A benchmark of Spanish language datasets for computationally-driven research

Journal Title
XX(X):1–11

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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

Abstract

In the domain of Galleries, Libraries, Archives and Museums (GLAM) institutions, creative and innovative tools and methodologies for content delivery and user engagement have recently gained international attention. New methods have been proposed to publish digital collections as datasets amenable to computational use. Standardized benchmarks can be useful to broaden the scope of machine-actionable collections and to promote cultural and linguistic diversity. In this article, we propose a methodology to select datasets for computationally-driven research applied to Spanish text corpora. This work seeks to encourage Spanish and Latin American institutions to publish machine-actionable collections based on best practices and avoiding common mistakes.

Keywords

GLAM Labs, Collections as data, Data Quality Metrics, Digital Libraries

1 Introduction

Cultural heritage institutions have traditionally provided access to digital collections. They are an excellent example of public engagement, bringing together materials, people and services with a multidisciplinary perspective. The materials represent rich sources of information that include text, maps, images, metadata, video and audio, among others. Digital collections differ in several ways: for example, in terms of copyright, the number of formats available and the accessing method, i.e., using an API or bulk downloads.

Meanwhile, Labs have emerged in GLAM institutions that work on the reuse of digital collections in inspiring and creative ways.[1] New scholarship programs encompassing all disciplines, such as Computer Science and Digital Humanities, are being adopted by GLAM institutions with the goal of improving their services by involving researchers and understanding how they use the data.[2] In addition, institutions are producing innovative models for supporting cloud-based research computing based on their digital collections and identifying requirements as well as possibilities. Examples include the Library of Congress, the National Library of the Netherlands and the National Library of Scotland. In this way, Labs can reinforce and maintain the relevance of GLAM institutions and their digital collections by engaging researchers.

GLAM institutions are starting to explore the benefits of new approaches to the publication of their digital collections to encourage computational use. Most of the documentation and examples of machine-actionable collections, however, are in English, including the text data.[3] In this sense, Spanish and Latin American institutions such as the Biblioteca Digital del Patrimonio Iberoamericano (BDPI),[4] and Mexicana, as well as, project-based initiatives are taking a step forward by making openly available digital materials. In order to foster machine-actionable collections in Spanish and Latin American institutions, best practices and guidelines are required to make their content available and reusable by researchers. Some efforts have recently been made regarding the translation of documentation into Spanish to encourage the use and publication of machine-actionable collections[5] as well as several research projects based on Spanish literature. Examples include Mnemosine and Unlocking the Colonial Archive.[6,7]

Digital collections often come in the form of hard-to-access data silos and this impedes their reuse by researchers. In addition, identifying a dataset for reuse

[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

is not an easy task for various reasons, such as copyright restrictions, coverage or quality.

In this regard, benchmarks provide an experimental process for comparing and assessing the performance of processes, services, databases and many other technologies with those regarded as the best. Benchmarking allows the identification of opportunities for improvement as well as the replication of the results. In this way, benchmarking can be adapted to datasets in order to identify the best datasets amenable to computationally-driven research. [8,9]

The purpose of this study was to introduce and extensible methodology to create a benchmark of digital collections amenable to computationally-driven research. The methodology was applied to several Spanish language datasets to encourage Spanish and Latin American institutions to publish machine-actionable collections based on best practices and avoiding common mistakes.

The main contributions of this paper are as follow: (a) a methodology for selecting datasets for computationally-driven research; (b) a benchmark of Spanish language datasets for computationally-driven research; and (c) the description of a practical and reproducible example of how to reuse the benchmark.

The paper is organized as described next. After a brief review of the state of the art in Section 2, Section 2.1 describes the methodology to create a benchmark of datasets. Section 3 introduces the benchmark of datasets for computationally-driven research, gives an example of reuse, based on a collection of Jupyter Notebooks and discusses the results. The paper concludes with an outline of the results, general guidelines on how to use the results and future work.

2 Background

For preservation purposes and to improve ease of access, cultural heritage institutions have digitized the vast and rich collections that represent cultural diversity. Digital technologies and the internet have unleashed unprecedented and unique opportunities to access the rich materials hosted by institutions as well as to create engaging programs to reuse the contents. [10,11]

Cultural heritage institutions have recently started to explore research applied to digital collections, based on computationally-driven methods. They are investigating the feasibility of data analytics approaches to improve the access to their digital collections. [12] New approaches such as Collections as Data provides a framework to create machine-actionable collections ready for reuse. [13] The Library of Congress (LC) recommends creating digital collections usable for computation as well as building institutional capacity for digital scholarship and for expanding user services. [14] The Online Computer Library Center (OCLC) recently published a study on community engagement with data science, machine learning, and artificial intelligence. [3]

Nevertheless, designing a sustainable data extraction workflow to publish machine-actionable collections is a challenging task. [15] The National Library of Scotland is exploring the opportunities and challenges of publishing datasets that support computational access including data management, rights and required skills. [16] Other approaches are based on datasets published by several relevant GLAM institutions including a detailed step-by-step guide. [17] KU Leuven Libraries are exploring new ways of creating, sharing and using the libraries' digitised collections as data. [18,19]

