (Cordell 2020; Padilla 2019; Padilla *et al.* 2019b; Strien *et al.* 2021). In this sense, new initiatives on advancing the use of Artificial Intelligence have emerged such as Artificial Intelligence for Libraries, Archives and Museums (AI4LAM) [1] and NewsEye [2]. Several aspects regarding data quality and transparency in terms of how the data is available for the public (e.g., license, format, access, etc.) have become crucial elements for researchers wanting to reuse the contents (Candela *et al.* 2021). Many organisations such as the Bibliothèque nationale de France, the British Library and the Rijksmuseum focus on the application of new and advanced technologies to their digital materials (Bibliothèque nationale de France 2021; Dobbs and Ras 2022; British Library 2019). In addition, organisations have explored the benefits and challenges of using Application Programming Interfaces (APIs) to make available their digital collections as well as advanced vocabularies to describe the metadata (Harvard Art Museums 2022; Museum of Modern Art 2022; Koho *et al.* 2021; Smith-Yoshimura 2020). Moreover, features such as data cleaning and enrichment, the use of expressive controlled vocabularies instead of traditional metadata formats, using advanced and widely used APIs and the use of common and known open licenses have become crucial to facilitate the reuse of the contents. These technological innovations are relevant to the efforts in building data spaces for cultural

heritage and the need to meet the needs of different types of users (Dobreva, Stefanov, and Ivanova 2022).

Despite all these efforts, there is still room for improvement and consistency regarding the publication of digital collections suitable for computational use (Candela *et al.* 2021). Adopting these new initiatives from scratch is difficult for organisations for several reasons, e.g., the absence of dedicated personnel, a limited budget or the lack of advanced technical skills.

In this context, a checklist can prove to be a powerful tool, as it presents a list of tasks, activities, and behaviours that need to be followed to achieve a systematic result. The creation of checklists has emerged as an innovative method to provide best practices and guidelines. Several initiatives have already discussed the definition and creation of checklists in other domains, for instance for the improvement of the reliability of artificial intelligence systems in terms of the life cycle (Han and Choi 2022) and the evaluation of software process line approaches (Agh, García, and Piattini 2022). Here, a checklist publication workflow was proposed including aspects such as source data management, reproducible data transformation, version control, data documentation and publication (Reyserhove *et al.* 2020). Other initiatives include a checklist for developing a machine learning project based on cultural heritage data (Lee 2022) or a checklist for a Data Management Plan (Digital Curation Centre 2013).

Regarding Collections as data at GLAM institutions, previous work has proposed a methodology to select datasets for computationally driven research applied to Spanish text corpora in order to encourage Spanish and Latin American institutions to publish machine-actionable collections (Candela *et al.* 2021). A compilation of actions that can be carried out to stimulate conversation, and to encourage and generate ideas and new possibilities concerning the publication of digital collections suitable for computational use was recently published (Padilla *et al.* 2019a).

The use of advanced technologies such as Artificial Intelligence in combination with rich data made available by GLAM organisations raised important ethical issues (Romein *et al.* 2022; Boyd, Keller, and Tijerina 2016). These include, for example, control over the data, including the terms of service requirements, the long-term subsistence of the organisation sharing the data, the anonymous release of data and the threat of potential reidentification, and awareness of potential uses of the data. While clearer guidelines and better coordination are needed (Padilla *et al.* 2019a; Boyd, Keller, and Tijerina 2016), libraries and universities are in the position of playing a crucial role in training and education concerning unknown and future ethical issues.

These efforts provide an extensive demonstration of how to make available digital collections suitable for computational use, giving particular attention to data quality, planning and experimentation. Nevertheless, to our best knowledge, none of the work to date provides an easy-to-follow and robust checklist to publish collection datasets in GLAM institutions. This checklist intends to encourage small and medium-sized institutions to adopt the Collection as data principles in their daily workflows.

3 A checklist to publish Collections as data in GLAM institutions

Making available digital collections suitable for computational use is a complex process. Examples in the literature follow different approaches, making it difficult to adopt and standardise the process. In this sense, institutions may face challenges when addressing the adoption of Collections as data due to the lack of expertise, guidelines and best practices. This section introduces the methodology that was used to create an easy to follow checklist to publish collections as data in the GLAM sector.

The checklist was constructed in four steps, which are described in further detail in the following sections: i) identifying relevant topics to support the publication of digital collections suitable for computational use based on existing implementations of the Collections as data principles and on a literature review (Section 3.1); ii) conducting a survey to identify potential issues and needs regarding how to make collections available as data from practitioners in GLAM and researchers (Section 3.2); iii) synthesising and analysing the results of relevant research literature, articles and studies and the issues and needs identified in the survey to create the checklist (Section 3.3); and finally iv) evaluating the checklist by applying it both as a tool for assessing a selection of datasets made available by GLAM institutions and as a supporting tool for developing Collections as data platform, to further improve the checklist prior to publication (Section 4).

3.1 Previous works based on data published by GLAM

The first step in creating the checklist is based on a literature review encompassing existing work on publishing checklists in different domains and data management plans, institutional reports from GLAM organisations about digital collection publication for the public, and projects based on the reuse of the digital collections with innovative and creative approaches. In addition, recent research articles were searched in repositories (e.g., ACM Digital Library and dblp) about the impact and reuse of digital collections in GLAM institutions. Appendix A shows the list of studies included in the review. The items were classified into five categories as shown in Table I.

Table I: Literature review to create the checklist classified into categories.

Category	References
Best practices	(Padilla *et al.* 2019b; Cordell 2020; Padilla *et al.* 2019a; Sherratt 2019; Candela *et al.* 2022b; Mahey *et al.* 2019; SmithYoshimura 2020; Godby *et al.* 2019; Padilla 2019; Gebru *et al.* 2021; Harris, Potter, and Zwaard 2020; Averkamp *et al.* 2021; Romein *et al.* 2022; Boyd, Keller, and Tijerina 2016)
Data quality	(Candela *et al.* 2021; Candela *et al.* 2022a; Király 2019; Strien *et al.* 2020)
Checklist definition	(Agh, García, and Piattini 2022; Digital Curation Centre 2013; Lee 2022; Han and Choi 2022; Reyserhove *et al.* 2020)
Strategy & data plan	(British Library 2019; National Library of Scotland 2019; Bibliothèque nationale du Luxembourg 2014; Bibliothèque nationale de France 2021; National and State Libraries Australasia 2022; Europeana 2020; Digital Preservation Coalition 2022; Research Libraries UK 2022; LIBER 2018; National Library of Scotland 2022)
Examples & experiments	(Dobbs and Ras 2022; Dijkshoorn *et al.* 2018; Museum of Modern Art 2022; Harvard Art Museums 2022; Koho *et al.* 2021; Candela and Carrasco 2022; Lee *et al.* 2020; Biblioteca Nacional de España 2020; British Library 2020; Bibliothèque nationale du Luxembourg 2021; Data Foundry 2020; Australian Cultural Data Engine 2022; Candela *et al.* 2018; Lorang, Soh, and Pack 2020; Jakeway *et al.* 2020; Royal Library of Belgium 2020)

The items classified as *best practices* include literature regarding guidelines to adopt Collections as data from different projects and authors, as well as recent approaches in publishing and reusing machine-actionable datasets. *Data quality* encompasses research articles discussing the assessment of datasets with various methods as well as considering different types of contents such as text and metadata. *Checklist definition* includes methodologies for creating checklists and examples applied to several domains. *Strategy and data plan* describes reports made available by large institutions and initiatives regarding digital change. The *examples and experiments* category introduces several examples of datasets and how they can be accessed and reused in innovative and creative ways.

3.2 Identifying issues and information needs when implementing the Collections as data principles

The second step in creating the checklist corresponds to an observational study regarding the knowledge about and uptake of the Collections as data principles in GLAM institutions using an online survey during the period 10-30 October 2022. Participation was voluntary and open to all interested GLAM institutions. Participants could provide the name of their institution and contact information or opt for anonymity. Consent was obtained from all respondents to include the survey results anonymously.

A first core set of questions aimed to understand the respondents' existing experience with Collections as data, including the issues encountered in the early implementation phases, and collect examples of datasets already published. A second core set of questions was included to identify to what extent the respondents felt sufficiently informed when starting to implement the Collections as data principles and to understand their information needs.

Table II shows the questions used in the form sent to the participants.

Table II: Survey employed to retrieve information regarding the publication of Collections as data in GLAM institutions.

Category	Question	Type
Introduction	Goal of the survey	-
Contact Information	Institution, email, name, etc.	Text
Experience with Collections as data	What is the level of your experience with preparing Collections as data?	Scale 1-5 (1 = no experience; 5 = we have datasets ready and are confident that we know what to do)
	Feel free to include (a) link(s) to your collection data sets here	Text
	What were the main issues that you encountered when starting to prepare Collections as data?	Text
Learning to prepare Collections as data	How well-informed do you feel / did you feel when starting to move towards Collections as data?	Scale 1-5 (1= not well-informed at all; 5 = very well-informed)
	Main sources of information are / were (include as many as you wish)	Text
	What information would you like to have / have liked to have had when starting to work towards Collections as data? What knowledge would have made it easier?	Text
Summary	Acknowledgement and contact	-

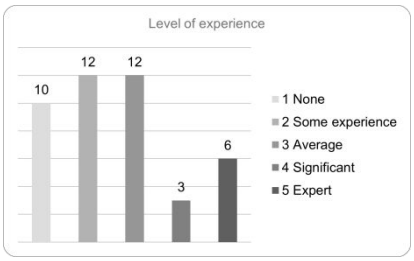


Figure 1: Survey results "What is the level of your experience with preparing Collections as data?"

The forty three unique responses came from GLAM and research institutions with a geographical spread across the USA (26) and Europe (14), complemented by one Asian and two fully anonymous contributions. Figure 1 shows that over half of the respondents indicated a low level of experience with preparing Collections as data and nine were significantly experienced or experts. Similarly, the majority of respondents felt ill-informed when starting work on Collections as data, with only two feeling very well-informed (Figure 2).

The core issues encountered when creating collections suitable for computational use were those of data preparation and dataset structure as well as matters of licensing and usage restrictions (Figure 3). Data preparation is hampered by data quality issues, particularly regarding OCR data, but also because of incoherent data and inconsistencies, e.g., in the resources' descriptive metadata. Decisions on ontologies, vocabulary reconciliation, identifiers, overall structuring and packaging are all identified as obstacles when creating the dataset structure.

In terms of information or knowledge that the respondents would like to have, Figure 4 reveals that institutions primarily name access to examples of implementation, followed by specific information on data preparation and general know-how about how to create collections as data.

The survey identifies a primary need for information and guidelines regarding the preparation of data and the structuring of datasets, in addition to the proposed checklist. Specific sought-after information on data preparation includes information on standards and best practices relating to file formats, metadata, data structure, and how to assess and (after selection) normalise the available data. General know-how should entail a user-friendly guide to tools, processes, decisions, and necessary policy choices and give insight into their implications on data modelling, data mapping and data reconciliation. A registry with Collections as data projects and descriptions of the dataset creation processes would inspire and support institutions with no relevant experience in the initial implementation stages.

Similarly, detailed accounts of the creation process for specific existing collections as data could help institutions make decisions when developing their own data for computational use. On an organisational level, resources such as use cases showcasing the added value of Collections as data could leverage strategic institutional support and encourage colleagues' and users' involvement.

