Journal Title
XX(X):1–16
~~©The Author(s) 2016~~
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Reusing digital collections from GLAM institutions

Gustavo Candela[1] and María-Dolores Sáez[1] and Pilar Escobar[1] and Manuel Marco-Such[1]

Abstract

For some decades now, Galleries, Libraries, Archives and Museums (GLAM) institutions have published and provided access to information resources in digital format. Recently, innovative approaches have appeared such as the concept of *Labs* within GLAM institutions that facilitates the adoption of innovative and creative tools for content delivery and user engagement. In addition, new methods have been proposed to address the publication of digital collections as datasets amenable to computational use. In this article, we propose a methodology to create machine actionable collections following a set of steps. This methodology is then applied to several use cases based on datasets published by relevant GLAM institutions. It intends to encourage institutions to adopt the publication of datasets that support computationally-driven research as a core activity.

Keywords

GLAM Labs, Collections as data, Linked Open Data, Digital Libraries

1 Introduction

Galleries, Libraries, Archives and Museums (GLAM) institutions have traditionally provided access to resources and services. They provide a rich and unique environment to bring together collections, services and people from different backgrounds.

As technologies have evolved over the years, GLAM organisations need to adapt to remain relevant in this new context. New skills are required regarding digital innovation ranging from service design, data science, digital research, and artificial intelligence. Moreover, the research community has highlighted the need for reproducible research by providing articles, as well as, data and code.[1,2]

As information service providers, GLAM institutions are strongly positioned to lead the effective exploitation of digital technologies within their institutions.[3]

In this sense, new approaches have recently appeared such as the concept of *Labs* within GLAM institutions that facilitates the adoption of innovative and creative tools for content delivery and user engagement.[4] In addition, new methods seek to address the publication of digital collections as datasets amenable to computational use such as Collections as data.[5] Making digital collections available as data and ready for computational analysis have an impact on user engagement since collections are more accessible and interoperable, using open licences and reactivating legacy material held by GLAM institutions.

New software solutions and methodologies provide an alternative and complementary means to search and access information available in digital collections. For instance, applying the Linked Open Data (LOD) concepts to the cultural heritage domain has become an active and challenging field. Innovative vocabularies to describe digital collections have appeared to enhance the expressiveness but also the interoperability. In addition, new frameworks such as International Image Interoperability Framework[6] (IIIF) provide access to high quality image resources providing a rich user experience. Several public repositories containing information that provides a cross-domain coverage have been created that can be used as a rich source of information, some of the most popular of which are DBpedia[7] and Wikidata.[8] The Jupyter Notebook is an interactive computing environment that enables users to create executable documents –called notebooks–

[1] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, carretera Sant Vicent s/n, 03690 Sant Vicent del Raspeig, Alicante (Spain)

Corresponding author:
Gustavo Candela.
Email: gcandela@ua.es

including live code, widgets, narrative text, images and plots. Notebooks have emerged as a powerful tool for transparent, collaborative, reproducible, and reusable data analysis that can be converted to various formats and shared with others.[9]

Although many institutions are currently exploring new ways of publishing their content to facilitate user engagement and promote its reuse, the publication of digital collections ready for computational analysis currently is not a core activity in the publication flow of GLAM institutions. There are several practical and structural barriers, such as required skills, dedicated staff, traditional and established trends, and copyright issues. In addition, while GLAM organizations provide access to digital collections in several ways, they are built upon traditional systems that are not suitable for new environments based on computational usage.

The purpose of this paper is to illustrate a methodology to create machine actionable collections following a set of steps. The methodology is tested by means of several use cases providing reproducible notebooks. The results of this study could then be used for researchers to create their own datasets and promote the creation of notebooks in the research community.

The main contributions of this paper are the following: (a) a methodology to create machine actionable datasets; (b) a collection of notebooks based on datasets published by relevant GLAM institutions; and (c) we show a practical example of exploitation of LOD.

The paper is organized as follows: after a brief description of the state of the art in Section 2, Section 3 introduces the methodology to create a machine actionable dataset. Section 4 presents the application of the methodology by means of several GLAM institutions and shows the results. The paper concludes with an outline of the methodology adopted, general guidelines for the use of the results and future work.

2 Related work

2.1 Overview

Traditional methods of accessing digital collections includes contacting the institution or using websites with limited search and content functionalities. In order to support digital scholarship, a growing number of institutions have adopted semantic web technologies and creating LOD projects.[10,11] However, the use of semantic web technologies require complex technical skills and knowledge of the contents, and the vast majority of the projects are focused on providing rich metadata described with global standards and controlled vocabularies. The prefixes in Table 1 are used to abbreviate namespaces throughout this paper.

In parallel, new approaches have emerged to give researchers more flexible access to collections including the underlying data. The FAIR data principles provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital collections.[12] Collections as data provides best practices for turning digital collections into datasets amenable to computational use and novel research methods.[5] As a result, many organizations recommend publishing machine actionable collections, combined with documentation about digital collections.[13,14]

A growing number of institutions are starting to publish datasets openly and in re-useable formats (e.g. CSV, XML, text, images). For instance, the National Library of Scotland Data Foundry[15] publishes metadata, image and as well as map collections including provenance. The Bibliothèque nationale de France (BnF) published in 2017 *Bnf API et jeux de données* including datasets and the API documentation. The British Library[16] is making some of its datasets available for research and creative purposes including images, full text and metadata. The Library of Congress provides machine-readable access to its digital collections including APIs and datasets.[17] In some cases, the institutions maintain access to datasets based on previous snapshots.[*] An overview of GLAM institutions providing access to datasets can be found in Table 2.

A great effort has been made regarding access to high quality image resources. The International Image Interoperability Framework (IIIF) is a protocol for standardized image retrieval collaboratively created by a community including libraries and archives to produce an interoperable framework for image delivery.[18] Some examples adopting IIIF are the Smithsonian Institution[19], Europeana[20] and the Staatsbibliothek zu Berlin.[21]

According to Collections as data final report,[22] institutions are recommended to share prototypes with the research community to provide examples of use of their collections as well as being relevant to new ways of conducting scholarship. Following this practices, many institutions such as the Royal Danish Library[23] in Denmark, provide prototypes based on digital methods including topic modeling, artificial intelligence, machine learning and exploration and visualization tools. In addition, some publications are focused on the development of machine actionable collections from several digital collections providing case studies to test digital humanities methods such as text mining, topic modeling and GIS (Geographic Information System).[24] Furthermore, machine actionable datasets

[*] https://github.com/tategallery/collection

Table 1. Common prefixes used to designate RDF vocabularies.

prefix	URI
blt	http://www.bl.uk/schemas/bibliographic/blterms#
c4dm	http://purl.org/NET/c4dm/event.owl#
dcmitype	http://purl.org/dc/dcmitype/
dcterms	http://purl.org/dc/terms/
foaf	http://xmlns.com/foaf/0.1/
ps	http://www.wikidata.org/prop/statement/
rdagroup2elements	http://rdvocab.info/ElementsGr2/
rdam	http://rdaregistry.info/Elements/m/
rdarelationships	http://rdvocab.info/RDARelationshipsWEMI/
rdaw	http://rdaregistry.info/Elements/w/
rdfs	http://www.w3.org/2000/01/rdf-schema#
schema	http://schema.org/
wdt	http://www.wikidata.org/prop/direct/
wd	http://www.wikidata.org/entity/
wikibase	http://wikiba.se/ontology#

Table 2. A selection of GLAM institutions providing datasets using diverse types of content and formats.