While the number of machine-actionable collections for computation has increased, most of them are hosted and published by large Western institutions, where the use of English predominates. [3] Standardized benchmarks can be useful to broaden the scope of machine-actionable collections and to promote cultural and linguistic diversity. They can also help practitioners select, reuse and improve the right datasets, and provide objective feedback to the research community. [20]

The identification of a dataset for reuse is not an easy task for various reasons, including vague copyright and terms of use, coverage, completeness, or ease of understanding. Even if the dataset is available, in some cases, it may require some preprocessing and cleaning to be ready for computational purposes. In addition, when working with large datasets, researchers can obtain manageable slices of the data.

In this sense, the LC Selected Datasets Collection provides an initial series of 20 datasets to support emerging styles of data-driven research, such as text mining and machine learning. [21] Chronicling America provides access to information about historic newspapers and a selection of digitized newspaper pages in the USA. [22] The publication of text of a collection of books in computer readable format was funded by the Faculty of Arts and Social Sciences and the Digital Humanities Hub of Lancaster University (UK). [23–25] A collection of datasets released by the British Library includes several openly available repositories. [26] In 2017, the Bibliothèque nationale de France (BnF) published *Bnf API et jeux de données*, including datasets and the API documentation. Mexicana is an open platform that provides access to available digital collections of the Ministry of Culture in Mexico. [27] GLAM Labs usually publish data openly and in reuseable data ready for computational use. Examples include the National Library of Scotland data, [28] the Austrian National Library [29] and the Dutch National Library. [30] Other approaches are based on Linked Open Data (LOD) using standard vocabularies and providing SPARQL [31] endpoints to access

the data.[32–34] However, LOD repositories published by libraries are mainly dedicated to publishing metadata retrieved from their main catalogues using several controlled vocabularies. Moreover, additional examples are based on international aggregators including BDPI, Europeana[35] and the Atlas of Digitised Newspapers and Metadata.[36]

Organizations, publishers and the community promote the sharing and reuse of datasets for research to encourage scientific progress. In this sense, several factors, such as sustainability, availability and discoverability have become crucial to support a collaborative research environment.[37] As a result, several platforms enable researchers to cite, locate and identify datasets, such as DataCite and Zenodo.

The final report of Collections as Data[38] recommends that institutions share prototypes and examples of use of their collections with the research community. The popularity of Jupyter Notebooks[39] has significantly increased in recent years. A notebook combines software code, multimedia resources, narrative text, visualizations and results in a single document that researchers can use and share. The combination of Jupyter Notebooks and machine-actionable collections provide an innovative and interactive environment for collaborative, transparent and reproducible data analyses.[17,40,41]

Although some approaches reuse datasets published by GLAM institutions, to the best our of knowledge, no benchmark of datasets for computationally-driven research exists based on Spanish text corpora. Benchmarks based on machine-actionable datasets are relevant because: (i) they help to compare the available datasets and to meet the needs of the users; (ii) researchers can address new challenges, improving the features and including new datasets; and (iii) organizations can benefit from shared best practices when publishing their datasets.[20]

2.1 *A methodology for selecting datasets for computationally-driven research*

The main goal of this study was to provide the research community with a benchmark to compare and evaluate machine-actionable datasets in cultural heritage institutions. Since the publication of digital collections has become popular and the number of datasets has increased, identifying candidates for the assessment, known as subjects, is an essential factor in a benchmark's success and performance. Other approaches propose methodologies to identify subjects that consider a variety of attributes ranging from more advanced technical issues to general cultural aspects.[42,43]

We defined our benchmark's criteria based on previous works.[20,44–46] Each feature can be given a score according to a criterion that consists of a function, with values ranging from 1-0. The definition of each criterion is described below.

Table 1. Possible scores according to the accuracy criterion.

Description	Score
OCR reviewed by curators	1
OCR reviewed by the community	0.5
OCR without editing	0.25
Otherwise	0

Licensing. In general, licenses range from very permissive with none or few obligations and known as open, to very restrictive or closed that include restrictions for reuse. The most permissive open licenses are Creative Commons CC0 1.0 Universal Public Domain Dedication,* and Public Domain Mark (PDM). Open licenses such as CC BY (Creative Commons Attribution License), CC BY-SA (Creative Commons Attribution-Share Alike) and other types require attribution and appropriate credit, as well as the indication of whether changes were made. Close licenses are less permissive and limit the usage. Other approaches are based on national policies regarding the publication of open data.† This criterion is defined as follows:

$$m_{\text{license}} = \begin{cases} 1 & \text{public domain licenses/CC0} \\ 0.5 & \text{open licenses (CC BY, CC BY-SA and other types)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Accuracy. Based on the literature,[47] this criterion determines the extent to which data are correct, reliable, and certified free of error.