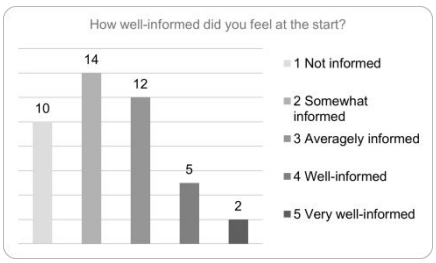


Figure 2: Survey results "How well-informed do you feel/did you feel when starting to move towards Collections as data?"

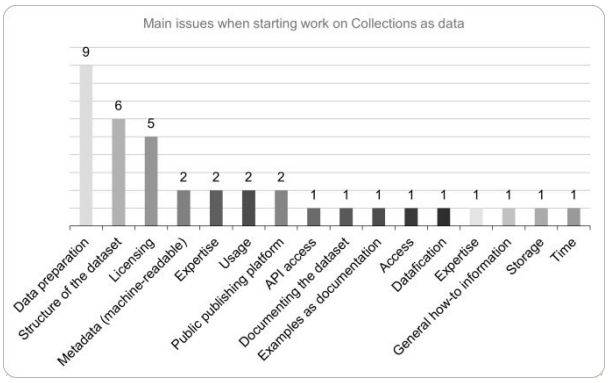


Figure 3: Survey results "What were the main issues that you encountered when starting to prepare Collections as data?"

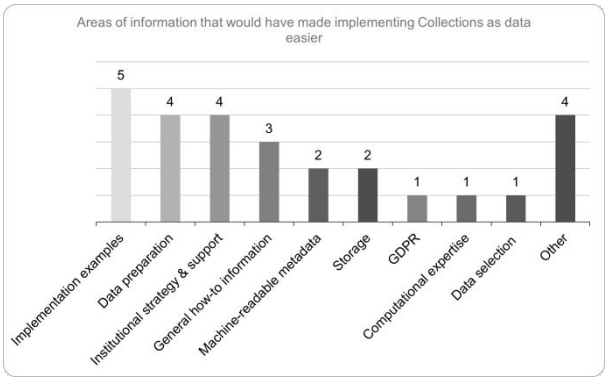


Figure 4: Survey results "What information would you like to have/have liked to have had when starting to work towards Collections as data? What knowledge would have made it easier?"

3.3 Synthesising and analysing the results of the literature review and survey to create the checklist

Building on the previous steps, a checklist to publish Collections as data was created as shown in Table III. In response to the survey results, the checklist offers GLAM institutions a much sought-after overview of elements that must be considered and, if relevant for the specific data collection or for the institutional context, developed during the process of preparing data collections for computational use. Whilst detailed discussion of each of the elements is outside the scope of this article, the main elements surfaced in the survey as areas in which potential uptakers require guidance such as data structure, metadata and examples of use, are expanded upon.

A preliminary version of the checklist was presented and discussed during an international webinar organised by the International GLAM Labs Community held on 25 October 2022 (Candela *et al.* 2022c; International GLAM Labs Community 2022b). An overview of each item is described below.

Table III: Checklist to publish Collections as data in GLAM institutions.

Item	Description	Yes	No
1	Provide a clear license allowing reuse of the dataset without restrictions (e.g., CC0, CC BY)		
2	Provide a suggestion of how to cite your dataset		
3	Include documentation about the dataset		
4	Use a public platform to publish the dataset		
5	Share examples of use as additional documentation		
6	Give structure to the dataset		
7	Provide machine-readable metadata (about the dataset itself)		
8	Include your dataset in collaborative edition platforms		
9	Offer an API to access your repository		
10	Develop a portal page		
11	Add a terms of use		

3.3.1 Provide a clear license allowing reuse of the dataset without restrictions

The adoption of licenses that allow reuse strengthens and expands the role of GLAM institutions in innovative scholarly communication. The use of permissive licenses is crucial to ease an understanding of the reuse possibilities and to facilitate the reuse of the digital collections (Padilla *et al.* 2019a; Candela *et al.* 2022a). Researchers expect a clear and reliable statement about the terms under which the dataset can be used.

During the past few years, organisations have started to publish and promote the use of metadata and digital objects in their collections (or part of them) under open licenses (National Library of Scotland 2019; Candela *et al.* 2021; Bibliothèque nationale du Luxembourg 2021; Europeana 2020; LIBER 2018; British Library 2019). Creative Commons licenses are a popular and widely-used tool. Some examples of licenses, statements and tools used by GLAM institutions are the following:

- The Creative Commons Public Domain Mark indicates that data is in the public domain. For instance, the Moving Image Archive published by the Data Foundry at the National Library of Scotland is published under this tool (National Library of Scotland n.d.[a]).

- The Creative Commons Public Domain Dedication (CC0) removes copyright restrictions on the use of the content. For instance, the British Library and the Library of Congress provide a selection of datasets published under this tool (British Library n.d.[e]; Library of Congress n.d.[c])
- CC BY data can be used when giving the appropriate credit to the source. For instance, the organisational data provided by the National Library of Scotland (National Library of Scotland n.d.[b]) is published under a CC BY license.
- National standards: other approaches are based on national licenses that describe how the data can be reused. For example, the Bibliothèque nationale de France made data available for the public on data.bnf.fr under the French Open license that enables the reuse and requires an attribution.
- Rights Statement "No known copyright" indicates that it is likely to be free from copyright restrictions but the public domain cannot be entirely confirmed.

In addition, publication platforms such as GitHub [3] and Zenodo [4] allow users to select an appropriate license when publishing the contents. License information can be provided as textual information, including a link to the appropriate license [5] or using metadata fields to describe copyright details such as the properties dc:rights and dcterms:license in the Dublin Core metadata schema.
Licensing the dataset must take into account the license of each of the resources contained in the dataset as these may vary.

3.3.2 Provide a suggestion of how to cite your dataset

A suggestion for the citation both promotes access and reusability of data and helps reusers to properly cite the dataset. Best practices recommend to include a preferred citation for the dataset (Padilla *et al.* 2019a).
A citation can be improved by using a permanent identifier to uniquely identify a resource such as a dataset (Candela *et al.* 2021). Digital Object Identifiers (DOI) are widely used by the community. For example, the datasets made available by the British Library and the National Library of Scotland provide a DOI as well as suggestions for citation. In fact, platforms such as Zenodo and DataCite [6] provide a DOI for all published resources, including a citation in the most common citation formats such as BibTeX and APA.
Another practice is to describe the publication of a dataset in a research article that then can be used as a citation since journals provide citations in several formats. Several examples include the description of the transformation of a dataset into Linked Open Data (LOD) that have been made available as a research article (Dijkshoorn *et al.* 2018; Koho *et al.* 2021).

3.3.3 Include documentation about the dataset

Documentation is a key element to foster the reuse by the community (Padilla *et al.* 2019a). Documentation may include details about the original sources as well as the cleaning and transformation principles and actions performed, information about how to access and use the dataset, or a description of the quality in terms of the content provided (Europeana 2020).
The documentation can be provided in several ways such as a blog post, README files and tutorials. For example, Chronicling America provides information about the dataset by means of a dedicated website (Library of Congress n.d.[d]). Other examples are based on the use of README files, as is the case for the British Library (British Library n.d.[e]).

3.3.4 Use a public platform to publish the dataset

Public platforms to make available datasets enable reusers to download the contents in bulk (Padilla *et al.* 2019a). Some examples of free platforms are GitHub, Zenodo, Hugging Face [7] and DataCite. However, some platforms may have limitations in terms of size for which paid services may be required. For example, the National Library of Scotland and the British Library use cloud storage services for their datasets (National Library of Scotland 2020).

These platforms provide additional features such as release management that can be useful to publish different versions of the same dataset (Romein *et al.* 2022).

3.3.5 Share examples of use as additional documentation

Examples of use of the contents provided by a digital collection are useful to inspire researchers (Padilla *et al.* 2019a; Mahey *et al.* 2019).
In particular, a Lab environment within a GLAM organisation is the place where reusers are able to find examples and prototypes based on the digital collections that in many cases are made available under open licenses. For example, the KB Labs (National Library of the Netherlands n.d.[b]) from the National Library of the Netherlands provides a list of tools and the LC Labs from the Library of Congress include the experimental tool Newspaper Navigator that allows users to browse the images extracted from the digitised newspapers database Chronicling America (Lee *et al.* 2020).
In other cases, reproducible Jupyter Notebooks [8] are used to introduce researchers to how to access and reuse the datasets. A Jupyter Notebook combines textual descriptions and code in the form of cells that can be run step by step. Some examples are the GLAM Workbench (Sherratt 2021) and the GLAM Jupyter Notebooks from Biblioteca Virtual Miguel de Cervantes (Candela *et al.* 2022b).
Other approaches entail the publication of tutorials on platforms such as The Programming Historian (Crymble, MacEachern, and Turkel 2012) and Library Carpentry (Baker *et al.* 2016), and research articles in journals describing how the dataset was created and reused.

3.3.6 Give structure to the dataset

A coherent internal distribution of a dataset is essential for researchers wishing to explore and query that dataset. Depending on the size and the type of contents, the structure will differ. Digital materials include a wide variety of content types, including images, maps, metadata, text, music and video amongst others.
There are some rules that will allow for a better understanding of the content provided by the dataset. One way to enhance this understanding is, for example, using self-describing folder names (e.g., text or images). Another approach could be based on the file format of the files provided (e.g., txt and XML). Each file included in the dataset may be named with the local identifier in the GLAM organisation. When having different formats for each resource (e.g., XML and JSON), a new root folder can be created clustering each of the formats.
For example, the Bibliothèque nationale du Luxembourg made available historical newspapers as open data using a zip file (Bibliothèque nationale du Luxembourg 2021). Each journal is included in a folder named with the title and the date. Each folder provides a set of folders according to different type of contents (images, pdf, text, thumbnails), the complete pdf and a xml file. Other approaches are based on metadata and provide a set of documents with different formats (e.g., Dublin Core and MARC) such as the Moving Image Archive.

More advanced initiatives such as BagIt File Packaging Format (Kunze *et al.* 2018), describes a set of hierarchical file layout conventions for storage and transfer of arbitrary digital content.
When providing large-size images, which is often the result of a digitization process, it can be interesting to provide reduced-size thumbnails based on the original images to be able to visualise them easier and faster. One additional aspect to consider is the cleaning of the data before publication. For example, sometimes postcorrection OCR data is included in the case of digitization datasets, or metadata collections may require cleaning to remove unnecessary metadata fields.

3.3.7 Provide machine-readable metadata

There is a wide variety of forms and formats to make metadata about digital resources (e.g., a dataset) available. The use of interoperable machine-readable metadata enhances discoverability and use since the data is readily processed by a computer (World Wide Web Consortium 2017). Some examples of vocabularies to provide metadata are MARC, Dublin Core, Vocabulary of Interlinked Datasets (VoID) (World Wide Web Consortium 2011) and Data Catalog Vocabulary (DCAT) (World Wide Web Consortium 2020). Other initiatives are based on Resource Description Framework (RDF) and

schema.org. For example, the machine readable metadata description using the vocabulary DCAT for the dataset National Bibliography of Scotland published by the Data Foundry is shown in Listing 1.

Listing 1: Machine-readable metadata description using the vocabulary DCAT for the dataset National Bibliography of Scotland published by the Data Foundry

```
@prefix dcat : <http : //www.w3. org/ns/dcat#> .
@prefix dct : <http : // purl . org/dc/terms/> .

<https://doi.org/10.34812/7cda-ep21> a dcat : Distribution ;
   dcat:downloadURL <https://nlsfoundry.s3.amazonaws.com/metadata/nls-nbs-v2.zip> ;
    dct:license <https://creativecommons.org/licenses/by/4.0/> ;
    dcat:mediaType <https://www.iana.org/assignments/media-types/text/xml> ;
    dcat:compressFormat <http://www.iana.org/assignments/media-types/application/zip>
.
```

3.3.8 Include your dataset in collaborative edition platforms

Collaborative edition platforms have become increasingly relevant in the GLAM context to create links and enrich their collections (Padilla *et al.* 2019a; Godby *et al.* 2019; Candela *et al.* 2022a). Crowdsourcing approaches enable the community to contribute to the content in a collaborative environment.