Institution	Dataset collections	URL
Bibliotèque nationale de France	BnF API et jeux de données	http://api.bnf.fr/
British Library	data.bl.uk	https://data.bl.uk/
Biblioteca Virtual Miguel de Cervantes	BVMC Labs	http://data.cervantesvirtual.com/blog/labs
Det Kgl. Bibliotek	KB Labs	https://labs.kb.dk/
Europeana	Europeana IIIF APIs	https://pro.europeana.eu/page/iiif
Ghent University	Ghent University library	https://lib.ugent.be/
History Trust of South Australia	Learn section	https://history.sa.gov.au/
Impact	Tools and resources	https://www.digitisation.eu/tools-resources/
National Library of Netherlands	KB Lab	https://lab.kb.nl/
National Library of Scotland	Data Foundry	https://data.nls.uk/
Library of Congress	LC for Robots	https://labs.loc.gov/lc-for-robots/
Österreichische Nationalbibliothek	ONB Labs	https://labs.onb.ac.at/en/
Smithsonian Institution	International Image Interoperability Framework	https://iiif.si.edu/
Staatsbibliothek zu Berlin	SBB Labs	https://lab.sbb.berlin/?lang=en
State Library New South Wales	DX Lab	https://dxlab.sl.nsw.gov.au
The British Museum	The British Museum Collection	https://www.britishmuseum.org/collection
Tate Gallery	The Tate Collection	https://github.com/tategallery/collection

combined with Jupyter Notebooks[25] provide transparent, collaborative, reproducible and reusable data analyses. A notebook integrates detailed workflows, narrative text and visualization of results. For instance, the GLAM Workbench[26] provides notebooks based on digital collections showing collection of tools, tutorials and examples. The Archives Unleashed Notebooks is a prototype to interactively explore and filter the information, extracted full text, and network visualization data.[27] Other approaches are based on data visualizations including graphs, charts and maps in order to help users better understand the scope and coverage of Chronicling America.[28]

Cultural changes and innovation in collections are transforming GLAM institutions in research centres. User engagement is becoming increasingly important and GLAM institutions are starting to develop fellowship programs of all disciplines to reuse the collections.[†] These programs open up new opportunities to interact with collections to researchers which are assisted by computer scientists and data specialists within the institution. Other approaches are defined as partnership between the institutions and the growing community of volunteers to help transcribing, tagging and editing.[29] In this context, free and open knowledge bases such as Wikidata have raised interest in GLAM institutions. Wikidata allows the participation of the community by editing information and enriching the digital collections. Recently, many properties have been created in order to establish links to records published by GLAM institutions.[1] Many are the benefits of being linked such as automatically enriching the datasets by including geographic and contextual information (see Figure 1).

Making data discoverable is essential to create the highest impact in the research community. In this sense, several online platforms provide services to cite, identify and share datasets such as DataCite and Zenodo. In addition, since sustainability of repositories raises a number of challenging issues in several areas including technical, financial and legal, recent reports published by the research community describe the characteristics of trustworthy repositories in order to ensure the reliability and durability of data repositories.[30]

Although best practices and guidelines are crucial in the creation of machine actionable collections, to our best knowledge, the creation of machine actionable datasets and Jupyter Notebooks currently is not a core activity and limited to some extent to the Labs domain, within GLAM institutions. In this sense, the publication of notebooks describing the process of creation and exploitation of machine actionable collections based on digital collections from GLAM institutions can help encouraging them to adopt these methods in their daily workflows.

2.2 *Methodologies and best practices for publishing datasets*

The main goal of this study is to introduce a methodology to publish machine actionable datasets. Since creating detailed and consistent datasets is a challenge common to most institutions, several approaches are applied when publishing datasets regarding its domain, quality and features.

Linked data are becoming very popular as a model used to publish data on the Web due to the possibility to connect datasets and being suitable to fully exploit the nature of the Web. Some examples are national libraries and historical photographic archives[31]. Several frameworks have been proposed to publish and enrich LOD datasets.[32,33] In general, these approaches work in a number of steps such as preprocessing, modeling, enrichment and disambiguation. However, its use requires complex skills, such as the use of Resource Description Framework[34] (RDF), in order to reproduce and fully exploit the information provided.

Open Refine[35] is a open source tool that allows users data cleanup and transformation to other formats. Open Refine provides semi-automated methods for processing large volumes of metadata requiring no computer programming experience. It also allows the user to export sequences of operations that can be applied later to other datasets. Several approaches based on Open Refine includes a set of steps: (i) data cleaning (e.g. trimming white spaces), (ii) migration to a computer friendly format, (iii) validation of data (e.g. accuracy) and (iv) data improvement (e.g. Named Entity Recognition).[36,37] While Open Refine supports reproducible research, exported sequences of operations do not include any edits made manually to individual cells.[38] In addition, Open Refine runs locally –not in the cloud– using a browser, but no web connection is needed.

Other approaches explore service models to support researchers accessing collections in the cloud, combining library staff and researchers and providing access in a controlled environment.[39] These experimental models aim at identifying new ways of using collections and methods of support, risks and costs.

3 A framework to create machine actionable collections

We have adapted the methodologies introduced in Section 2.2 in order to describe our framework to create machine actionable collections.

The framework described in Figure 3 works in 6 steps, which are detailed in the following subsections: (i) identification, (ii) access and retrieve, (iii) cleaning and reformat, (iv) data enrichment, (v) package dataset and (vi) analysis. The output of the data enrichment step

[†] See, for example, https://lab.kb.nl/news/call-kb-researcher-residence-2020-closed and https://data.nls.uk/projects/ai-in-residence/

```
SELECT DISTINCT ?resource ?place ?title ?date
WHERE {
  ?resource ?role
  <http://bnb.data.bl.uk/id/person/ShakespeareWilliam1564-1616> ;
  dct:title ?title ;
  schema:datePublished ?date .
  OPTIONAL {
    ?resource blt:publication ?publication .
    ?publication c4dm:place ?place .
    FILTER regex(?place, "geonames", "i")
  }
}
```

Figure 1. A SPARQL query retrieving the works of William Shakespeare from the British National Bibliography SPARQL endpoint, including the publication place represented by its GeoNames identifier. The output is shown in Table 3 in which the GeoNames identifier could be used to retrieve the attributes latitude and longitude from Wikidata, as is shown in Figure 2.

Table 3. Overview of the results obtained using the SPARQL query in Fig. 1.

Resource	Place	Title	Date
http://bnb.data.bl.uk/id/resource/013310275	http://sws.geonames.org/6269131/	Macbeth : teachit KS3 interactive pack	2006-01
http://bnb.data.bl.uk/id/resource/013310276	http://sws.geonames.org/6269131/	Much ado about nothing : teachit KS3 interacti...	2006-01
http://bnb.data.bl.uk/id/resource/013315368	http://sws.geonames.org/6269131/	Hamlet	2006-01
http://bnb.data.bl.uk/id/resource/019599478	http://sws.geonames.org/6269131/	Twelfth night	2020-01
http://bnb.data.bl.uk/id/resource/019599479	http://sws.geonames.org/6269131/	The tempest : the alexander text	2019-11

```
SELECT ?idgeonames ?lat ?lon
WHERE {
  values ?idgeonames { "2643743" }
  ?x wdt:P1566 ?idgeonames ;
   p:P625 [
     psv:P625 [
       wikibase:geoLatitude ?lat ;
       wikibase:geoLongitude ?lon ;
       wikibase:geoGlobe ?globe ;
     ];
     ps:P625 ?coord
   ]
}
```

Figure 2. A SPARQL query retrieving the latitude and longitude from Wikidata by means of an identifier of GeoNames. The identifier 2643743 corresponds to London in GeoNames.

is a machine actionable collection that can be analyzed and exploited as is shown in the last step.