Optical Character Recognition (OCR) is an automated process that transforms an image into computer-readable text. However, OCR is not 100 percent accurate, and may contain errors for various reasons, e.g., the use of small fonts.[48] Many institutions, such as the Library of Congress and Europeana are considering crowdsourcing approaches, thus allowing volunteers to create and review transcriptions to improve search and discovery.[49–52] As a result, this criterion is defined as shown in Table 1.

Provenance. The fulfillment of this criterion means that provenance is used to describe the

*http://creativecommons.org/publicdomain/zero/1.0/

†See, for example, https://data.bnf.fr/docs/Licence-Ouverte-Open-Licence-ENG.pdf.

creation process and the derived data. For instance, provenance information can be encoded by using the `dcterms:provenance` and `dcterms:source` properties in Dublin Core. This criterion is defined as follows:

$$m_{\text{provenance}} = \begin{cases} 1 & \text{provenance on dataset level} \\ 0.5 & \text{provenance on a website} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Language. Datasets are usually provided in the organization's original language. But sometimes the text is provided in several languages such as in the case of an international aggregator. Let A be the set of languages in which we are interested, then:

$$m_{\text{language}} = \frac{|\{x \epsilon A\}|}{|A|} \quad (3)$$

Permanent identifier. Regarding the identification of the datasets, several methodologies and platforms can be used. For instance, when using Zenodo, each dataset is assigned a DOI. This criterion is defined as follows:

$$m_{\text{identifier}} = \begin{cases} 1 & \text{permanent identifier provided} \\ 0.5 & \text{URL provided} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Prototypes and documentation. Providing prototypes and examples of use in addition to documentation can facilitate the reuse of the datasets by potential researchers.[38,53] In this sense, Jupyter Notebooks has become very popular in the community and has helped to lower barriers and include reproducible code as well as documentation.[17] This criterion is defined as follows:

$$m_{\text{examples}} = \begin{cases} 1 & \text{providing examples of use} \\ 0.5 & \text{providing documentation} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Formats. It is relevant to providing datasets in a variety of formats because it allows compatibility with commonly used methods and tools.[54,55] Machine readable formats can be automatically read and processed by a computer, such as CSV and TXT. However, organizations often provide PDF files that are not machine-readable, or that use proprietary formats, such as Microsoft Word (.doc).

The number of formats provided can be computed by exploring their websites as well as open science repositories such as Zenodo and FigShare. This criterion is defined as follows:

$$m_{\text{formats}} = \begin{cases} 1 & \text{machine-readable text and further formats are supported} \\ 0.75 & \text{machine-readable text is supported} \\ 0.5 & \text{text is supported} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Terms of use and code of conduct. Adding terms of use to the datasets is crucial to facilitate their reuse.[38] A code of conduct aims at ensuring a respectful and productive environment for reuse and research based on the datasets. These policies are applicable to all users and they may cover several aspects, such as the conditions of use, rules, responsibilities and proper practices.[‡] This criterion is defined as follows:

$$m_{\text{terms}} = \begin{cases} 1 & \text{providing terms of use and code of conduct} \\ 0.5 & \text{providing terms of use} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Technical aspects. Several technical aspects need to be considered including the use of an API such as a public endpoint SPARQL or the protocol OAI-PMH. This criterion is defined as follows:

$$m_{\text{technical}} = \begin{cases} 1 & \text{providing a public harvesting method.} \\ 0.5 & \text{providing a website} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The list of potential subjects can be evaluated using diverse techniques and methods. For instance, the alternatives to alternatives scorecard consists of a matrix in which candidates for benchmarking (known

[‡]See, for example, https://www.bl.uk/about-us/governance/policies/code-of-conduct

as alternatives) are shown in rows and attributes based on criteria are shown in columns. Another example is that of *polar charts*, which are circular graphs where rays associated to attributes are drawn from the centre of a circle and their length is proportional to the rating. The best choice would be the subject that covers the largest area.[45]

3 Benchmarking Spanish language datasets

This section introduces the datasets that will serve as benchmark. This approach is based on the methodology proposed in Section 2.1 in order to extend the research value of the digital collections, encourage GLAM institutions to embrace Collections as Data as a core activity, and to promote greater linguistic diversity in terms of the texts provided.

There is a wide range of means of publication of datasets that provides a machine-actionable collection ready for reuse. Approaches based on APIs enable reuse of data by multiple applications for different purposes (e.g., embedding images in HTML or enhancing images with transcriptions).[56] In addition, by using APIs the user is able to identify and download a slice of the dataset according to the requirements of the research to be conducted. Nevertheless, general APIs users may face the challenge of a steep learning curve. In addition, APIs can be vulnerable to attacks and additional resources are necessary in order to adopt security protocols and maintenance. Other approaches are based on conventional websites, as well as open and free platforms, such as GitHub and Zenodo. The latter provide a link to the dataset, including OCR text.