Wikidata, for example, enables the creation of resources known as entities, adding properties to describe the entities. The edition is performed using an easy and accessible web interface. For example, the section dedicated to computational access to digital collections at the International GLAM Labs Community website includes a selection of Jupyter Notebooks projects made available by relevant institutions that have been published in Wikidata (International GLAM Labs Community 2022a). Wikidata provides a public API to access the data, enabling users to retrieve the contents. Table IV shows an overview of Wikidata properties that can be useful to describe datasets.

Table IV: Overview of Wikidata properties to describe a dataset as an entity.

Property	Identifier	Description
full work available at URL	P953	full work available at URL
instance of	P31	that class of which this subject is a particular example and member. For example, newspaper archive and data set.
language of work or name	P407	language associated with this creative work
owned by	P127	owner of the subject
publication date	P577	date or point in time when a work was first published or released
title	P1476	published name of a work

3.3.9 Offer an API to access your repository

The use of an API to make available the dataset is a key element to foster reuse (Padilla *et al.* 2019a). APIs allow systems to communicate and to access and retrieve the entire dataset. In some cases, only a portion of the dataset may be retrieved for analysis using the API.

The use of an API to publish the digital contents may require additional features to be considered. For instance, when using IIIF, each resource should include a manifest.json describing the contents of this resource. For LOD, the adoption of URL patterns for the resources (e.g., author/id or author/name) is required as well as an analysis of how the data will be modelled (e.g, classes used and number of properties) according to the controlled vocabularies used to describe the metadata.

Table V introduces an overview of digital collections made available by institutions using a wide variety of APIs.

Table V: Overview of digital collections made available by relevant institutions using a wide variety of APIs.

Dataset	API	URL
Biblioteca Virtual Miguel de Cervantes	Linked Data	https://data.cervantesvirtual.com/sparql
Bibliothèque nationale du Luxembourg	OAI-PMH	https://data.bnl.lu/apis/oai-pmh/
Chronicling America	JSON	https://chroniclingamerica.loc.gov/about/api
Deutsche Nationalbibliothek	OAI-PMH	https://www.dnb.de/EN/oai
Harvard	IIIF	https://iiif.harvard.edu/about-iiif/
Harvard Art Museums	JSON	https://harvardartmuseums.org/collections/api
Library of Congress	JSON/YAML	https://www.loc.gov/apis/
Museum of Modern Art	JSON or XML	https://api.moma.org/
Victoria and Albert Museum	IIIF	https://developers.vam.ac.uk/guide/v2/images/iiif.html
WarSampo	Linked Data	https://www.ldf.fi/

3.3.10 Develop a portal page

Using a portal page for the dataset enhances the visibility and facilitates additional information about reusing the data (Padilla *et al.* 2019a). This information may include references to links for the dataset, visualisations, awards received, contact information, etc. For example, the dataset Chronicling America includes a dedicated website to access the contents but also to understand how the API can be used and to provide information about the original sources and license.
In addition, platforms such as GitHub provide free services to publish websites that are stored as a code repository and enables the use of several themes [9].

3.3.11 Add a terms of use

Best practices show the importance of adding terms of use describing the conditions of use for the data (Padilla *et al.* 2019a). The content can be provided as an additional section on a portal page or as a text document.
For example, the British Library EThOS dataset includes a terms of use section that details copyright, liability and access statements (British Library n.d.[f]). Other examples describe additional aspects such as how to report content as inappropriate in situations where people's rights are violated (Royal Danish Library n.d.).

4 Evaluation and application of the checklist

The checklist is intended as a tool for institutions to start implementing the Collections as data principles by giving a list of actions that can be performed so as to make collection data ready for computational use and reuse. While it is not necessary to carry out all the steps on the checklist to publish Collections as data, they give a clear direction when deciding on which actions to prioritise and which to defer to a later stage or to consider as unfitting to the specific collection. The checklist also serves as a tool for assessing existing collections as data, indicating the level of readiness for potential computational use.
The following section provides examples for both of the above use cases. It presents the results of the application of the checklist to assess a selection of datasets made available by relevant GLAM institutions. Furthermore, it provides details of three case studies where the checklist provided a useful framework to help implement the Collection as data principles in an institutional context. A case study is included that specifically presents an example regarding the complex issue of providing a clear license. The institutions were selected according to the following criteria: i) they are members of the International GLAM Labs Community; ii) they provide content for the public under open licenses; and iii) they are interested in the adoption of the Collections as data principles. As a result, a wide diversity of datasets in terms of content, access and use, from different types of institutions in a variety of geographical locations is provided. These can be useful examples for other institutions wishing to use and publish Collections as data.

4.1 The checklist as a tool for assessing GLAM collections as data

A selection of datasets made available by a wide variety of GLAM institutions in terms of size has been assessed against the checklist (see Table VI). The institutions and datasets are listed below.

- The British Library has a number of data services available to support different use cases, for example the content showcased on data.bl.uk and hosted on the open access British Library Research Repository (British Library n.d.[a]). Some of these datasets will also have a corresponding collection guide (British Library n.d.[c]).
- Data Foundry is the National Library of Scotland's open data platform, which includes digitised datasets, metadata, spatial data and organisational data. For this example, we have chosen 'Encyclopaedia Britannica', the most-used dataset. This covers the first 8 editions (100 years) of the Encyclopaedia (National Library of Scotland 2020).
- The Library of Congress (LC) recently published data.labs.loc.gov, as an experimental sandbox for sharing data packages compiled as part of LC Labs' Mellon Foundation-funded Computing Cultural Heritage in the Cloud (CCHC) initiative (Library of Congress 2019). In this context, the Stereograph Card dataset consists of 39,526 stereograph card images from the 1850s through 1924, a subset of what was available online in the collection in the catalog in August 2022.
- The Royal Danish Library made available an API for its digitised collection as a result of a newspaper digitization project running from 2014 to 2017. The construction of the API has been a way to experiment with the OpenAPI standard [10].
- Art In Flanders (AIF) is a dataset supported by Meemoo that includes more than 20.000 images of objects from Flemish museums and cultural institutions, comprising paintings, sculptures, archaeological artefacts, design objects, and more. Digital reproductions and descriptive metadata are being made available through the artinflanders.be platform.
- Miguel de Cervantes Virtual Library (BVMC) made available its main catalogue as Linked Open Data using Resource, Description and Access (RDA) as its main vocabulary (Candela *et al.* 2018).

Table VII shows the results obtained after the assessment in terms of the items provided by the checklist introduced in Section 3.

Table VI: Overview of the datasets and organisations used for the assessment.

Organisation	URL	Dataset description	license
British Library	https://www.bl.uk/collection-guides/digitised-printed-books	Digitised printed books (18th-19th century)	Public Domain Mark
Library of Congress	https://data.labs.loc.gov/stereographs/	39,526 stereograph card images from the 1850s through 1924	Library's statement
Miguel de Cervantes Virtual Library	https://data.cervantesvirtual.com/datos-enlazados	Main catalogue as LOD	Creative Commons CC0 1.0 Universal Public Domain Dedication
Meemoo	https://artinflanders.be/en	20.000 images of objects from Flemish museums and cultural institutions	Public Domain Mark and in-copyright
National Library of Scotland's Data Foundry	https://doi.org/10.34812/cg4r-dn40	Encyclopaedia Britannica	Public Domain Mark
Royal Danish Library	https://www2.statsbiblioteket.dk/mediestream/avis	Digitised newspaper collection	Public Domain Mark

Table VII: Overview of the results obtained when evaluating the checklist against a list of datasets made available by relevant GLAM institutions.

Organisation	1	2	3	4	5	6	7	8	9	10	11
British Library	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Library of Congress	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓
Meemoo	✓	✓	✓	✓	-	✓	✓	-	-	✓	✓
Miguel de Cervantes Virtual Library	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
National Library of Scotland	✓	✓	✓	✓	✓	✓	✓	-	-	✓	✓
Royal Danish Library	✓	-	✓	✓	-	✓	-	✓	✓	✓	✓

Checklist item 1: License. All the datasets and platforms assessed provide a clear license. For example, the British Library has a formal access and reuse process to identify if works are out of copyright or in copyright, the National Library of Scotland's Data Foundry provides the license for each dataset (National Library of Scotland 2020) and the Library of Congress bases its reuse policies on its rights statement on the source collection. Table VI introduces the licenses used in the dataset.
Checklist item 2: Suggested citation. In general, most of the datasets provide a persistent identifier such as a DOI. Other examples such as the National Library of Scotland's Data Foundry provide a suggested citation. The Library of Congress offers citation details for the source collections and the dataset creators and contributors. In other cases, an article - for example in a data journal - can be used to describe a data. In this case, the article can therefore be used to cite the dataset.

Checklist item 3: Documentation about the dataset. All the datasets provide dataset information and metadata as documentation in a wide diversity of manners (e.g., website, README) and granularity (e.g., collection and individual level). Other approaches are based on the use of machine-readable vocabularies based on RDF to describe the datasets such as the Vocabulary of Interlinked Datasets (VoID) (World Wide Web Consortium 2011). Table VIII shows an overview of the approaches followed by the institutions selected in this work.

Table VIII: Overview of strategies to provide documentation about machine-actionable collections in GLAM institutions.

Organisation	Documentation strategies
British Library	Website and datasheets
Library of Congress	Website, README, cover sheet, data processing plan
Meemoo	Website
Miguel de Cervantes Virtual Library	Website, research journal, VoID file
National Library of Scotland	Website
Royal Danish Library	Website

For example, the Library of Congress provides three levels of documentation for the datasets made available on data.labs.loc.gov by means on different files: i) a README file with a technical overview of how the data set was created (e.g., details of the dataset source collection, computational readiness and possible uses, dataset field descriptions and rights statement); ii) a data cover sheet file with a more substantive overview of the data and the collection from which it is derived (e.g., version information, background of collection, original format, reading room details, contact and metadata types); and iii) a data processing plan describing the goal of the experiment and a description of intended use, and data documentation regarding different aspects such as composition, provenance, compilation methods, preprocessing steps and potential risks to people and communities, amongst others (Eileen J. Manchester 2022).

In addition, the BL have explored innovative approaches such as Datasheets for Datasets by including a datasheet in the datasets, documenting its motivation, composition, collection process, recommended uses, etc. to facilitate better communication between dataset creators and dataset consumers (British Library 2022).

Checklist item 4: Use of a public platform to publish the dataset. All the datasets are available by means of public platforms. However, there are differences across the institutions regarding the use of institutional and third party platforms. For example, the BL uses both institutional and third party platforms, including British Library Research Repository (British Library n.d.[a]), Flickr, Wikimedia, Hugging Face, and secondary publishers, depending on the type/format of data. In the case of the other

institutions, the datasets are available by means of an institutional website (e.g., Lab section and dedicated website) as is the case for Meemoo, BVMC and Data Foundry.
Other organisations have different approaches depending on the content provided. For example, the Library of Congress provides access to datasets that have been officially acquired in the Selected Datasets Collection (Library of Congress n.d.[f]). For experimental or temporary datasets, access is provided on LC for Robots (Library of Congress n.d.[e]) or on data.labs.loc.gov which hosts datasets using cloud service providers.