Digital collections provided by GLAM institutions may be different in terms of the content and the structure of the information. A digital collection may include unstructured information such as plain text. Others may include images in different resolutions. Moreover, some collections may include structured information such as metadata in CSV format. As a result, regarding the content and its structure, the processing may vary and be adapted to exploit the digital collections to its full potential. Providing a one-for-all solution is a complex task since GLAM institutions use different means to produce their digital collections, and depending on the institution and the collection, each step of the framework can be applied to some extent or being bypassed.

In particular, copyright is a critical aspect no only when identifying and reusing digital collections, but also when publishing a machine actionable collection based on one or more original datasets that may have different licences. Recent approaches aim at creating opportunities to reuse materials in education, research and cultural heritage.[40,41] However, there is still room to improvement regarding the use of open licenses and, in particular, the case of derivative materials. As a result, the use of the framework has to be adapted to each case in order to avoid legal issues.

3.1 Identification

The identification of digital collections is not an easy task due to several reasons including copyright, quality, accessibility, completeness or ease of understanding. Fortunately, in the last years, GLAM institutions have made a great effort in this direction, adopting open

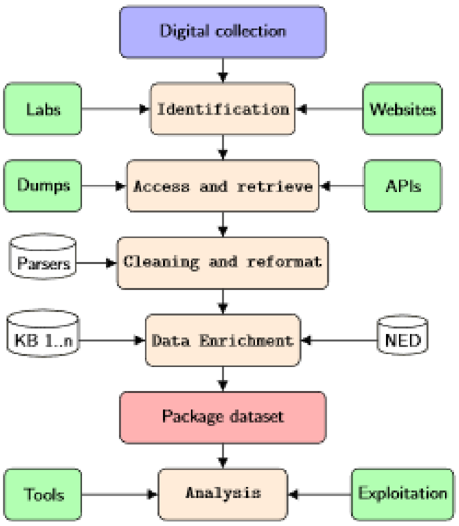


Figure 3. Framework employed to create machine actionable collections - KB stands for Knowledge Base, while NED stands for Named Entity Disambiguation.

licenses and providing documentation describing its collections.

Traditionally, websites have been explored in order to identify records based on specific criteria provided by the search functionality. In general, records have been classified by thematic areas easing the process of identifying. Some examples may include authors and works of a movement, portraits of artists and, language and literature. In addition, many Labs have been recently published including a section devoted to open and transparent datasets ready for its reuse. The datasets may be classified in terms of the type of the content such as text, images, maps and metadata.

3.2 Access and retrieve

Metadata and digital collections services in GLAM institutions have traditionally provided access to digital contents for research by means of a proposal.[‡] However, GLAM institutions require time to make a decision and to prepare the data, causing a delay in the initiation of the research.

Other approaches are based on automated processes to harvest the content. A first approach is accessing the websites using a web crawler to fetch the pages and extracting the data.

Advances in technology, standardization and interoperability have enhanced the definition of APIs. For instance, Open Archives Initiative Protocol for Metadata Harvesting[42] (OAI-PMH) is a low-barrier mechanism to expose structured metadata that can be requested by means of operations. More advanced systems are based on more complex and expressive languages such as SPARQL[43] that enable the execution of federated datasets. With regard to image-based resources –including images, books, newspapers, manuscripts–, IIIF defines a set of common APIs that support interoperability between image repositories and explores new ways of viewing, comparing, manipulating and annotating images. There are several ways to consume an API using programming languages such as Python, Java and JavaScript.

When working with large datasets, researchers are able to get manageable slices of the data by means of APIs, obtaining a subset data. For instance, by using SPARQL the researcher can indicate the number of records and fields to be retrieved before downloading the data.

GLAM institutions provide detailed documentation explaining how to harvest the digital collections by means of dedicated websites, PDF documents and tutorials.

Other approaches are based on the publication of data dumps that are created on a regular basis. The size of the dump –ranging to 1 gigabytes to several terabytes–, even when using compressed files, can pose some challenges in the import and exploitation process.

3.3 Cleaning and reformat

GLAM institutions publish digital collections from different sources and in different formats. In this sense, data cleaning involve different techniques based on the problem and the data type, which make it difficult to generalize this process and provide one-for-all solutions. However, we introduce a systematic approach including common steps that can be extended in order to address a particular issue. Some of the possible challenges and the measures taken are listed below. In addition, each measure may require the specification of a particular language according to the original datasets.

- Removing special characters. Depending on the project, the task will remove special characters that do not affect the meaning of the text, such as punctuation marks and symbols.
- Lowercasing. In some cases, it will be required to converts all uppercase characters to lowercase.
- Lemmatization. The different inflected forms of a word are grouped together so they can be analysed as a single item.
- Stopwords. Words which do not add meaning to a sentence are ignored. For example, the words like, the, he, have, etc.

[‡] See, for instance, https://www.bl.uk/help/proposing-a-new-research-collaboration-with-the-british-library

- OCR quality. Digital collections may include the Optically Character Recognised (OCR) derived text. Due to reasons of low quality printing and scanning, may result in corpora of poor quality, making the OCR task notoriously difficult. In some cases, it might be useful removing non existent words as a result of the digitisation process.
- Specific metadata formats. GLAM institutions use different standards to describe its collections such as MARCXML, Dublin Core, METS and TEI. Specific parsers are required to extract the data.

There are different technologies that have been traditionally used in this sense. For instance, regular expressions consist on a sequence of characters that provide a search pattern to perform find or replace string operations in text. Based on this background, technological advances have made available tools aimed at data cleaning using visual interfaces and avoiding code such as Open Refine. Other examples includes Extraction, Transformation and Load (ETL) tools that allow to create a workflow including several steps, each of them performing a particular task. However, sometimes using a programming language is required because of a specific need that can only be met via custom coding. In this sense, Python is becoming very popular as a software providing libraries to set up an ETL pipeline such as pandas[44] and Apache Airflow.[45]

Institutions such as the World Wide Web Consortium (W3C) recommends the use of machine-readable standardized data formats, providing data in multiple formats and reusing vocabularies.[5,46] Making data available in machine-readable relates to the use of standardized data formats that are easily parseable including but not limited to CSV, XML, JSON and RDF.

3.4 Data enrichment

Data enrichment refers to a set of manual and automatic processes to enhance and refine a dataset. Some examples include enriching data by generating new data, as well as, to incorporate links from external repositories such as Wikidata and GeoNames.[22,46]

Many are the techniques for data enrichment that can be used based on machine learning such as disambiguation, entity recognition and sentiment analysis, among others. New data values may be derived from the data originally provided such as word frequencies, part-of-speech-tagged tokens, and token counts.