In the present case, we were interested in the Spanish language for the criterion m_{language} since it is the second-most spoken language in the world.[57] The analysis of how institutions handle and publish Spanish-language collections could help librarians and curators to improve their skills.[58]

Moreover, there is variety of reasons to exclude a dataset: full text lacking, the language of the text or copyrighted material.

A collection of Jupyter Notebooks based on the datasets provided by the benchmarking was created. The project is openly available in GitHub[§] as a collection of interactive notebooks and the code is runnable and reproducible in a cloud environment such as Binder.[59] The notebook collection was assigned a DOI with the data archiving platform Zenodo.[¶] Table 3 shows the main features of the datasets used in the Jupyter Notebooks collection. In addition, Figures 2, 3 and 4 show the results obtained after reusing the datasets.

3.1 Results

In order to find suitable subject datasets, we applied the methodology described in Section 2.1. We identified datasets provided by GLAM Labs, Google Public Datasets and Zenodo whose descriptions contained terms such as *library* or were included in Section 2. Some subjects were removed because they were out of date or because their URLs were invalid. International aggregators sometimes include items that are out of date.[‖] Table 2 presents a preliminary list of candidates.

We then used polar charts to identify which machine-actionable datasets were most suitable for the study. Every axis on the polar chart corresponds to one criterion. The global score is computed as the area of the polar chart –as shown in Figure 1 for Chronicling America. If the subject does not provide content in Spanish, the area is not computed.

As a result of the evaluation, six datasets (see grey cells in Table 2) were selected which support computationally-driven research and their contents are based on text in Spanish. Although the dataset features vary considerably among the datasets, these datasets all mainly publish metadata, images and full text.

The highest value was obtained by Chronicling America because this latter repository: provides its content in several languages, including Spanish; uses a permanent identifier; includes machine-readable text; and provides its data under the CC0 license. Mexicana, Corpus general de poesía lírica castellana del Siglo de Oro and Biblioteca Digital Hispánica obtained a very similar value, above 13. The three of them present their contents in Spanish and provide an URL to download the text. However, regarding licenses, Biblioteca Digital Hispánica offers its data under the CC0 license, while the other two provide the contents under a CC-BY license. The BDPI obtained the lowest value.

According to the evaluation results, only two datasets achieved the maximum criterion accuracy score. The reason may be that it is time-consuming for institutions to edit large text corpora.

3.2 Discussion

Regarding the use of open licenses, there is still room for improvement, since institutions tend to publish digital collections under CC-BY and other types of licenses. In some cases, the licenses were not clear, and were difficult to find or interpret. In this sense, Creative Commons and platforms such as FigShare and

[§] https://github.com/hibernator11/notebook-spanish-corpus

[¶] https://doi.org/10.5281/zenodo.3989221

[‖] See, for example, the download link at https://www.europeana.eu/es/item/2022705/lod_oai_www_bibliotecavirtualmadrid_org_8066_ent1

Table 2. Benchmark of datasets for computationally-driven research.

Subject	m_{license}	m_{accuracy}	$m_{\text{provenance}}$	m_{language}	$m_{\text{identifier}}$	m_{examples}	m_{formats}	m_{terms}	$m_{\text{technical}}$	Total
Austrian National Library - Historical Newspapers	0.5	0	0.5	0.33	0	0	0.75	0.5	0.5	-
Biblioteca Digital del Patrimonio Iberoamericano	0.5	0	0.5	0.67	0.5	0.5	0.5	0	0.5	5.03
Biblioteca Digital Hispánica	1	0	1	0.67	0.5	0.5	1	0.5	1	13.04
Bnf API et jeux de données	0,5	0	1	0.33	1	1	1	1	0.5	-
British Library datasets	0.5	0.25	1	0.33	1	1	1	0.5	1	-
Chronicling America	1	0.25	1	0.67	1	1	1	1	1	20.92
Corpus general de poesía lírica castellana del Siglo de Oro	0.5	1	1	0.67	0.5	1	1	0.5	1	13.25
Dutch National Library	0.5	1	1	0.33	1	1	1	0.5	0.5	-
Europeana Newspapers	0.5	0.25	1	0.33	1	1	1	1	1	-
Lancaster University - Transcripción del Catálogo Monumental de España	0.5	1	0	0.67	1	0.5	1	0.5	0.5	7.51
LC Selected Datasets	1	0.25	0.5	0.33	1	1	1	1	1	-
Mexicana	0.5	0.25	0.5	0.67	0.5	0.5	1	1	1	13.93
National Library of Scotland - Data Foundry	1	0.25	1	0.33	1	1	1	1	0.5	-