Checklist item 5: Share examples of use. Many of the datasets assessed include examples of use as additional documentation to show how to reuse the contents. However, the approaches differ from one institution to another. For example, the National Library of Scotland's Data Foundry provides examples based on reproducible Jupyter Notebooks and the project includes collaboration initiatives based on the reuse of the datasets (National Library of Scotland n.d.[c]). The BL shares examples of dataset reuse on its Digital Scholarship blog (British Library n.d.[b]). The Library of Congress includes a section "Computational Readiness and possible uses" in the README files (Library of Congress 2022).

Checklist item 6: Give structure to the dataset. While datasets are structured according to different requirements and contents, in general, datasets are structured with reuse and data management in mind. For example, the National Library of Scotland's Data Foundry provides the datasets as zip files including folders per file format that can be easily identified by potential reusers. The Library of Congress has explored different ways to create and communicate coherence in datasets structure. Some examples are including dataset field descriptions in the README files, metadata and manifests for scripted and API access, and providing sample data and guidance on each data package page for ways to download the OCRed text, documentation and metadata.

Checklist item 7: Provide machine-readable metadata. All the datasets provide machine-readable metadata to describe the digital collections based on Dublin Core and more advanced approaches based on controlled vocabularies. The metadata is provided in form of additional files (e.g., XML) or through an API.

Checklist item 8: Include your dataset in collaborative edition platforms. While many institutions already provide information about their datasets in Wikidata, they are also interested to develop further Wikidata opportunities. Table IX shows an overview of the Wikidata approaches in GLAM organisations. However, it is important to notice that in some cases the information included is not related to datasets but to other initiatives such as projects and notebooks. In addition, other approaches are based on Wikimedia approaches. For example, a subset of the BL dataset is currently on Wikimedia Commons [11], which offers a useful introduction to the collection, including a Synoptic Index, as well as projects to georeference maps found in the texts.

Table IX: Overview of Wikidata links related to GLAM organisations.

Organisation	Wikidata link
British Library	https://www.wikidata.org/wiki/Wikidata:British_Library
National Library of Scotland	https://www.wikidata.org/wiki/Q111411199
Miguel de Cervantes Virtual Library	https://www.wikidata.org/wiki/Q111396572

Checklist item 9: Offer an API to access your repository. Most of the institutions have adopted APIs as a means to make available their collections based on different standards and tools. For example, the BVMC provides the datasets through a SPARQL public endpoint. Others provide a blend of this recommendation, using APIs to source collections and using manifests from data packages to gather those packages via a JSON/YAML API such as the Library of Congress.

However, some institutions have decided to provide the collections with simple, straightforward access through downloads to cater for those users whose technical skills are limited, such as students and artists.

Checklist item 10: Develop a portal page. All the institutions provide a data portal page including detailed information about the collections. Several examples included in this selection of organisations and datasets are the result of previous experimental data access points that have evolved to a new section such as Data.labs.loc.gov (Library of Congress n.d.[a]) and data.cervantesvirtual.com.

Checklist item 11: Terms of use. Terms of use are provided by the organisations. In some cases, the information includes contact details, for example the BL (British Library n.d.[g]). In other examples, the information is provided only in the country's official language (e.g., Spanish) (Biblioteca Virtual Miguel de Cervantes 2015). Table X shows an overview of the terms of use provided by the institutions.

Table X: Overview of terms of use provided by the organisations selected in this work.

Institution	URL
British Library	https://bl.iro.bl.uk/terms
Miguel de Cervantes Digital Library	https://data.cervantesvirtual.com/condiciones-de-uso/
Data Foundry	https://data.nls.uk/about/rights/
Meemoo	https://artinflanders.be/en

4.2 Case study: providing a clear license for meemoo and the Art in Flanders dataset

Providing a clear license allowing reuse of the dataset with as few restrictions as possible is one of the more complex items on the checklist. Managing these rights over time only adds to this complexity. The case of meemoo's Art in Flanders (AIF) dataset offers an example of this checklist item, presenting the selection and management of clear licensing information in AIF.

Rights management has gained much attention in the heritage sector over the last few years, in relation to developments in computing (e.g., Linked Open Data) and copyright legislation (e.g., EU directive on copyright). In line with this, the need arose to review the rights labelling policy for the AIF platform. The ambition was twofold: i) to make the dataset as freely available for access and reuse as possible; and ii) to adopt more appropriate standard rights labels for communicating the rights information. In this sense, some issues were identified:

- inadequacies in the rights information (e.g., a painting that had fallen into the public domain because its creator died more than 70 years ago, but with a copyright waiver (CC0) in the metadata and a photo credit © name-photographer on the picture).
- new copyright is claimed on digital surrogates of public domain works (e.g., photographic reproductions of 3-dimensional objects).
- owners may impose restrictions on the use of reproductions made of works in their collections, even if they are in the public domain.

The group of photographic reproductions of 3-dimensional artworks were particularly problematic, as the project was confronted with double-layered rights statuses. Two solutions were successively considered: 1) using separate rights labels: one for the rights status of the artwork and one for the rights status of the photo, and 2) using a single rights label that communicates the rights status and usage conditions for the resource as a whole. In parallel, the possibility of recontacting the photographers in question, and asking them to waive their rights, was considered.

For images where the owner of the cultural objects imposes use restrictions on reproductions, it was decided to adopt a *rightstatements.org* label. These labels have been specifically devised for heritage institutions to communicate rights information in a standardised way when they do not own the copyright and therefore using a copyright license is legally not possible.

For the reproductions of two-dimensional works, it was proposed to use updated labels for the three main groups. Firstly, the majority are in the public domain and can be released with a public domain mark instead of a copyright waiver. These images are freely downloadable in high resolution and reusable for any purpose without any restrictions. Secondly, where collection owners have imposed use conditions, it was proposed to use the rights statement "contractual restrictions". However, since this is not a legally binding tool, the user still needs to agree to user terms before downloading the images. In parallel, these contracts are currently being reviewed with the goal of minimizing and

standardizing the user restrictions far as possible. Thirdly, when works are under copyright, images get the "in copyright" mark.

For the reproductions of 3-dimensional works, a one-label approach was chosen for a number of reasons. Firstly, a single rights label is more user-friendly. It requires less pre-existing knowledge and leaves less room for (mis)interpretation by the user than the multiple-label approach. So in a sense, this is also the more secure and controlled approach. Secondly, not all labels are entirely compatible. For instance, a CC BY license which allows reuse of an image of an artwork, but which is also under full copyright and cannot be used without permission of the rights owner. In these cases, it was proposed to use the most restrictive label. In this way, pictures of artworks that are in copyright are tagged as such, even when the photographer agrees to a more open license. Additionally, pictures in the public domain get their label from the license agreed upon with the photographer.

In parallel, it was decided to include a maximum waiver of rights in (new) photography contracts. The process of clearing rights for older contracts is also in progress.

In summary, the updated 'open' licensing policy on the AIF platform proposes 4 rights labels for the main categories of images:

- Public Domain mark for images of 2D artworks in the public domain.
- CC0 for images of 3D artworks in the public domain.
- No-copyright - contractual restrictions for images of artworks in the public domain restricted by the collection owner (2D and 3D).
- In Copyright mark for images of artworks that are under copyright (2D and 3D).

The majority of images on AIF belong to the first two groups. In addition, providing access to the AIF dataset through an API is on the roadmap.

4.3 Case study: the checklist as a tool for implementing the Collections as data principles at KU Leuven Libraries

As apparent from the survey results presented in Section 3.2, when working towards disclosing collections as data, GLAM institutions are looking for inspiration and general insight in how to approach the implementation of the Collections as data principles. The checklist can give guidance to the process, allowing to make informed decisions on what to focus on first. To give potential users of the checklist insight into this process, this section describes the case of KU Leuven Libraries' (KU Leuven n.d.[b]) work on creating datasets for computational use.

Parameters for creating datasets depend on the context, the collection content, target users and intended use (Padilla *et al.* 2019b). Six datasets (see Table XI) were created as part of the preparation for a hackathon aimed at researchers and postgraduate students from within KU Leuven (*KU Leuven BiblioTech Hackathon* 2023). The hackathon organisation was a collaborative effort of KU Leuven Libraries and the university's Faculty of Arts (KU Leuven n.d.[a]). Considering this context and as this was the library's first endeavor in creating datasets for computational use, it was decided to (at least temporarily) offer access to the datasets (item 4, *use a public platform to publish the dataset*) through KU Leuven's internal data portal ManGO. *Share examples of use* (item 5) and *include your dataset in collaborative edition platforms* (item 8) were irrelevant to this context as participants were expected to work directly and independently with the source data. Regarding *-providing a clear license allowing reuse of the dataset without restrictions* (item 1), it was decided to only include resources in the public domain in order to allow for full reuse by participants outside of the hackathon. The Public Domain mark was provided on a resource-level in the descriptive metadata of the resources included in the dataset. *Add a terms of use* (item 11) was satisfied by including a statement in the hackathon Code of Conduct.

To start, the metadata and data were identified and extracted from their respective repositories. The dataset was subsequently structured according to *give structure to the dataset* (item 6): each dataset contained a separate

Table XI: Overview of datasets created by KU Leuven Libraries for the BiblioTech 2023 Hackathon (13-23 March 2023). At the time of writing, the datasets are only accessible for KU Leuven staff and students.

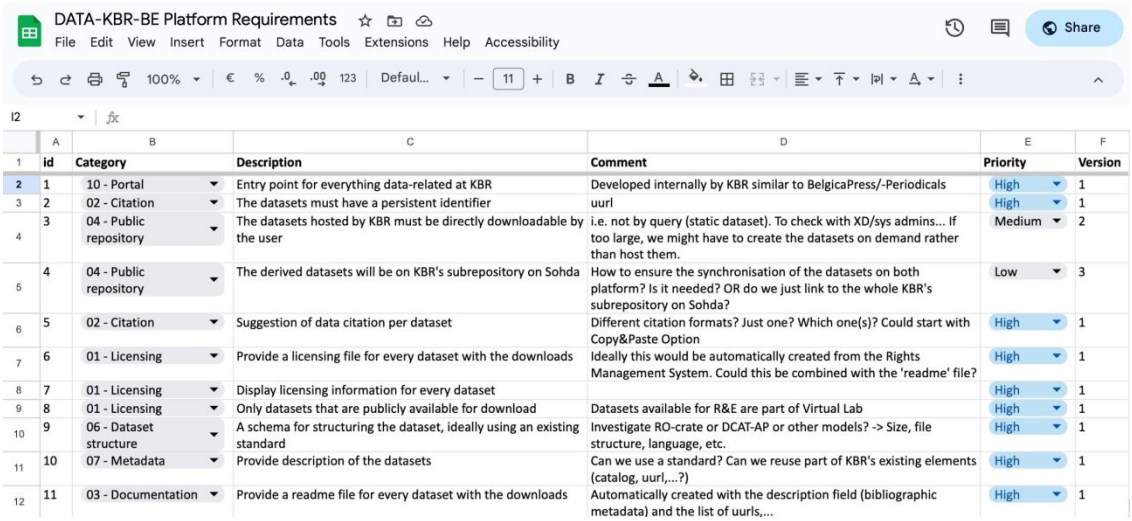
Dataset	URL to physical collection	Hackathon team poster DOI
Lecture Notes from the Old University of Leuven	https://kuleuven.limo.libis.be/ discovery/ collectionDiscovery?vid= 32KUL_KUL:KULeuven&inst= 32KUL_KUL&collectionId= 81411248550001488	https://doi.org/10.5281/zenodo.7762478
Anjou Bible	https://kuleuven.limo.libis.be/ discovery/fulldisplay?docid= alma9983846510101488&context= L&vid=32KUL_KUL:KULeuven&lang=en	https://doi.org/10.5281/zenodo.7762363
Postcards Belgium	https://kuleuven.limo.libis.be/ discovery/ collectionDiscovery?vid= 32KUL_KUL:KULeuven&inst= 32KUL_KUL&collectionId= 81411181930001488	https://doi.org/10.5281/zenodo.7769559
Wartime Posters	https://kuleuven.limo.libis.be/ discovery/ collectionDiscovery?vid= 32KUL_KUL:KULeuven&inst= 32KUL_KUL&collectionId= 81411182030001488	https://doi.org/10.5281/zenodo.7761872
Historical Censuses Belgium	https://kuleuven.limo.libis.be/ discovery/ collectionDiscovery?vid= 32KUL_KUL:KULeuven&inst= 32KUL_KUL&collectionId= 81423334580001488	https://doi.org/10.5281/zenodo.7764048
Academic Collection of the Old University of Leuven	https://kuleuven.limo.libis.be/ discovery/ collectionDiscovery?vid= 32KUL_KUL:KULeuven&inst= 32KUL_KUL&collectionId= 81411248640001488	https://doi.org/10.5281/zenodo.7762634

folder for each resource within which there were subfolders for each of the representations of these resources, e.g., a folder for page-level OCR data, for page-level jp2 images, and for PDF. On the level of the resource, a JSON manifest was included describing the resource. At the dataset level, the descriptive metadata was included as a xml metadata dump and as a partially cleaned csv and excel. A final csv was also provided, revealing the concordance between all the files in the dataset.