3.5 Package dataset

Although there is no a general form to publish collections as data, there are some recommendations to encourage researchers to engage such as providing license information and detailed documentation including examples of reuse.[22]

In addition to publishing the data itself, it is fundamental to provide a human and machine readable description of the data, including its structure – for instance, the name of the columns–, and guidance on how to use it. In this sense, a dataset should be accompanied by a README plain text file that documents information about how to find, use and interpret the data.

The Data Package[47] specification –developed by Open Knowledge International– is a format aimed at data publishing that provides an interface to datasets, especially those containing tabular data (e.g. CSV). The specification defines a descriptor as the main file in a Data Package providing general metadata (e.g. title, author, license, etc.), a list of data resources, including their location in disk and structure. Several software libraries in different programming languages have been developed that can interpret and read the Data Package specification.

Data packages can be automatically validated in terms of structure and contents identifying potential issues by means of tools such as goodtables.io.[48] Sharing the data is crucial to create the highest impact in the research community.

3.6 Analysis

This step includes the use of the dataset by means of the creation of tools and services, but also to conduct research by applying emergent trends in order to fully exploit the potential of the digital collections. For instance, providing prototypes based on digital methods including topic modeling and text mining, and including data visualizations that ease the discovery of hidden data patterns.

The Collections as data final report recommends to share samples projects, ideas and create training workshops addressed to researchers to introduce them how the collections can be relevant for research.[22]

4 Application of the framework using real digital collections

This section introduces 6 use cases to create Collections as data from datasets published by GLAM institutions. We then tested the datasets by employing computational methods such as text mining, topic modeling and GIS.

The variety of options adopted by GLAM institutions in order to publish digital collections may hindering their reuse. In general, metadata repositories, including LOD, provide the description of resources that can be used in several ways such as GIS and data visualization. Metadata can include links to externals files such as images and text. In particular, LOD repositories are more appropriate for network graphs due to its design as a graph by means of RDF. On the other hand, datasets including corpora text are best suited for other methods such as topic modeling and text mining. The main features of the datasets reused in the application of the framework as well as the research methods and transformations applied are illustrated in Table 4.

This approach is based on the methodology proposed in Section 3 in order to extend the research value of the digital collections and encourage GLAM institutions to embrace Collections as data as a core activity.

GLAM institutions provide open datasets in different formats and types. Each dataset may require different transformation to address the issue involved (see Table 4). In this sense, we have selected the datasets according to the following criteria: (i) available as open access, (ii) containing metadata, images and text, or a combination of all three, and (iii) available as a dump or by means of an API. In the particular case of dumps, the size of the dataset has been taken into consideration (i.e. megabytes) in order to support the reproducibility in a limited period of time.

Since the research community have highlighted the need for reproducible research by providing articles, as well as, data and code, we have decided to use Jupyter Notebooks. The project is available in GitHub[§] as a collection of interactive notebooks classified by content and the code is reproducible in an executable environment in the cloud based on Binder.[¶] The notebook collections have been assigned a Digital Object Identifier (DOI) with the data archiving tool Zenodo.

Regarding the code and programming skills, language is a contentious subject, and many good choices are possible. While Jupyter Notebooks support several programming languages, we have used Python since it is easy to learn and offer a relatively low barrier to enter for researchers. Python continues to be the most preferred language for scientific computing, data science and machine learning.[49,50] The notebooks use specialized tools for handling data such as NumPy[51] and pandas, Matplotlib[52] for visualizations and Data Package to create the data packages. In addition, more common packages are used such as the package requests to make HTTP requests, and the packages CSV and JSON to read and write data in CSV and JSON formats.

4.1 Increasing the value of LOD through interactive maps

Introduction. The BNB Linked Data Platform[53] provides access to the British National Bibliography (BNB) published as LOD and made available through SPARQL services.

The dataset has been modeled upon RDF vocabularies, such as Dublin Core, the Bibliographic Ontology (BIBO) and Friend of a Friend (FOAF). The full dataset is available for download.

Librarians and academic users are interested in filtering works by place, country or location.[33,54] For example, the following is a typical scenario: find all books published in London in the twentieth century. A traditional digital catalogue may often have this information as textual content. However, LOD allows to include links to external repositories such as GeoNames and Wikidata that can be reused to obtain additional information regarding the locations.

Application. This notebook explains how to query a LOD repository in order to retrieve places of publication (fields `blt:publication` and `blt:projectedPublication`). As the works are linked to GeoNames, the records can be linked to external repositories. This notebook obtains information from Wikidata (see Figure 1 and Figure 2), showing a practical example of exploitation of LOD.

The dataset generated is finally used to create an interactive map based on Folium[55] in which the publication places are shown (see Figure 4).

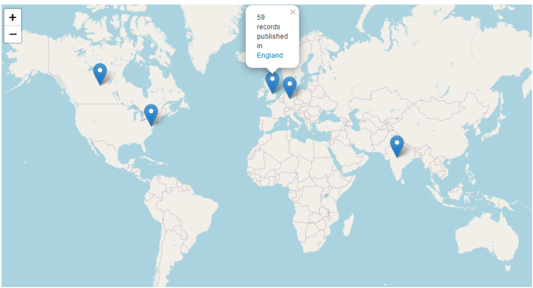


Figure 4. Interactive map representing the publication places of works written by William Shakespeare published as LOD in the BNB.

4.2 Exploiting covers: A zoomable mosaic from most relevant authors at BVMC

Introduction. The potential for reusing LOD repositories is limited to some extent due to the use of complex

[§] https://github.com/hibernator11/notebook-iiif-images, https://github.com/hibernator11/notebook-lod-libraries, https://github.com/hibernator11/notebook-texts-example,

[¶] https://mybinder.org/

Table 4. Main features of the datasets used to apply the framework. Note that LOD repositories may include metadata in several languages.

Dataset	Language	Type	Access	Method	Transformations
BNB Linked Data Platform	English	Metadata - LOD	SPARQL	GIS	Metadata parsing and enrichment
data.cervantesvirtual.com	Spanish	Metadata - LOD	SPARQL	Visualization	Metadata parsing and image treatment
data.bnf.fr	French	Metadata - LOD	SPARQL	Network graph	Metadata parsing
BL digitised theatrical playbills	English	Text - OCR	Dump	Topic modeling	Removing punctuation,lowercase words, stopwords
Smithsonian	English	Images	API - IIIF	Face detection	Metadata parsing, image processing
Moving Image Archive	English	Metadata - MARC	Dump	Visualization	Parsing metadata

technologies. However, LOD repositories contain not only meaningful metadata, but also images such covers of books which are hidden to most users.

Application. This notebook starts by using the SPARQL API from `data.cervantesvirtual.com/sparql` to retrieve the bibliographic records, including the covers. Following other approaches,[26] it creates a thumbnail for each work retrieving its cover based on an a list of authors, and obtaining as a result a folder full of thumbnails.

The next stage is creating a mosaic using all the covers. The output is accessible at easyzoom (see Figure 5).[‖]

The SPARQL query is configured to retrieve the covers of the authors Miguel de Cervantes, Lope de Vega and Calderón de la Barca (see Figure 6). The number of results was configured to 2500 in the SPARQL to retrieve a minimum number of covers to create the mosaic.

This notebook uses the Python Imaging Library (PIL)[56] to process the images. It is important to take into account the number and size of the images to create the mosaic.