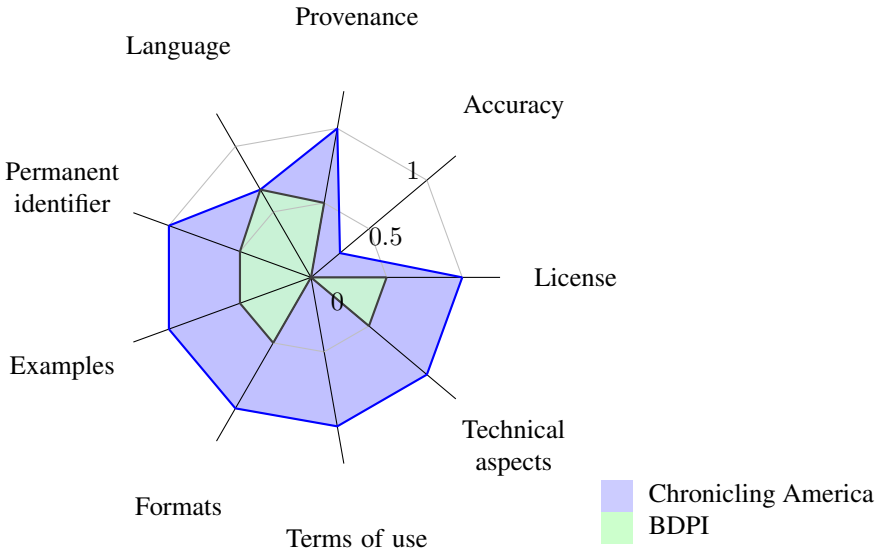

Figure 1. Polar chart that shows Chronicling America and Biblioteca Digital del Patrimonio Iberoamericano that obtained the highest (20.92) and lowest (5.03) scores, respectively.

Table 3. Main features of the datasets used and methods applied in the collection of Jupyter Notebooks.

Dataset	Language	Type	Access	Method	Transformations
Biblioteca Digital Hispánica	Spanish	Text	OCR output text files	Topic modelling	Text preprocessing
Chronicling America	Spanish	Text	JSON API	Topic modelling	Text preprocessing
Lancaster University - Transcripción del Catálogo Monumental de España	Spanish	Text	Text files	N-gram language models	Text preprocessing

Zenodo facilitate an environment for the adoption of open licenses when publishing datasets.

Many institutions and aggregators (e.g., BDPI) include platforms that offer metadata and links, but in some cases, the OCR text is not available. Other institutions provide the original OCR output, but in a non-edited format, because editing is a difficult task that requires considerable resources. Crowdsourcing approaches could thus allow engaging with the public while improving the quality of the contents. Smaller-scale approaches based on a particular work or author are more affordable.

Generally, all benchmark subjects provide documentation about the production process. In some examples, aggregators consist of websites that provide content retrieved from several institutions. Locally-generated DOIs are used in some subjects while in others, the DOI is provided by publication platforms.

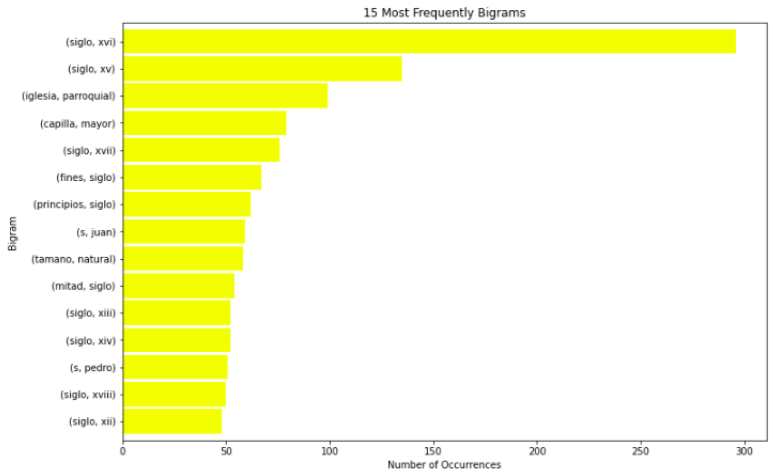

Figure 2. Overview of the most frequent bigrams for the Lancaster University dataset.

```
(0, '0.003*"guzmán" + 0.001*"conde" + 0.001*"sancho" + 0.001*"áfrica"')
(1, '0.000*"indios" + 0.000*"condestable" + 0.000*"pizarro" + 0.000*"alvaro"')
(2, '0.005*"nacional" + 0.004*"indios" + 0.002*"condestable" + 0.002*"¡ue"')
(3, '0.003*"indios" + 0.002*"franceses" + 0.002*"gonzalo" + 0.001*"cortes"')
(4, '0.003*"indios" + 0.002*"pizarro" + 0.002*"josé" + 0.002*"manuel"')
```

Figure 3. Topics and words obtained after applying the LDA model to the dataset from Biblioteca Digital Hispánica. Each topic and their corresponding words are related to a common theme (e.g., topic 3 is related to *franceses* and *cortes*).

```
(0, '0.001*"niño" + 0.001*"junio" + 0.001*"julio" + 0.001*"medina"')
(1, '0.001*"niño" + 0.001*"libro" + 0.001*"mayo" + 0.001*"medina"')
(2, '0.001*"rusia" + 0.001*"julio" + 0.001*"niño" + 0.001*"comisión"')
(3, '0.001*"independencia" + 0.001*"niño" + 0.001*"septiembre" + 0.001*"trabajadores"')
(4, '0.001*"diciembre" + 0.001*"gonzález" + 0.001*"julio" + 0.001*"compañía"')
```

Figure 4. Topics and words obtained after applying the LDA model to the journal *About Hispano América* from Chronicling America collection. Each topic and their corresponding words are related to a common theme (e.g., topic 3 is related to *independencia* and *trabajadores*)

According to Collections as Data, the datasets should include documentation and examples of use to demonstrate how they can be used for research. Documentation is usually provided, but there is still room for improvement regarding the inclusion of prototypes and examples of use as part of the datasets.