The full dataset was uploaded to the internal KU Leuven active data portal ManGO (item 4, *use a public platform to publish the dataset*), where hackathon participants could access the data, execute downloads or (providing the necessary infrastructure to work with the large datasets) connect the high performance computing infrastructure to the data through an API (item 9, *offer an API to access the repository*). In ManGO dataset metadata was added but for query purposes only. Documentation to each of the datasets included full information on the dataset structure, the descriptive metadata model, and some information on the level of the physical collection on the basis of which these datasets were created.

The library plans to further develop the datasets, improving there (re-)useability by including a Terms of use on a dataset level (item 11, *add a terms of use*), investigating possible locations to store and access the datasets for non-KU Leuven users (items 4, 9 and/or 10) as well as citation information which is so far lacking as the data is non-accessible to non-KU Leuven users (item 2, *provide a suggestion of how to cite your dataset*), and improving the structure and completeness of the datasets' documentation (item 3, *include documentation about the dataset*). It will also include the DOI of the hackathon team posters to inspire potential users (item 5, *share examples of use as additional documentation*), and investigate and develop a metadata model for dataset metadata (item 7, *provide machine-readable metadata*). The library has already pushed other data to a collaborative editing

platform such as Wikidata and considers doing the same for these datasets [12]. Yet, due to a lack of personnel, this will be postponed to a later date.
As a whole, the checklist directed the data preparation phase by providing a concise overview of which actions lead to collections as data, allowing for timely reflection on priorities. It now also supports decisions regarding next steps to take.


DATA-KBR-BE Platform Requirements

File Edit View Insert Format Data Tools Extensions Help Accessibility

Share

	A	B	C	D	E	F
1	id	Category	Description	Comment	Priority	Version
2	1	10 - Portal	Entry point for everything data-related at KBR	Developed internally by KBR similar to BelgicaPress/-Periodicals	High	1
3	2	02 - Citation	The datasets must have a persistent identifier	uurl	High	1
4	3	04 - Public repository	The datasets hosted by KBR must be directly downloadable by the user	i.e. not by query (static dataset). To check with XD/sys admins... If too large, we might have to create the datasets on demand rather than host them.	Medium	2
5	4	04 - Public repository	The derived datasets will be on KBR's subrepository on Sohda	How to ensure the synchronisation of the datasets on both platform? Is it needed? OR do we just link to the whole KBR's subrepository on Sohda?	Low	3
6	5	02 - Citation	Suggestion of data citation per dataset	Different citation formats? Just one? Which one(s)? Could start with Copy&Paste Option	High	1
7	6	01 - Licensing	Provide a licensing file for every dataset with the downloads	Ideally this would be automatically created from the Rights Management System. Could this be combined with the 'readme' file?	High	1
8	7	01 - Licensing	Display licensing information for every dataset		High	1
9	8	01 - Licensing	Only datasets that are publicly available for download	Datasets available for R&E are part of Virtual Lab	High	1
10	9	06 - Dataset structure	A schema for structuring the dataset, ideally using an existing standard	Investigate RO-crate or DCAT-AP or other models? -> Size, file structure, language, etc.	High	1
11	10	07 - Metadata	Provide description of the datasets	Can we use a standard? Can we reuse part of KBR's existing elements (catalog, uurl,...?)	High	1
12	11	03 - Documentation	Provide a readme file for every dataset with the downloads	Automatically created with the description field (bibliographic metadata) and the list of uurls,...	High	1

Figure 5: A collaborative spreadsheet for capturing the technical and functional requirements for the DATA-KBR-BE platform

4.4 Case study: towards a Collections as data platform at KBR, the Royal Library of Belgium

Inspired by the Collections as data movement (Padilla *et al.* 2019a; Padilla *et al.* 2019b), in 2020, KBR, the Royal Library of Belgium, embarked on a 48 month project [13] called DATA-KBR-BE (2020-2024) (Royal Library of Belgium 2020). The aim of the project is to optimise KBR's ICT infrastructure to stimulate sustainable data-level access to KBR's digitised and born-digital collections for digital humanities research. A key output of the project is to design and implement an Open Data Platform (data.kbr.be), for publishing KBR's Collections as data datasets.
Work on conceptualising the future DATA-KBR-BE platform began at an early stage in the project, during an initial *Brainstorming Workshop* held in November 2020. It was clear from the outset that a researcher-centred and iterative approach was needed to gather requirements for the design and development of the DATA-KBR-BE platform. A first important step in this process was to review some of the existing library data platforms, such as the national libraries of Luxembourg (Bibliothèque nationale du Luxembourg 2021), the Netherlands (National Library of the Netherlands n.d.[a]), Scotland (National Library of Scotland 2021) and The British Library (British Library n.d.[d]). Questions such as: *What data is offered? How? What format? What did people like, dislike about the platforms which were explored?* The outcomes of this workshop were used as the basis for iteratively developing a checklist of needs. The emergence of the "Checklist to publish Collections as data in GLAM Institutions" introduced in Section 3.3 provided the project team with an ideal framework to help structure the development of the functional and technical requirements for the platform.
To prepare for the webinar in October 2022 (Candela *et al.* 2022c; International GLAM Labs Community 2022b), an initial analysis of the checklist was undertaken to assess which checklist items were most relevant for the DATA-KBR-BE project. Initially, *develop a portal page* (item 10) and *give structure to the dataset* (item 6) were identified as the most relevant items for the project team, as our aim was to develop the DATA-KBR-BE platform and that we would need to understand how to structure the datasets that will be published there. However, it soon became relevant that many, if not all, the checklist items would support the development of the DATA-KBR-BE platform. For example, *provide a suggestion of how to cite your dataset* (item 2) and *provide a license allowing reuse of the dataset* (item 1) were quickly seen as essential.

To use the checklist more systematically, a collaborative spreadsheet was designed to capture each of the functional and technical requirements for the DATA-KBR-BE platform, as shown in Figure 5. Column B, is used for categorising each of the requirements based on the checklist list, e.g. requirement 1: *entry point for everything data-related at KBR*, has been categorised in relation to checklist item *develop a portal page* (item 10). This approach enabled us to: a) group the requirements by category, b) to ensure that our requirements analysis was as exhaustive as possible by considering each of the checklist items, and c) to provide feedback to the International GLAM Labs Community to further improve the checklist.

When reviewing the checklist items, the difference between *use a public platform to publish the dataset* (item 4) and *develop a portal page* (item 10) was not initially clear without further explanation. We were also not sure about *include your dataset in a collaborative edition platform* (item 8) and how it relates to items 4 and 10.

Additionally, *Provide machine-readable metadata* (item 7) caused quite some discussion in the DATA-KBR-BE project team. For example, what about human-readable metadata? What does machine-readable mean in this context? Why is it prioritised? Is this descriptive metadata? Are any particular standards recommended? How does this relate to structural metadata? Is this covered under *give structure to the dataset?* (item 6). The training and documentation aspects of the checklist *include documentation about the dataset* (item 3) and *share examples of use as additional documentation* (item 5) were both seen as very relevant to the development of the DATA-KBR-BE platform. However, they would likely be added following the initial development of the platform itself. Finally, *offer an API access to your repository* (item 9) was out of the scope of the current DATA-KBR-BE project, and will be addressed in a follow-up project, the KBR Virtual Lab.

In conclusion, the checklist was a valuable tool as it helped ensure that the DATA-KBR-BE project team had considered as many of the aspects of the checklist as possible when developing our Collections as data platform.

4.5 Discussion

While various institutions have made available their collections, there are still barriers hindering the adoption of the Collections as data principles, e.g., a lack of resources and of institutional support to make the collections easily available to a broad user group by means of simple access and downloads. The GLAM datasets which were selected for assessment in this article present some similarities but also some differences, e.g., the type of content, the formats and standards used for digital delivery, how they can be accessed, the licensing, and the documentation provided.

The checklist is informed by the issues and needs identified within the literature review and is complemented by the contributions of the practitioners who considered all the items included in the checklist relevant. In general, the practitioners observed a balance between simplicity and depth of practice. Some of them remarked that each of the items requires a different degree of maturity and prioritisation, which in some cases necessitates joint efforts by the community.

With regard to the application of the checklist as an assessment tool, and taking into account the wide variety of datasets provided by the GLAM institutions, the results obtained after the application of the checklist may differ amongst adopters of this approach. Initial results showed that the checklist is useful for identifying which aspects are relevant for a particular institution and, to some extent, easy to apply when making available datasets for computational use. In general, we observed that there is no order when applying the items in the checklist. Rather, as the case of KU Leuven Libraries demonstrates, priorities depend on the context, the content, the intended use and target users of the dataset. Furthermore, the checklist can facilitate the development of infrastructures related to Collections as data, as shown in the case of the DATA-KBR-BE platform. In general, future work based on the items in the checklist is a common goal across the institutions wishing to make their collections available as data.

While the institutional journeys into the delivery of Collections as data differ significantly, an additional layer of complexity in the computational use of cultural data which needs to be accommodated is the evolution towards data spaces for cultural heritage or large research infrastructures in the humanities which will both be using GLAM data. This is, for example, the case in Europe with the common European data space for cultural heritage, the European Cultural Heritage Cloud, and the European

Open Science Cloud (European Commission 2021). There is an ongoing effort to identify use cases for the data space for cultural heritage, and it would be helpful to coordinate this work with the future refinement of the checklist.

5 Conclusions

Over the past few years, there has been a growing interest in making available the digital collections published by GLAM organisations for computational use.

Based on previous work, we defined a methodology described in Section 3 to build a checklist for the publication of Collections as data. Our evaluation showed several examples of applications that can be useful to encourage other institutions to publish their digital collections for computational use.

Future work to be explored includes the improvement of the methodology by including additional features such as carbon footprint assessment, ethical issues and quality, as well as the inclusion of additional collections as data provided by organisations as use cases.