4.3 Analysing the editions of ”Les fleurs du mal de baudelaire” from data.bnf.fr

Introduction. The Bibliothèque nationale de France published data.bnf.fr including resources by aggregating information scattered among its various catalogues and the Gallica digital library on dedicated HTML pages. Data.bnf.fr is an open data project based on semantic web standards and tools, based on Functional Requirements for Bibliographic Records (FRBR) to describe the resources which are identified using Archival Resource Key (ARK) identifiers. The information is stored in different formats, including RDF, JSON and HTML.

The visualization of graphs has proven to be very useful for exploring relationships in different application domains including culture heritage. Several approaches have been proposed to reveal meaningful connections between works and data related to the personal and professional lives of the creators.[57] In this sense, the RDF is a graph-based representation format for data publishing. Graphs are structures that map relations between objects that in RDF are represented by resources. The objects are referred to as nodes and the connections between them as edges.

Exploring data.bnf.fr contents, we identified that the particular case of the work *Les fleurs du mal de baudelaire*, includes several editions contributed by several authors and published in several years and languages. Data.bnf.fr provides the editions of *Les fleurs du mal de baudelaire* as well as the contributors, including their place of birth. The Wikidata entity[**] also reflects this situation.

Application. In particular, we explore attributes related to the personal and professional lives of the contributors. Data.bnf.fr describes authors with several properties such as `foaf:name`, and `fieldOfActivityOfThePerson` and `placeOfBirth`, both defined in the vocabulary `rdagroup2elements`. The SPARQL query in Figure 7 shows how the metadata is retrieved from data.bnf.fr.

Once the metadata is ready for use, we explore the relationships between the contributors of the editions and the attributes. Figure 8 shows how the contributors of the editions are related according to the field of activity attribute. The most relevant activity *Littératures* is shown in the centre of the graph.

The Python package NetworkX[††] is used for the creation and study of the structure of complex networks.

[‖] `https://www.easyzoom.com/imageaccess/acab10e5187b45d8ad802d8d302a4901`

[**] `https://www.wikidata.org/wiki/Q216578`

[††] `https://networkx.github.io/`

Figure 5. Mosaic built using covers retrieved from the Biblioteca Virtual Miguel de Cervantes.

```
SELECT ?m ?cover ?label
WHERE { ?m rdfs:label ?label .
 ?m foaf:depiction ?cover .
 ?m rdam:workManifested ?w .
 {
  { ?w rdaw:author <.../person/40>}
    UNION
  { ?w rdaw:author <.../person/72>}
    UNION
  { ?w rdaw:author <.../person/79>}
 }
}
```

Figure 6. A SPARQL query retrieving the works and covers of Miguel de Cervantes, Lope de Vega and Calderón de la Barca from Biblioteca Virtual Miguel de Cervantes LOD repository –dots stand for the common prefix data.cervantesvirtual.com.

4.4 *Implementing a topic model based on digitised theatrical playbills*

Introduction. Topic Modeling is an approach to understand and summarize large collections of text data. The goal is to find a sets of words that frequently occur together called *topics* in a text document. These topics can be used to describe the entire document.

There are many approaches for obtaining topics from a text document such as Latent Dirichlet Allocation (LDA). For instance, the MAchine Learning for LanguagE Toolkit (MALLET) provides a Java-based package for topic modeling and other machine learning applications to text.[58] Additional developments include a graphical user interface for MALLET to be run as a native application.[59] Other implementations are based on other programming languages such as Gensim[‡‡] Python library.

This example is based on the a dataset provided by the British Library that comprises 264 volumes of digitised theatrical playbills published between 1660 and 1902 (mostly 19th century) from England, Scotland, Wales and Ireland. Digitised from the British Library's physical collection of over 500 volumes of playbills, the dataset contains text files in OCR format.[60]

Application. A preprocessing on the original OCR text is performed in order to make the texts more amenable for analysis and obtain reliable results. Regular expressions are used to remove punctuation marks, lowercase the text and remove stopwords. In addition, non existent English words have been removed. Figure 9 shows an overview of the text provided and how it has been processed.

The LDA model is then created based on Gensim Python library. As a result, based on the text provided, the numbers of topics requested and the words within each topic which best describes them are created (see Figure 10). Each topic and their corresponding words are related to a common theme (e.g., topic 2 is related to bologna and money).

4.5 *Face detection based on Smithsonian collections*

Introduction. Recently, significant progress has been made in the field of face detection that tries to identify human faces in digital images and videos. Different approaches have been introduced in order to provide accurate and efficient methods.[61]

This example is based on the dataset published by the Smithsonian Institution Archives that recently has implemented IIIF. The images are identified by self described URIs including the Server, Prefix, Identifier, Region, Size, Rotation and Quality. The API provides a search function to retrieve the images.[19] The collection

[‡‡] https://radimrehurek.com/gensim/

```
SELECT DISTINCT ?edition ?titre ?date ?editeur
       ?contributor ?contributor_name ?activity ?placeOfBirth
WHERE {
  <http://data.bnf.fr/ark:/12148/cb11947965f> foaf:focus ?Oeuvre .
  ?edition rdarelationships:workManifested ?Oeuvre.
  OPTIONAL{
    ?edition dcterms:date ?date
  }OPTIONAL{
    ?edition dcterms:title ?titre
  }OPTIONAL{
    ?edition dcterms:publisher ?editeur
  }OPTIONAL{
    ?edition rdarelationships:expressionManifested ?exp.
    ?exp dcterms:contributor ?contributor.
    ?contributor foaf:name ?contributor_name .
    ?contributor rdagroup2elements:fieldOfActivityOfThePerson ?activity .
    FILTER (!regex(str(?activity), "dewey", "i")) .
    ?contributor rdagroup2elements:placeOfBirth ?placeOfBirth
  }
}
```

Figure 7. A SPARQL query retrieving the editions of *Les fleurs du mal de baudelaire* from data.bnf.fr including several attributes such as place of birth and field of activity.

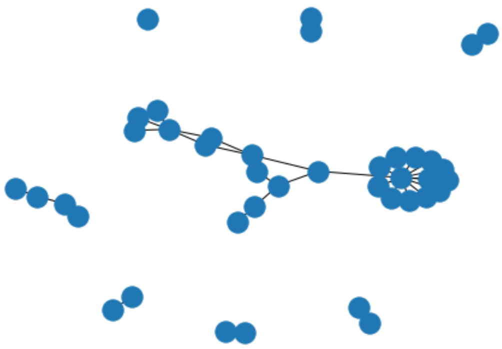


Figure 8. Network graph to visualise how contributors of the editions are related according to the field of activity attribute. The most relevant activity *Littératures* is shown in the center of the graph.

includes several themes such Art, Design, History, Culture and Science.

Application. The first step consists of retrieving the images of the digital repository by using its API. We have used the query string *Theodore Roosevelt* in order to retrieve portraits based on him. The results are based on a JSON file which is preprocessed in order to identify the portratis. Some analysis of the results is made by means of pandas such as the authors and types of documents. Figure 11 shows an overview of the images obtained as a result.

Then, several Python libraries are used in order to apply face detection to the images. Open Source Computer Vision Library[62] (OpenCV) is an open source computer vision and machine learning software library. In this example, images are treated as an standard NumPy array containing pixels of data points. Figure 12 presents an example of face detection based on an image from the repository.