OCR quality is a crucial factor when reusing a dataset. Poor quality OCR requires preprocessing tasks (e.g., removing OCR errors based on non-existent words) and the latter can generate unexpected results. The texts provided by the subjects in the benchmark are different in terms of how they have been created and made available to the public (e.g., OCR output or manually reviewed). In general, the errors generated by OCR tools increase with the age of the documents. There are multiple reasons for this, such as the state of the print medium, the quality of the paper and the scan.[60,61] In this way, OCR software can help to improve quality regarding the use of machine learning-based neural networks, as well as the adoption of post-correction tools.[62,63]

In some cases, there is no option to retrieve the datasets by means of an API, hindering the reuse of the digital collections locked inside siloed repositories. In addition, institutions publish the information as PDF files instead of plain text files amenable to computational use. In this sense, tools such as the International Image Interoperability Framework (IIIF) provides an environment to facilitate the publication and reuse of the digital collections by means of APIs.

Datasets based on Linked Open Data principles provide rich metadata described using standard vocabularies. In these cases, the content is often provided as PDF files by means of URIs and using properties of the vocabularies such as Functional Requirements for Bibliographic Records (FRBR)[64] and Resource Description and Access (RDA).[65] As a result, users are required to understand the vocabularies. And this is sometimes a complex task for beginners. Documentation and examples can be useful in this case.

Regarding the language, and in the particular case of Spain, the contents provided by a digital collection

can be expressed in the co-official languages spoken in different geographical areas of the country, such as Catalan, Basque or Galician. Although this work focused on Spanish, the methodology to design the benchmark is flexible and can be adapted to language requirements, allowing the use of one or more languages.

Criteria regarding technical aspects can be improved by means of additional features, such as the use of an API key or the size of the collection. For example, some repositories require registration in order to be accessed and reused, such as the Rijksmuseum API.** In addition, the benchmarking can be improved through additional criteria adapted to assess datasets such as completeness, representativeness or timeliness.[45,66]

4 Conclusions

Cultural heritage institutions are starting to adopt Collections as Data in order to publish machine-actionable datasets that can be reused in innovative and creative ways.

The methodology described in Section 2.1 describes a series of steps to create a benchmark of machine-actionable datasets in the Spanish language that can be extended and adapted to other scenarios. In addition, recommendations and best practices are provided based on the results obtained for the benchmark. These examples encourage the adoption of Collections as Data within cultural heritage institutions. They also help to promote greater linguistic diversity regarding the texts provided.

The figures in Table 2 help select the machine-actionable collection that best fits a specific purpose. For instance, if the most relevant features for an institution are accuracy, using a permanent identifier and providing machine-readable text, the University of Lancaster dataset may be the best choice regarding reuse.

Future work could focus on further generalizing and automating the creation of the benchmark and the inclusion of additional features to compare datasets. In addition, the results of the benchmark and recommendations will be used to improve OCR tools and methods currently being used at the Biblioteca Virtual Miguel de Cervantes digital library to publish machine-actionable collections.

Acknowledgements

This research has been funded by the AETHER-UA (PID2020-112540RB-C43) Project from the Spanish Ministry of Science and Innovation.