6 Acknowledgements

Text removed

7 Notes

1. http://www.ai4lam.org
2. https://www.newseye.eu
3. https://github.com
4. https://zenodo.org
5. See, for example, https://creativecommons.org/choose
6. https://datacite.org
7. https://huggingface.co
8. https://jupyter.org
9. https://pages.github.com
10. https://swagger.io/specification
11. https://commons.wikimedia.org/wiki/Commons:British_Library/Mechanical_Curator_collection
12. See, for example, https://commons.wikimedia.org/wiki/Category:Glass_diapositives_Egyptology,_KU_Leuven_Libraries and https://www.wikidata.org/wiki/Q112958007
13. 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. Originally foreseen as a 24-month project, in February 2022, the project was extended until 15th March 2024.

8 References

Agh, H., García, F. and Piattini, M. (2022), “A checklist for the evaluation of software process line approaches”, *Information and Software Technology.* 146, p. 106864, available at: https://doi.org/10.1016/j.infsof.2022.106864.

Australian Cultural Data Engine (2022), “Data Outputs”, available at: https://www.acd-engine.org/datasets; (accessed 20 June 2023).

Averkamp, Shawn *et al.* (2021), “Humans-in-the-Loop Recommendations report”, available at: https://labs.loc.gov/static/labs/work/reports/LC-Labs-Humans-in-the-Loop-Recommendations-Reportfinal.pdf; (accessed 20 June 2023).

Baker, J. *et al.* (Nov. 2016), "Library Carpentry: software skills training for library professionals", *Liber Quarterly: The Journal of European Research Libraries* 26.3, pp. 141–162, available at: https://doi.org/10.18352/lq.10176.

Biblioteca Nacional de España (2020), "Data", available at: https://bnelab.bne.es/en/data/; (accessed 20 June 2023).

Biblioteca Virtual Miguel de Cervantes (2015), "Condiciones de uso", available at: https://data.cervantesvirtual.com/condiciones-de-uso/; (accessed 5 May 2023).

Bibliothèque nationale de France (2021), "BnF Roadmap on AI", *2021-2026*, available at: https://www.bnf.fr/sites/default/files/2022-01/Poster_AI\%20Roadmap_BnF_202112.pdf; (accessed 3 May 2023).

Bibliothèque nationale du Luxembourg (2014), "BnL's Technical Requirements", available at: https://downloads.bnl.lu/bnlbooks2014/technical_requirements_and_appendixes.pdf; (accessed 3 May 2023).

— (2021), "Historical Newspapers, available at: https://data.bnl.lu/data/historicalnewspapers; (accessed 5 May 2023).

Boyd, D., F. Keller, E. and Tijerina, B. (2016), "Supporting Ethical Data Research: An Exploratory Study of Emerging Issues in Big Data and Technical Research", available at: https://datasociety.net/wp-content/uploads/2016/09/SupportingEthicsDataResearch_Sept2016.pdf; (accessed 1 May 2023).

British Library (n.d.[a]), "British Library's Research Repository", available at: https://bl. iro.bl.uk/; (accessed 5 May 2023).

— (n.d.[b]), "Digital scholarship blog. Enabling innovative research with British Library digital collections", available at: https://blogs.bl.uk/digital-scholarship/; (accessed 7 May 2023).

— (n.d.[c]). "Digitised printed books (18th-19th century)", available at: https://www.bl.uk/collection-guides/digitised-printed-books; (accessed 5 May 2023).

— (n.d.[d]), "Experiment with British Library's Digital Collections and Data", available at: https://data.bl.uk; (accessed 5 May 2023).

— (n.d.[e]), "Free dataset downloads", available at: https://www.bl.uk/collectionmetadata/downloads; (accessed 5 May 2023).

— (n.d.[f]), "Terms and Conditions of Use of EThOS", available at: https://ethos.bl.uk/ViewTerms.do; (accessed 5 May 2023).

— (n.d.[g]), "Terms of Use", available at: https://bl.iro.bl.uk/terms; (accessed 5 May 2023).

— (2019), "Foundations for the Future. The British Library's Collection Metadata Strategy 2019-2023", available at:https://www.bl.uk/bibliographic/pdfs/british-library-collection-metadatastrategy-2019-2023.pdf; (accessed 5 May 2023).

— (2020), "British Library Datasets", available at: https://data.bl.uk/bl_labs_datasets/; (accessed 5 May 2023).

— (2022), "Making British Library collections (even) more accessible", available at: https://blogs.bl.uk/digital-scholarship/2022/04/making-british-library-collections-evenmore-accessible.html; (accessed 5 May 2023).

Candela, G. and Carrasco, R. (2022). "Discovering emerging topics in textual corpora of galleries, libraries, archives, and museums institutions". *J. Assoc. Inf. Sci. Technol.* 73.6, pp. 820–833, available at: https://doi.org/10.1002/asi.24583.

Candela, G. *et al.* (2018), "Migration of a library catalogue into RDA linked open data", *Semantic Web* 9.4, pp. 481–491, available at: https://doi.org/10.3233/SW-170274.

Candela, G. *et al.* (2021), "A benchmark of Spanish language datasets for computationally driven research", *Journal of Information Science* 0.0, pp. 1–13, available at: https://doi.org/10.1177/01655515211060530.

Candela, G. *et al.* (2022a), "Evaluating the quality of linked open data in digital libraries", *Journal of Information Science.* 48.1, pp. 21–43, available at: https://doi.org/10.1177/0165551520930951.

Candela, G. *et al.* (2022b), "Reusing digital collections from GLAM institutions", *Journal of Information Science.* 48.2, pp. 251–267, available at: https://doi.org/10.1177/0165551520950246.

Candela, G. *et al.* (Oct. 2022c), "*Towards implementing Collections as Data in GLAM institutions*", available at: https://glamlabs.io/events/collections-data/; (accessed 5 May 2023).

Cordell, R. (2020), "Machine learning and libraries: a report on the state of the field", available at: https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf; (accessed 5 May 2023).

Crymble, A., MacEachern, A., and Turkel, W. (2012), "The Programming Historian 2: A Participatory Textbook", *7th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2012, Hamburg, Germany, July 16-22, 2012, Conference Abstracts*, p. 162, available at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/the-programminghistorian-2-aparticipatorytextbook.1.html.

Data Foundry (2020), "Data", available at: https://data.nls.uk/data/; (accessed 5 May 2023).

Digital Curation Centre (2013), "Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre", available at: http://www.dcc.ac.uk/resources/data-management-plans; (accessed 5 May 2023).

Digital Preservation Coalition (2022), "A sustainable future for our digital assets 2022 – 2027", available at: https://www.dpconline.org/docs/miscellaneous/about/2607-dpc-strategic-plan-2022-2027/file; (accessed 5 May 2023).

Dijkshoorn, C. *et al.* (2018), "The Rijksmuseum collection as Linked Data", *Semantic Web* 9.2, pp. 221–230, available at:https://doi.org/10.3233/SW-170257.

Dobbs, T. and Ras, Z. (2022), "On art authentication and the Rijksmuseum challenge: A residual neural network approach", *Expert Syst. Appl.* 200, p. 116933, available at: https://doi.org/10.1016/j.eswa.2022.116933.

Dobreva, M., Stefanov, K. and Ivanova, K. (2022), "Data Spaces for Cultural Heritage: Insights from GLAM Innovation Labs", *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries - 24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, November 30 - December 2, 2022, Proceedings*, pp. 492–500, available at: https://doi.org/10.1007/978-3-031-21756-2_41.

Eileen J. M. (2022), "Announcing LC Labs Data Sandbox and 3 New Data Packages", available at: https://blogs.loc.gov/thesignal/2022/12/announcing-lc-labs-datasandbox-and-3-new-data-packages; (accessed 5 May 2023).

European Commission (2021), "Commission proposes a common European data space for cultural heritage", available at: https://digital-strategy.ec.europa.eu/en/news/commission-proposes-commoneuropean-data-space-cultural-heritage; (accessed 5 May 2023).

— (2022), "The Cultural Heritage Cloud", available at: https://research-and-innovation. ec.europa.eu/research-area/social-sciences-and-humanities/cultural-heritage-andcultural-and-creative-industries-ccis/cultural-heritage-cloud_en; (accessed 5 May 2023).

Europeana (2020), "Strategy 2020-2025. Empowering digital change", available at: https://pro.europeana.eu/page/strategy-2020-2025-summary; (accessed 5 May 2023).

— (2022), "Common European data space for cultural heritage", available at: https://pro.europeana.eu/page/common-european-data-space-for-cultural-heritage; (accessed 5 May 2023).

Gebru, T. *et al.* (2021), "Datasheets for datasets", *Commun. ACM* 64.12, pp. 86–92, available at: https://doi.org/10.1145/3458723.

Godby, J. *et al.* (2019), "Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage", available at: https://doi.org/10.25333/faq3-ax08; (accessed 5 May 2023).

Han, S. and Choi, H. (2022), "Checklist for Validating Trustworthy AI", *IEEE International Conference on Big Data and Smart Computing, BigComp 2022, Daegu, Korea, Republic of, January 17-20, 2022*, pp. 391–394, available at: https://doi.org/10.1109/BigComp54360.2022.00088.

Harris, G., Potter, A. and Zwaard, K. (2020), "Digital Scholarship at the Library of Congress", available at: https://labs.loc.gov/static/labs/work/reports/DHWorkingGroupPaper-v1.0.pdf; (accessed 5 May 2023).

Harvard Art Museums (2022), "Application Programming Interface (API)", available at: https://harvardartmuseums.org/collections/api; (accessed 5 May 2023).

International GLAM Labs Community (2022a), "Computational access to digital collections", available at: https://glamlabs.io/computational-access-to-digital-collections/; (accessed 5 May 2023).
— (2022b), "Towards implementing Collections as Data in GLAM institutions", available at: https://glamlabs.io/events/collections-data; (accessed 5 May 2023).

Jakeway, E. *et al.* (2020), "Machine Learning + Libraries Summit Event Summary", available at: https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf; (accessed 5 May 2023).

Király, P. (2019), "Validating 126 million MARC records", *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, DATeCH 2019, Brussels, Belgium, May 08-10, 2019*, pp. 161–168, available at: https://doi.org/10.1145/3322905.3322929.

Koho, M. *et al.* (2021), "WarSampo knowledge graph: Finland in the Second World War as Linked Open Data", *Semantic Web* 12.2, pp. 265–278, available at: https://doi.org/10.3233/SW-200392.

KU Leuven (n.d.[a]), "Digital Humanities at the Faculty of Arts", available at: https://www.arts.kuleuven.be/digitalhumanities/english; (accessed 5 May 2023).
— (n.d.[b]), "KU Leuven Libraries", available at: https://bib.kuleuven.be/english; (accessed 5 May 2023).

KU Leuven Libraries (2023), "KU Leuven BiblioTech Hackathon", available at: https://zenodo.org/communities/bibliotech-hackathon-kuleuven/; (accessed 5 May 2023).

Kunze, J. A. *et al.* (2018), "The BagIt File Packaging Format (V1.0)". *RFC* 8493, pp. 1–25, available at: https://doi.org/10.17487/RFC8493.

Lee, B. (2022), "The "Collections as ML Data" Checklist for Machine Learning & Cultural Heritage". *CoRR* abs/2207.02960, available at: https://doi.org/10.48550/arXiv.2207.02960.

Lee, B. *et al.* (2020), "The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America", *CoRR* abs/2005.01583, available at: https://arxiv.org/abs/2005.01583.
LIBER (2018), "Research Libraries Powering Sustainable Knowledge in the Digital Age", available at: https://libereurope.eu/wp-content/uploads/2020/10/LIBER-Strategy-2018-2022.pdf; (accessed 5 May 2023).