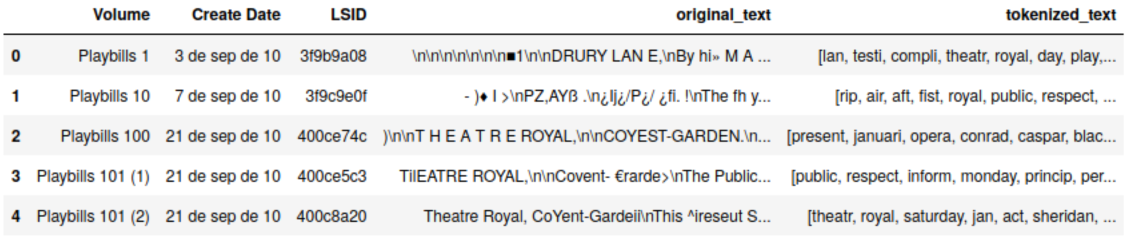
4.6 Scotland's national collection of moving images

Introduction. Machine Readable Cataloging (MARC) has been traditionally used by libraries to describe their collections. Many library systems are based on MARC[63] standards such as MARCXML.

This example is based on the Moving Image Archive dataset published by the Data Foundry in the National Library of Scotland. The dataset is based on the Scotland's national collection of moving images and includes more than six thousand records. The metadata is described in MARCXML, as well as Dublin Core.

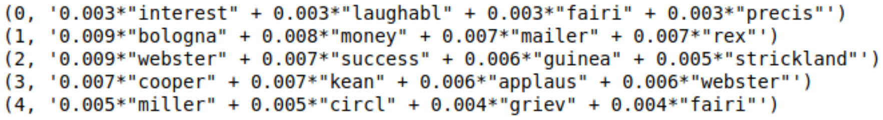
Application. The metadata describing the records is read in order to extract the most relevant information. MARC records are described by means of fields and subfields. The Python library Pymarc[64] is used to retrieve the metadata. Figure 13 shows an overview of the Python code to extract the metadata.

Then, Python data analysis library pandas is used to analyse the records such as identifying how often is a topic used or the number of records. In addition, a list of unique topics is provided. Figure 14 shows the most representative topics in the dataset.



	Volume	Create Date	LSID	original_text	tokenized_text
0	Playbills 1	3 de sep de 10	3f9b9a08	\n\n\n\n\n\n\n■1\n\nDRURY LAN E,\nBy hi» M A ...	[lan, testi, compli, theatr, royal, day, play,...
1	Playbills 10	7 de sep de 10	3f9c9e0f	-)♦ I >\nPZ,AYß .\n¿lj¿/P¿/ ¿fi. !\nThe fh y...	[rip, air, aft, fist, royal, public, respect, ...
2	Playbills 100	21 de sep de 10	400ce74c	)\n\nT H E A T R E ROYAL,\n\nCOYEST-GARDEN.\n...	[present, januari, opera, conrad, caspar, blac...
3	Playbills 101 (1)	21 de sep de 10	400ce5c3	TiIEATRE ROYAL,\n\nCovent- €rarde>\nThe Public...	[public, respect, inform, monday, princip, per...
4	Playbills 101 (2)	21 de sep de 10	400c8a20	Theatre Royal, CoYent-Gardeii\nThis ^ireseut S...	[theatr, royal, saturday, jan, act, sheridan, ...

Figure 9. Overview of the theatrical playbills dataset from British Library including a column with a clean version of the text.



```
(0, '0.003*"interest" + 0.003*"laughabl" + 0.003*"fairi" + 0.003*"precis"')
(1, '0.009*"bologna" + 0.008*"money" + 0.007*"mailer" + 0.007*"rex"')
(2, '0.009*"webster" + 0.007*"success" + 0.006*"guinea" + 0.005*"strickland"')
(3, '0.007*"cooper" + 0.007*"kean" + 0.006*"applaus" + 0.006*"webster"')
(4, '0.005*"miller" + 0.005*"circl" + 0.004*"griev" + 0.004*"fairi"')
```

Figure 10. Topics and words obtained after applying the LDA model to the theatrical playbills dataset from British Library.

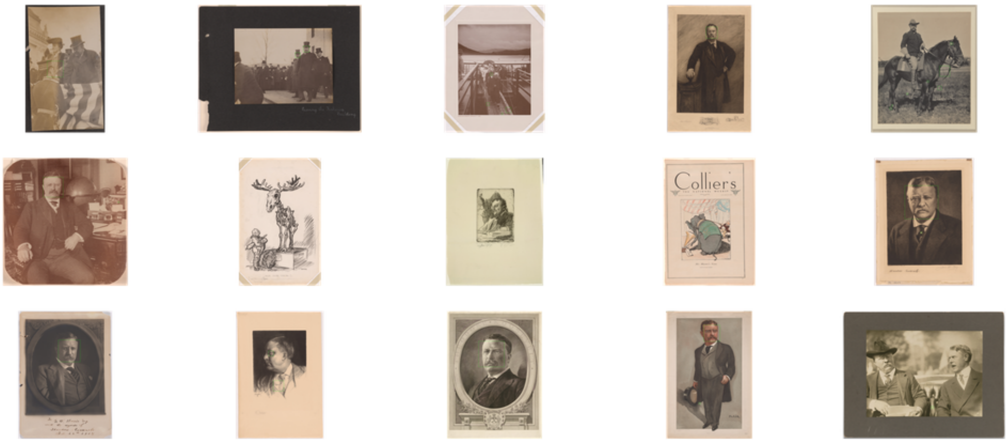


Figure 11. Overview of the portraits retrieved from the Smithsonian Institution Archives according to the querystring *Theodore Roosevelt*.

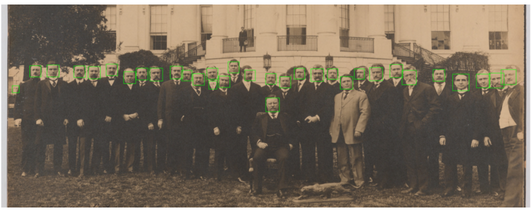


Figure 12. Face detection output based on a Smithsonian Institution Archives image.

5 Conclusions

GLAM institutions are starting to adopt Collections as data in order to publish datasets for reuse in innovative and inspiring ways.

The methodology described in Section 3 provides a set of steps to create machine actionable datasets within a GLAM institution. The application of the methodology described in Section 4 provides a comprehensive picture of how digital collections can be transformed into Collections as data. Six relevant repositories have been reused from GLAM institutions, including six Jupyter Notebooks available in GitHub. These examples are useful to promote the adoption of Collections as data within GLAM institutions.