References

1. Mahey M, Al-Abdulla A, Ames S et al. *Open a GLAM lab*. QU Press, 2019. ISBN 978-9927-139-07-9.
2. Library of Congress. Digital Scholarship at the Library of Congress: A Research Guide. URL https://guides.loc.gov/digital-scholarship/introduction.
3. Thomas Padilla. Responsible Operations: Data Science, Machine Learning, and AI in Libraries, 2019. DOI:https://doi.org/10.25333/xk7z-9g97. URL https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html. [Online; accessed 26-June-2020].
4. Silvia E Gutiérrez De la Torre and Miguel D Cuadros-Sánchez. Digital Resources: The Digital Library of Ibero-American Heritage, 2020. URL https://oxfordre.com/latinamericanhistory/view/10.1093/acrefore/9780199366439.001.0001/acrefore-9780199366439-e-798.
5. Mahey M, Al-Abdulla A, Ames S et al. *Open a GLAM Lab*. Alicante : Biblioteca Virtual Miguel de Cervantes, 2021. URL http://www.cervantesvirtual.com/nd/ark:/59851/bmc1066249.
6. Unlocking the Colonial Archive. Harnessing Artificial Intelligence for Indigenous and Spanish American Collections, 2021. URL https://unlockingarchives.com/research/.
7. González Soriano JM. Mnemosyne: A digital library of the other silver age (origins, contents, perspectives). *Signa: Revista de la Asociación Española de Semiótica* 2021; 30: 31–58. DOI:10.5944/signa.vol30.2021.29297. URL http://revistas.uned.es/index.php/signa/article/view/29297.
8. Sim SE, Easterbrook SM and Holt RC. Using Benchmarking to Advance Research: A Challenge to Software Engineering. In *Proceedings of the 25th International Conference on Software Engineering, May 3-10, 2003, Portland, Oregon, USA*. pp. 74–83. DOI:10.1109/ICSE.2003.1201189. URL https://doi.org/10.1109/ICSE.2003.1201189.
9. Spahiu B, Maurino A and Meusel R. Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned. *Semantic Web* 2019; 10(2): 329–348. DOI:10.3233/SW-180323. URL https://doi.org/10.3233/SW-180323.
10. European Commission. Cultural Heritage: Digitisation, online accessibility and digital preservation, 2018. URL https://ec.europa.eu/newsroom/

**https://data.rijksmuseum.nl/

dae/document.cfm?doc_id=60045. [Online; accessed 26-June-2020].
11. Europeana. Issue 16: Newspapers. URL https://pro.europeana.eu/page/issue-16-newspapers.
12. Lorang E, Soh LK, Liu Y et al. Digital libraries, intelligent data analytics, and augmented description: a demonstration project, 2020. URL https://labs.loc.gov/static/labs/work/experiments/final-report-revised_june-2020.pdf.
13. 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.
14. Harris G, Potter A, Zwaard K et al. Digital Scholarship at the Library of Congress, 2020. URL https://labs.loc.gov/static/labs/work/reports/DHWorkingGroupPaper-v1.0.pdf. [Online; accessed 26-June-2020].
15. Tasovac T, Chambers S and Tóth-Czifra E. Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper, 2020. URL https://hal.archives-ouvertes.fr/hal-02961317.
16. Ames S and Lewis S. Disrupting the library: Digital scholarship and Big Data at the National Library of Scotland. *Big Data & Society* 2020; 7(2): 2053951720970576. DOI: 10.1177/2053951720970576. URL https://doi.org/10.1177/2053951720970576. https://doi.org/10.1177/2053951720970576.
17. Candela G, Sáez MD, Esteban ME et al. Reusing digital collections from GLAM institutions. *Journal of Information Science* 0; 0(0): 0165551520950246. DOI:10.1177/0165551520950246. URL https://doi.org/10.1177/0165551520950246. https://doi.org/10.1177/0165551520950246.
18. Davids, André and Gabriels, Nele. Data-level access to Belgian historical censuses, 2020. URL https://enrichingheritage.wordpress.com/author/nelegabrielsoutlookcom/.
19. Ziku, Mariana and Gabriels, Nele. Opening up a little more: a minimal-computing approach for developing Git and machine-actionable GLAM open data, 2020. URL https://enrichingheritage.wordpress.com/2020/05/01/git-and-machine-actionable-data-pilot/.
20. Gijsbers P, LeDell E, Poirier S et al. An Open Source AutoML Benchmark. *arXiv preprint arXiv:190700909 [csLG]* 2019; URL https://arxiv.org/abs/1907.00909. Accepted at AutoML Workshop at ICML 2019.
21. Library of Congress. Selected Datasets: A New Library of Congress Collection, 2020. URL https://blogs.loc.gov/thesignal/2020/06/selected-datasets-a-new-library-of-congress-collection/. [Online; accessed 26-June-2020].
22. Library of Congress. Chronicling America. URL https://chroniclingamerica.loc.gov/about/.
23. Liceras-Garrido, Raquel and Comino, Alba and Murrieta-Flores, Patricia. Transcripción del Catálogo Monumental de España: Provincia de Ávila por Manuel Gómez Moreno (1900-1901), 2020. DOI: 10.6084/m9.figshare.12006318.v1.
24. Liceras-Garrido, Raquel and Comino, Alba and Murrieta-Flores, Patricia. Transcripción del Catálogo Monumental de la Provincia de Soria por Juan Cabré (1916-1917), 2020. DOI: 10.6084/m9.figshare.12006273.v1.
25. Liceras-Garrido, Raquel and Comino, Alba and Murrieta-Flores, Patricia. Transcripción del catálogo monumental y artístico de la provincia de burgos por narciso sentenach (1925), 2020. DOI: 10.6084/m9.figshare.12006327.v1.
26. British Library. A collection of datasets released by the British Library. URL https://data.bl.uk/.
27. Ministry of Culture. Mexicana, 2017. URL https://mexicana.cultura.gob.mx/en/repositorio/acerca.
28. National Library of Scotland. Data Foundry. Data collections from the National Library of Scotland. URL https://data.nls.uk/.
29. Austrian National Library. Data Sets. View, use and reuse the digital data sets of the ONB Labs. URL https://data.nls.uk/.
30. KB Labs. Datasets. URL https://lab.kb.nl/datasets.
31. World Wide Web Consortium. SPARQL 1.1 Query Language, 2013. URL https://www.w3.org/TR/sparql11-query/.
32. 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.
33. 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.
34. British Library. Basic RDF/XML. "http://www.bl.uk/bibliographic/datafree.html#basicrdfxml", 2014. [Online; accessed 26-June-2020].
35. Freire N, Voorburg R, Cornelissen R et al. Aggregation of Linked Data in the Cultural Heritage Domain: A Case