Library of Congress (n.d.[a]), "Computing Cultural Heritage in the Cloud Derivative Datasets", available at: https://data.labs.loc.gov/; (accessed 5 May 2023).
— (n.d.[b]), "About the site and API", available at: https://chroniclingamerica.loc. gov/about/api/; (accessed 5 May 2023).
— (n.d.[c]), "Atlas of historical county boundaries", available at: https://loc.gov/item/2018487899; (accessed 5 May 2023).
— (n.d.[d]), "Chronicling America", available at: https://chroniclingamerica.loc.gov; (accessed 5 May 2023).
— (n.d.[e]), "LC for Robots", available at: https://labs.loc.gov/lc-for-robots/; (accessed 5 May 2023).
— (n.d.[f]), "Selected datasets", available at: https://www.loc.gov/collections/selected-datasets/about-this-collection; (accessed 5 May 2023).
— (2019), "Computing Cultural Heritage in the Cloud", available at: https://labs.loc. gov/work/experiments/cchc; (accessed 5 May 2023).

Library of Congress (2022), "Stereograph Cards Dataset README", available at: https://data.labs.loc.gov/stereographs/README.txt; (accessed 5 May 2023).

Lorang, E., Soh, L., and Pack, C. (2020), "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A demonstration project", available at: https://labs.loc.gov/static/labs/ work/experiments/final-report-revised_june-2020.pdf; (accessed 5 May 2023).

Mahey, M. *et al.* (Sept. 2019), "*Open a GLAM lab"*. eng. Doha, Qatar: International GLAM Labs Community, Book Sprint, 164. isbn: 978-9927-139-07-9, available at: https://doi.org/10.21428/16ac48ec.f54af6ae.

Museum of Modern Art (2022), "MoMA API", available at: https://api.moma.org/; (accessed 5 May 2023).

National and State Libraries Australasia (2022), "Strategic plan 2020-2023", available at: https://www.nsla.org.au/index.php/about-nsla/strategic-plan (accessed 5 May 2023).

National Library of Scotland (n.d.[a]), "Moving Image Archive", available at: https:// data.nls.uk/data/metadata-collections/moving-image-archive; (accessed 5 May 2023).
— (n.d.[b]), "Organisational data", available at: https://data.nls.uk/data/organisationaldata; (accessed 5 May 2023).
— (n.d.[c]), "Projects", available at: https://data.nls.uk/projects/; (accessed 5 May 2023).
— (2019), "National Library of Scotland Open Data Publication Plan", available at: https://data.nls.uk/download/national-library-of-scotland-open-data-publication-plan.pdf; (accessed 5 May 2023).
— (2020), "Encyclopaedia Britannica", available at: https://doi.org/10.34812/cg4rdn40; (accessed 5 May 2023).
— (2021), "Data", available at: https://data.nls.uk/data; (accessed 5 May 2023).
— (2022), "Environmental scan: Artificial Intelligence, cultural heritage and the National Library of Scotland", available at: https://doi.org/10.34812/ag56-3820; (accessed 5 May 2023).

National Library of the Netherlands (n.d.[a]), "Dataset", available at: https://lab.kb.nl/products/product_type/dataset; (accessed 5 May 2023).
— (n.d.[b]), "KB LAB", available at: https://lab.kb.nl; (accessed 5 May 2023).

Padilla, T. (2019), "Responsible Operations: Data Science, Machine Learning, and AI in Libraries", available at: https://doi.org/10.25333/xk7z-9g97; (accessed 5 May 2023).

Padilla, T. *et al.* (May 2019a), "50 Things — Always Already Computational: Collections as Data", available at: https://doi.org/10.5281/zenodo.3066237.
— (May 2019b), *Final Report — "*Always Already Computational: Collections as Data". Version 1, available at: https://doi.org/10.5281/zenodo.3152935.

Research Libraries UK (2022), "Strategy 2022-2025", available at: https://strategy.rluk.ac.uk/wp-content/uploads/2022/01/RLUK-Strategy-2022-25.pdf; (accessed 5 May 2023).

Reyserhove, L. *et al.* (2020), "A checklist recipe: making species data open and FAIR", *Database J. Biol. Databases Curation* 2020, available at: https://doi.org/10.1093/database/baaa084.

Romein, A. *et al.* (Nov. 2022), "Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done", available at: https://doi.org/10.5281/zenodo.7267245; (accessed 5 May 2023).

Royal Danish Library (n.d.), "Terms of service for end users of the Library Open Access Repository", available at: https://loar.kb.dk/handle/1902/4291; (accessed 5 May 2023).

Royal Library of Belgium (2020), "DATA.KBR.BE - Facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research", available at: https://www.kbr.be/en/projects/data-kbr-be/; (accessed 5 May 2023).

Sherratt, T. (Nov. 2019), "From collection search to collections as data", available at: https://doi.org/10.5281/zenodo.3551405; (accessed 5 May 2023).

— (Oct. 2021). "GLAM Workbench", available at: https://doi.org/10.5281/zenodo.5603060; (accessed 5 May 2023).

Smith-Yoshimura, K. (2020), "Transitioning to the Next Generation of Metadata", available at: https://doi.org/10.25333/rqgd-b343; (accessed 5 May 2023).

Strien, D. *et al.* (2020), "Assessing the Impact of OCR Quality on Downstream NLP Tasks", *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*, pp. 484–496, available at: https://doi.org/10.5220/0009169004840496.

Strien, D. van *et al.* (2021), "An Introduction to AI for GLAM", *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop, September 8+13, 2021, Virtual Conference*, pp. 20–24, available at: https://proceedings.mlr.press/v170/strien22a.html.

World Wide Web Consortium (2011), "Describing Linked Datasets with the VoID Vocabulary", available at: https://www.w3.org/TR/void/ (accessed 5 May 2023).
— (2017). "Data on the Web Best Practices", available at: https://www.w3.org/TR/dwbp; (accessed 5 May 2023).
— (2020), "Data Catalog Vocabulary (DCAT) – Version 2", available at: https://www.w3.org/TR/vocab-dcat-2/; (accessed 5 May 2023).

9 A List of studies

Table XII presents the list of primary studies analysed as a literature review on the use of Collections as data in GLAM institutions.

Table XII: List of primary studies analysed as a literature review.

Title	Origin	Type	Description
50 Things — Always Already Computational: Collections as data	Collections as data project	Report	Best practices
A benchmark of Spanish language datasets for computationally driven research	Journal of Information Science	Research article	Best practices
A checklist for the evaluation of software process line approaches	Journal Information & Software Technology	Research article	Checklist
A checklist recipe: making species data open and FAIR	Database J. Biol. Databases Curation	Research article	Checklist
A sustainable future for our digital assets 2022 – 2027	Digital Preservation Coalition	Report	Strategy
Artificial Intelligence Roadmap 2021-2026	Bibliothèque nationale de France	Report	Strategy
Assessing the Impact of OCR Quality on Downstream NLP Tasks	ICAART 2020	Research article	Quality
BnL's Technical Requirements	Bibliothèque nationale du Luxembourg	Report	Requirements for digitization projects
British Library Datasets	British Library	Datasets	Data publication
Checklist for a Data Management Plan. v.4.0.	Digital Curation Centre	Report	Checklist
Checklist for Validating Trustworthy AI	Conference	Research article	Checklist
Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage	OCLC Research	Report	Best practices
Data	Biblioteca Nacional de España	Datasets	Data publication
Data	Data Foundry	Datasets	Data publication
Data outputs	Australian Cultural Data Engine	Report	Data publication
Datasheets for datasets	Communications of the ACM	Research article	Data publication
Digital Libraries, Intelligent Data Analytics, and Augmented Description: A demonstration project	Library of Congress	Report	Best practices
Digital Scholarship at the Library of Congress	Library of Congress	Report	Best practices
Discovering emerging topics in textual corpora of galleries, libraries, archives, and museums institutions	Jasist Journal	Research article	Examples of use
Environmental scan: Artificial Intelligence, cultural heritage and the National Library of Scotland	NLS	Report	Strategy
Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done	Zenodo	Research article	Best practices
Evaluating the quality of linked open data in digital libraries	Journal of Information Science	Research article	Quality
Facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research	Royal Library of Belgium	Website	Project description
Final Report — Always Already Computational: Collections as data	Collections as data project	Report	Outcomes of the project
Foundations for the Future. The British Library's Collection Metadata Strategy 2019-2023	British Library	Report	Strategy
From collection search to Collections as data	Tim Sherratt	Report	Best practices
Harvard Art Museums API	Harvard Art Museums	Website	Technical documentation
Historical Newspapers	Bibliothèque nationale du Luxembourg	Datasets	Data publication

Humans-in-the-Loop Recommendations report	Library of Congress	Report	Best practices
LIBER Europe Strategy 2018-2022	LIBER	Report	Strategy
Machine learning and libraries: a report on the state of the field	Library of Congress	Report	Best practices
Machine Learning + Libraries Summit Event Summary Library of Congress	Report	Best practices	
Migration of a library catalogue into RDA linked open data	Semantic Web Journal	Research article	Data publication
Next generation of metadata	OCLC Research	Report	Best practices
On art authentication and the Rijksmuseum challenge: A residual neural network approach	Rijksmuseum	Research article	Experiment
Open Data Plan	National Library of Scotland	Open Data Plan	Best practices
Open a GLAM Lab	International GLAM Lab community	Book	Best practices
Responsible Operations: Data Science, Machine Learning, and AI in Libraries	OCLC Research	Report	Best practices
Reusing digital collections from GLAM institutions	Journal of Information Science	Research article	Examples of use
RLUK Strategy 2022-2025	Research Libraries UK	Report	Strategy
Strategic Plan 2020-2023	National and State Libraries Australasia	Report	Strategy
Strategy 2020-2025	Europeana	Report	Strategy
Supporting Ethical Data Research:An Exploratory Study of Emerging Issues in Big Data and Technical Research	Data & Society	Report	Best practices
The "Collections as ML Data" Checklist for Machine Learning and Cultural Heritage	ArXiv	Research article	Checklist
The Museum of Modern Art (MoMA) API is a REST service	MoMA	Website	Technical documentation
The Newspaper Navigator Dataset	ArXiv	Research article	Data publication
The Rijksmuseum collection as Linked Data	Semantic Web Journal	Research article	Data publication
Validating 126 million MARC records	DATeCH 2019	Research article	Quality
WarSampo knowledge graph: Finland in the Second World War as Linked Open Data	Semantic Web Journal	Research article	Data publication

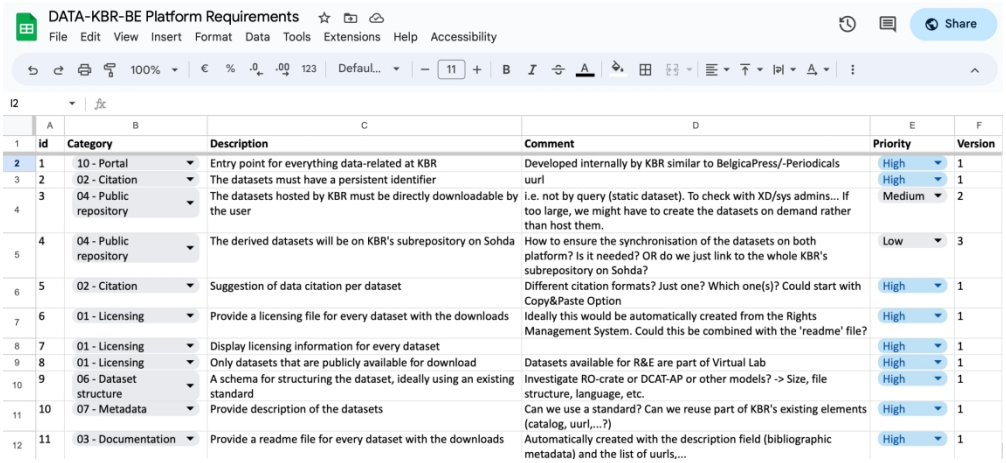
10 B Acronyms

Table XIII presents the list of acronyms used throughout the text.