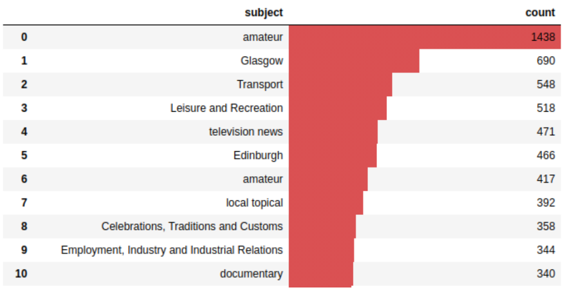
Although some GLAM institutions provide datasets that support computationally-driven research, the quality of the texts and metadata provided are crucial to obtain good results. Legacy OCR outputs may not have enough quality for new algorithms. Preprocessing and cleaning is a crucial task in order to prepare the data for computationally-driven research. In this sense, regular expressions and Python packages such as NLTK and Pandas provide a preliminary set of tools with which to process the original datasets.

```
# title
if record['245'] is not None:
  title = record['245']['a']
  if record['245']['b'] is not None:
    title +=" "+record['245']['b']

# determine author
if record['100'] is not None:
  author = record['100']['a']
elif record['110'] is not None:
  author = record['110']['a']
elif record['700'] is not None:
  author = record['700']['a']
elif record['710'] is not None:
  author = record['710']['a']

# place_production
if record['264'] is not None:
  place_prod = record['264']['a']
```

Figure 13. Overview of the Python code to extract metedata described using MARCXML.



	subject	count
0	amateur	1438
1	Glasgow	690
2	Transport	548
3	Leisure and Recreation	518
4	television news	471
5	Edinburgh	466
6	amateur	417
7	local topical	392
8	Celebrations, Traditions and Customs	358
9	Employment, Industry and Industrial Relations	344
10	documentary	340

Figure 14. Overview of the most representative topics in the Moving Image Archive.

Jupyter notebooks lower many barriers to reproducibility by enabling scientists to combine code, results and text. However, depending on the type of dataset and materials provided, some steps may require a significant amount of time to complete due to several reasons including the large size files or the use of external APIs.

LOD repositories are a rich source of information including images and metadata. Although its uses requires the knowledge of many technologies such as RDF and SPARQL, several datasets can be extracted from large repositories to create amenable datasets according to an author or a theme. This datasets can be combined and enriched by means of external repositories such as Wikidata and GeoNames.

With the advent of technology, face detection has become popular in the research community. The efficiency of face detection algorithms is crucial regarding its performance, which pose serious challenges to current models in terms of time and accuracy.

Traditional metadata formats are a rich source of information that can be used to extract and discover new knowledge. However, CSV formats are more friendly for general users. In addition, Python libraries such as pandas can be easily adopted to obtain a summary of the dataset in terms of features such as most relevant topics, authors and dates.

Future work to be explored includes the further generalization and automation of the creation procedures and the inclusion and redefinition of more steps. In addition, including more languages to provide the notebooks will be explored such as Spanish and Russian, as well as, to provide additional notebooks based on additional datasets.

Acknowledgements

The authors would like to acknowledge Sally Chambers, Digital Humanities Research Coordinator at Ghent Centre for Digital Humanities, Ghent University, Belgium and National Coordinator for DARIAH-BE for her support in writing this paper.

This work has been partially supported by ECLIPSE-UA RTI2018-094283-B-C32 (Spanish Ministry of Education and Science)

6 References

References

1. Baillieul B John, Hall O Larry, Moura M Jose et al. The first IEEE workshop on the Future of Research Curation and Research Reproducibility, 2017. URL https://open.bu.edu/handle/2144/39028.
2. Baillieul J, Grenier G and Setti G. Reflections on the Future of Research Curation and Research Reproducibility [Point of View]. *Proceedings of the IEEE* 2018; 106(5): 779–783.
3. Research Libraries UK. A manifesto for the digital shift in research libraries. https://www.rluk.ac.uk/digital-shift-manifesto/, 2020. [Online; accessed 20-April-2020].
4. Mahey M, Al-Abdulla A, Ames S et al. *Open a GLAM lab*. 2019. ISBN 978-9927-139-07-9.
5. Padilla T, Allen L, Frost H et al. Final Report — Always Already Computational: Collections as Data, 2019. DOI: 10.5281/zenodo.3152935. URL https://doi.org/10.5281/zenodo.3152935.
6. International Image Interoperability Framework. URL https://iiif.io/.
7. Auer S, Bizer C, Kobilarov G et al. DBpedia: A nucleus for a web of open data. In Aberer K, Choi K, Noy NF et al. (eds.) *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, *Lecture Notes in Computer*

Science, volume 4825. Springer, pp. 722–735. DOI: 10.1007/978-3-540-76298-0_52. URL https://doi.org/10.1007/978-3-540-76298-0_52.

8. Tanon TP, Vrandecic D, Schaffert S et al. From freebase to wikidata: The great migration. In Bourdeau J, Hendler J, Nkambou R et al. (eds.) *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. ACM, pp. 1419–1428. DOI:10.1145/2872427.2874809. URL https://doi.org/10.1145/2872427.2874809.
9. Rule A, Birmingham A, Zuniga C et al. Ten Simple Rules for Reproducible Research in Jupyter Notebooks. *CoRR* 2018; abs/1810.08055. URL http://arxiv.org/abs/1810.08055. 1810.08055.
10. Romero GC, Esteban MPE, Carrasco RC et al. Migration of a library catalogue into RDA linked open data. *Semantic Web* 2018; 9(4): 481–491. DOI:10.3233/SW-170274. URL https://doi.org/10.3233/SW-170274.
11. IFLA Information Technology Section ; IFLA Semantic Web Special Interest Group ; Bibliothèque nationale de France. *We grew up together: data.bnf.fr from the BnF and Logilab perspectives*. Paris, Bibliothèque nationale de France, Petit auditorium: IFLA Information Technology Section ; IFLA Semantic Web Special Interest Group ; Bibliothèque nationale de France, 2014. URL http://ifla2014-satdata.bnf.fr/program.html.
12. Wilkinson M, Dumontier M, Aalbersberg I et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data* 2016; 3. DOI:10.1038/sdata.2016.18.
13. Harris G, Potter A, Zwaard K et al. Digital Scholarship at the Library of Congress. User demand, current practices, and options for expanded services. https://labs.loc.gov/static/labs/meta/DHWorkingGroupPaper-v1.0.pdf, 2020. [Online; accessed 20-April-2020].
14. Association of European Research Libraries. Implementing FAIR Data Principles: The Role of Libraries. https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf, 2017. [Online; accessed 21-April-2020].
15. National Library of Scotland. Data Foundry. Data collections from the National Library of Scotland. URL https://data.nls.uk/.
16. British Library. A collection of datasets released by the British Library. URL https://data.bl.uk/.
17. Library of Congress. LC for Robots. URL https://labs.loc.gov/lc-for-robots/.
18. Freire N, Robson G, Howard JB et al. Cultural heritage metadata aggregation using web technologies: Iiif, sitemaps and schema.org. *Int J on Digital Libraries* 2020; 21(1): 19–30. DOI:10.1007/s00799-018-0259-5. URL https://doi.org/10.1007/s00799-018-0259-5.
19. Smithsonian. International Image Interoperability Framework at the Smithsonian Institution. URL https://iiif.si.edu.
20. Europeana. Europeana IIIF APIs. URL https://pro.europeana.eu/page/iiif.
21. Staatsbibliothek zu Berlin. Staatsbibliothek zu Berlin Labs. URL https://lab.sbb.berlin/dc/?lang=en.
22. Padilla T, Allen L, Frost H et al. 50 Things — Always Already Computational: Collections as Data, 2019. DOI: 10.5281/zenodo.3066237. URL https://doi.org/10.5281/zenodo.3066237.
23. Det Kgl Bibliotek. Det Kgl. Bibliotek Labs. URL https://labs.kb.dk/.
24. Wittmann R, Neatrour A, Cummings R et al. From digital library to open datasets. *Information Technology and Libraries* 2019; 38(4): 49–61. DOI:10.6017/ital.v38i4.11101. URL https://ejournals.bc.edu/index.php/ital/article/view/11101.
25. Project Jupyter. URL https://jupyter.org/.
26. Sherratt T. Glam-workbench/getting-started, 2019. DOI: 10.5281/zenodo.3549636. URL https://doi.org/10.5281/zenodo.3549636.
27. Deschamps R, Ruest N, Lin J et al. The archives unleashed notebook: Madlibs for jumpstarting scholarly exploration of web archives. In *19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, June 2-6, 2019*. pp. 337–338. DOI:10.1109/JCDL.2019.00059. URL https://doi.org/10.1109/JCDL.2019.00059.
28. Library of Congress. Chronicling America Data Visualizations, 2019. URL https://www.loc.gov/ndnp/data-visualizations/.
29. Library of Congress. By the People. URL https://crowd.loc.gov/.
30. Standards C and Board C. CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022, 2019. DOI:10.5281/zenodo.3632533. URL https://doi.org/10.5281/zenodo.3632533.
31. Robledano-Arillo J, Bonilla DN and Cerdá-Díaz J. Application of linked open data to the coding and dissemination of spanish civil war photographic archives. *Journal of Documentation* 2020; 76(1): 67–95. DOI:10.1108/JD-06-2019-0112. URL https://doi.org/10.1108/JD-06-2019-0112.
32. Villazón-Terrazas B, Vilches-Blázquez LM, Corcho O et al. *Methodological Guidelines for Publishing Government Linked Data*. New York, NY: Springer New York. ISBN 978-1-4614-1767-5, 2011. pp. 27–49. DOI:10.1007/978-1-4614-1767-5_2. URL https://doi.org/10.1007/978-1-4614-1767-5_2.
33. Romero GC, Esteban MPE, Carrasco RC et al. A linked open data framework to enhance the discoverability and impact of culture heritage. *J Inf Sci* 2019; 45(6). DOI:10.1177/0165551518812658. URL https://