Table XIII: List of acronyms used throughout the text.

Acronym	Description
AI4LAM	Artificial Intelligence for Libraries, Archives and Museums
ALTO	Analyzed Layout and Text Object
APA	American Psychological Association. Standard format used for citing sources.
API	Application Programming Interface
BibTeX	Tool and file format used to describe lists of references that are commonly used in LaTeX documents.
dblp	The DBLP Computer Science Bibliography
DCAT	Data Catalog Vocabulary
DOI	Digital Object Identifier
GLAM	Galleries, Libraries, Archives and Museums
IIIF	International Image Interoperability Framework
JSON	Javascript Object Notation
LOD	Linked Open Data
MARC	MAchine-Readable Cataloging
MARCXML	a framework for working with MARC data in a XML environment
METS	Metadata Encoding and Transmission Standard
OAI-PMH	Open Archives Initiative – Protocol for Metadata Harvesting

OCR	Optical Character Recognition
RDA	Resource, Description and Access
RDF	Resource Description Framework
VoID	Vocabulary of Interlinked Datasets
XML	Extensible Markup Language
YAML	Yet Another Markup Language



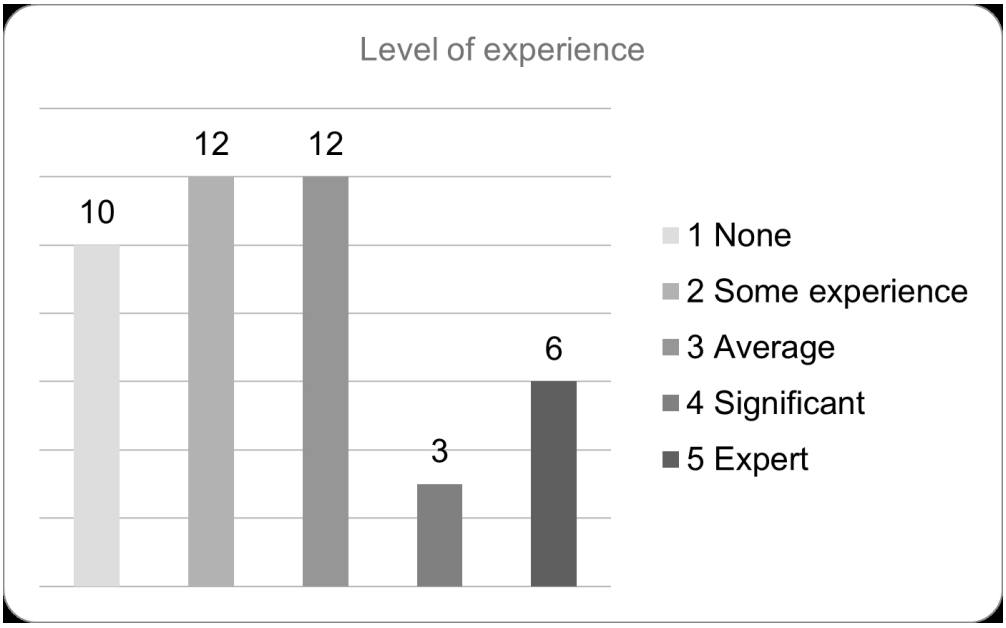
DATA-KBR-BE Platform Requirements

File Edit View Insert Format Data Tools Extensions Help Accessibility

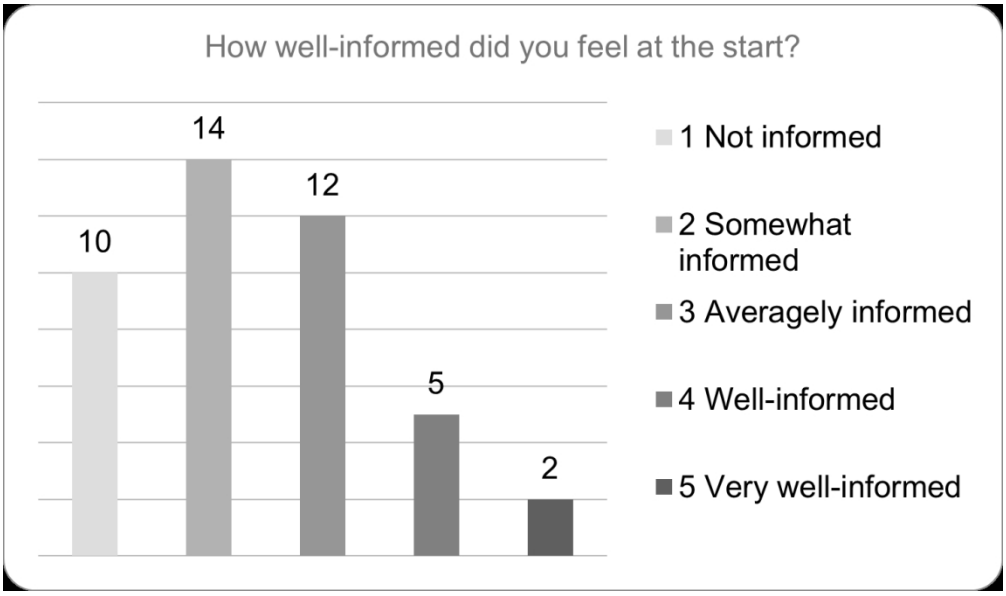
Share

	A	B	C	D	E	F
1	id	Category	Description	Comment	Priority	Version
2	1	10 - Portal	Entry point for everything data-related at KBR	Developed internally by KBR similar to BelgicaPress/-Periodicals	High	1
3	2	02 - Citation	The datasets must have a persistent identifier	uurl	High	1
4	3	04 - Public repository	The datasets hosted by KBR must be directly downloadable by the user	i.e. not by query (static dataset). To check with XD/sys admins... If too large, we might have to create the datasets on demand rather than host them.	Medium	2
5	4	04 - Public repository	The derived datasets will be on KBR's subrepository on Sohda	How to ensure the synchronisation of the datasets on both platform? Is it needed? OR do we just link to the whole KBR's subrepository on Sohda?	Low	3
6	5	02 - Citation	Suggestion of data citation per dataset	Different citation formats? Just one? Which one(s)? Could start with Copy&Paste Option	High	1
7	6	01 - Licensing	Provide a licensing file for every dataset with the downloads	Ideally this would be automatically created from the Rights Management System. Could this be combined with the 'readme' file?	High	1
8	7	01 - Licensing	Display licensing information for every dataset		High	1
9	8	01 - Licensing	Only datasets that are publicly available for download	Datasets available for R&E are part of Virtual Lab	High	1
10	9	06 - Dataset structure	A schema for structuring the dataset, ideally using an existing standard	Investigate RO-crate or DCAT-AP or other models? -> Size, file structure, language, etc.	High	1
11	10	07 - Metadata	Provide description of the datasets	Can we use a standard? Can we reuse part of KBR's existing elements (catalog, uurl,...?)	High	1
12	11	03 - Documentation	Provide a readme file for every dataset with the downloads	Automatically created with the description field (bibliographic metadata) and the list of uurls,...	High	1

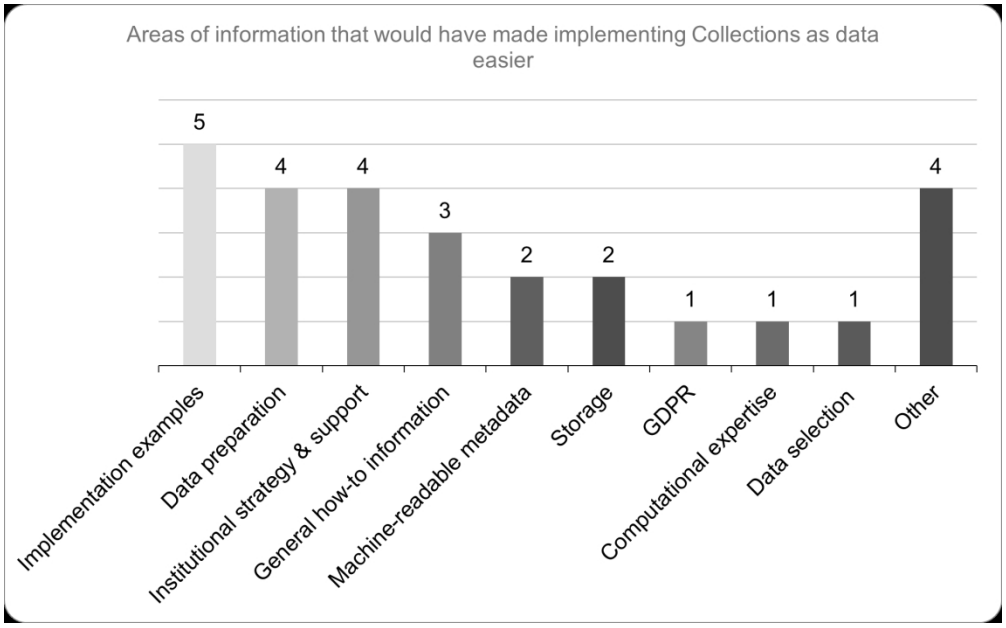
496x226mm (144 x 144 DPI)



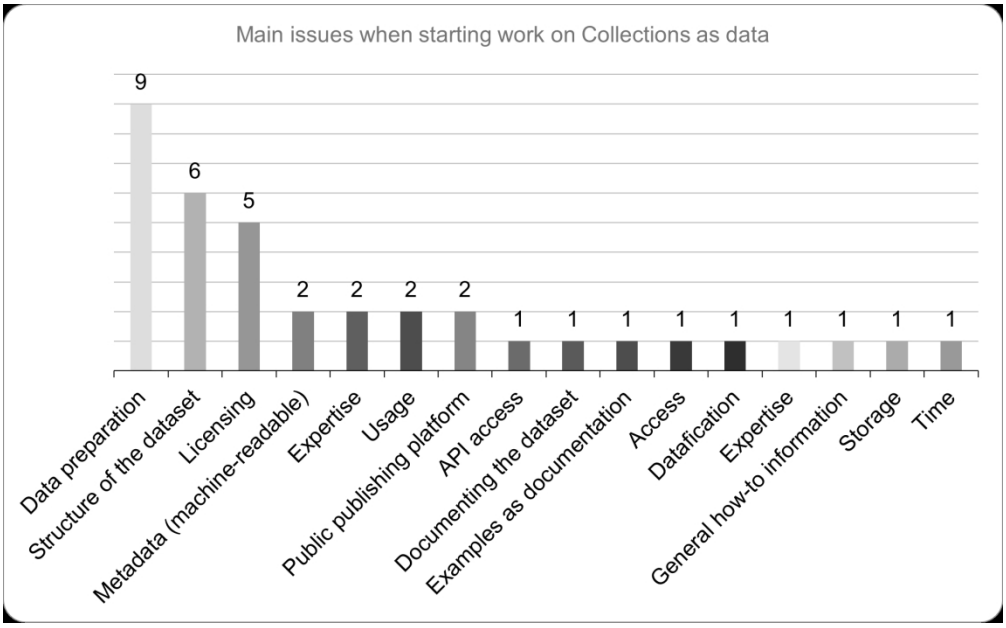
273x169mm (130 x 130 DPI)



288x169mm (130 x 130 DPI)



403x249mm (130 x 130 DPI)



403x249mm (130 x 130 DPI)