doi.org/10.1177/0165551518812658.

34. World Wide Web Consortium. RDF 1.1 Primer, 2014. URL https://www.w3.org/TR/rdf11-primer/.
35. OpenRefine. URL https://openrefine.org/.
36. Esteban MPE, Romero GC, Trujillo J et al. Adding value to linked open data using a multidimensional model approach based on the RDF data cube vocabulary. *Comput Stand Interfaces* 2020; 68. DOI:10.1016/j.csi.2019.103378. URL https://doi.org/10.1016/j.csi.2019.103378.
37. Jevon G. Clean. Migrate. Validate. Enhance. Processing Archival Metadata with Open Refine, 2020. URL https://blogs.bl.uk/digital-scholarship/2020/04/clean-migrate-validate-enhance-processing-archival-metadata-with-open-refine.html.
38. McPhillips T, Li L, Parulian N et al. Modeling provenance and understanding reproducibility for openrefine data cleaning workflows. In *11th International Workshop on Theory and Practice of Provenance (TaPP 2019)*. Philadelphia, PA: USENIX Association. URL https://www.usenix.org/conference/tapp2019/presentation/mcphillips.
39. Ferriter M. Introducing the Computing Cultural Heritage in the Cloud Project, 2019. URL https://blogs.loc.gov/thesignal/2019/11/introducing-the-computing-cultural-heritage-in-the-cloud-project/?loclr=blogsig.
40. European Comission. Modernisation of the EU copyright rules, 2019. Https://ec.europa.eu/digital-single-market/en/modernisation-eu-copyright-rules.
41. Wilkinson MD, Dumontier M, Aalbersberg IJ et al. The fair guiding principles for scientific data management and stewardship. *Scientific data* 2016; 3.
42. Open Archives Initiative. Open Archives Initiative Protocol for Metadata Harvesting. URL https://www.openarchives.org/pmh/.
43. World Wide Web Consortium. SPARQL 1.1 Query Language, 2013. URL https://www.w3.org/TR/sparql11-query/.
44. Pandas - Python Data Analysis Library. URL https://pandas.pydata.org/.
45. Hagedorn S. When sweet and cute isn't enough anymore: Solving scalability issues in python pandas with grizzly. In *CIDR 2020, 10th Conference on Innovative Data Systems Research, Amsterdam, The Netherlands, January 12-15, 2020, Online Proceedings*. URL http://cidrdb.org/cidr2020/gongshow2020/gongshow/abstracts/cidr2020_abstract76.pdf.
46. World Wide Web Consortium. Data on the Web Best Practices, 2017. URL https://www.w3.org/TR/dwbp/.
47. Walsh P and Pollock R. Data Package, 2007. URL https://specs.frictionlessdata.io/data-package.
48. Open Knowledge Foundation. goodtables.io. URL http://goodtables.io/.
49. Raschka S, Patterson J and Nolet C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *CoRR* 2020; abs/2002.04803. URL https://arxiv.org/abs/2002.04803. 2002.04803.
50. Koerner L, Caswell TA, Allan DB et al. A python instrument control and data acquisition suite for reproducible research. *IEEE Trans Instrumentation and Measurement* 2020; 69(4): 1698–1707. DOI:10.1109/TIM.2019.2914711. URL https://doi.org/10.1109/TIM.2019.2914711.
51. Numpy. URL https://numpy.org/.
52. Matplotlib: Visualization with python. URL https://matplotlib.org/.
53. British Library. Basic RDF/XML. "http://www.bl.uk/bibliographic/datafree.html#basicrdfxml", 2014. [Online; accessed 26-April-2020].
54. Bontcheva K, Kieniewicz J, Andrews S et al. Semantic enrichment and search: A case study on environmental science literature. *D-Lib Mag* 2015; 21(1/2). DOI:10.1045/january2015-bontcheva. URL https://doi.org/10.1045/january2015-bontcheva.
55. Folium. URL https://python-visualization.github.io/folium/.
56. Pillow. URL https://pillow.readthedocs.io/en/stable/.
57. Nurmikko-Fuller T, Bangert D, Hao Y et al. Swinging triples: Bridging jazz performance datasets using linked data. In *Proceedings of the 1st International Workshop on Semantic Applications for Audio and Music, SAAM@ISWC 2018, Monterey, CA, USA, October 9, 2018*. pp. 42–45. DOI:10.1145/3243907.3243914. URL https://doi.org/10.1145/3243907.3243914.
58. McCallum AK. Mallet: A machine learning for language toolkit, 2002. Http://mallet.cs.umass.edu.
59. Enderle JS, Balagopalan A, Li X et al. senderle/topic-modeling-tool: First stable release, 2017. DOI:10.5281/zenodo.496150. URL https://doi.org/10.5281/zenodo.496150.
60. British Library. Theatrical playbills from Britain and Ireland (OCR text only), 2015. URL https://doi.org/10.21250/pb2.
61. Hou S, Li Y, Pan Y et al. A face detection algorithm based on two information flow block and retinal receptive field block. *IEEE Access* 2020; 8: 30682–30691. DOI:10.1109/ACCESS.2020.2973071. URL https://doi.org/10.1109/ACCESS.2020.2973071.
62. OpenCV. URL https://opencv.org/.

63. Library of Congress. Marc standards. URL http://www.loc.gov/marc/.

64. Pymarc. URL https://pymarc.readthedocs.io/en/latest/.