Study in the Europeana Network. *Inf* 2019; 10(8): 252. DOI:10.3390/info10080252. URL https://doi.org/10.3390/info10080252.

36. Beals, Melodee and Bell, Emily. The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges, 2020. DOI:{10.6084/m9.figshare.11560059.v2}.
37. Rueda L, Fenner M and Cruse P. Datacite: Lessons learned on persistent identifiers for research data. *Int J Digit Curation* 2016; 11(2): 39–47. DOI:10.2218/ijdc.v11i2.421. URL https://doi.org/10.2218/ijdc.v11i2.421.
38. 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.
39. Project Jupyter. URL https://jupyter.org/.
40. Sherratt T. Glam-workbench/getting-started, 2019. DOI:10.5281/zenodo.3549636. URL https://doi.org/10.5281/zenodo.3549636.
41. Library of Congress. LC Maps for Robots, 2020. URL https://blogs.loc.gov/thesignal/2020/05/lc-maps-for-robots/.
42. Shen G and Liu G. The selection of benchmarking partners for value management: An analytic approach. *International Journal of Construction Management* 2014; 7. DOI:10.1080/15623599.2007.10773099.
43. Heckman SS and Williams L. On establishing a benchmark for evaluating static analysis alert prioritization and classification techniques. In *Proceedings of the Second International Symposium on Empirical Software Engineering and Measurement, ESEM 2008, October 9-10, 2008, Kaiserslautern, Germany*. pp. 41–50. DOI:10.1145/1414004.1414013. URL https://doi.org/10.1145/1414004.1414013.
44. Sarkar A, Yang Y and Vihinen M. Variation benchmark datasets: update, criteria, quality and applications. *Database* 2020; 2020. DOI:10.1093/database/baz117. URL https://doi.org/10.1093/database/baz117. Baz117, https://academic.oup.com/database/article-pdf/doi/10.1093/database/baz117/32322837/baz117.pdf.
45. Candela G, Escobar P, Carrasco RC et al. Evaluating the quality of linked open data in digital libraries. *Journal of Information Science* 0; 0(0): 0165551520930951. DOI:10.1177/0165551520930951. URL https://doi.org/10.1177/0165551520930951. https://doi.org/10.1177/0165551520930951.
46. Miksa T, Simms S, Mietchen D et al. Ten principles for machine-actionable data management plans. *PLOS Computational Biology* 2019; 15(3): 1–15. DOI:10.1371/journal.pcbi.1006750. URL https://doi.org/10.1371/journal.pcbi.1006750.
47. Wang RY and Strong DM. Beyond accuracy: What data quality means to data consumers. *J of Management Information Systems* 1996; 12(4): 5–33. URL http://www.jmis-web.org/articles/1002.
48. Library of Congress. OCR Data. URL https://chroniclingamerica.loc.gov/ocr/.
49. Library of Congress. By the People. URL https://crowd.loc.gov/.
50. Biblioteca Nacional de España. Comunidad BNE. URL https://comunidad.bne.es/.
51. British Library. Libcrowds. URL https://www.libcrowds.com/.
52. Europeana. Europeana Transcribe. URL https://europeana.transcribathon.eu.
53. Lee BCG, Mears J, Jakeway E et al. The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america, 2020. 2005.01583.
54. Thomas Padilla. On a Collections as Data Imperative. URL https://labs.loc.gov/static/labs/work/reports/tpadilla_OnaCollectionsasDataImperative_final.pdf.
55. Wilkinson MD, Dumontier M, Aalbersberg IJ et al. The fair guiding principles for scientific data management and stewardship. *Scientific data* 2016; 3.
56. Snydman S, Sanderson R and Cramer T. The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images, 2015. URL https://stacks.stanford.edu/file/druid:df650pk4327/2015ARCHIVING_IIIF.pdf.
57. Weiss A and James R. An examination of massive digital libraries' coverage of spanish language materials: Issues of multi-lingual accessibility in a decentralized, mass-digitized world. In *2013 International Conference on Culture and Computing*. pp. 10–14. DOI:10.1109/CultureComputing.2013.10.
58. Research Libraries UK. The role of academic and research libraries as active participants and leaders in the production of scholarly research, 2021. URL https://www.rluk.ac.uk/wp-content/uploads/2021/07/RLUK-Scoping-Study-Report.pdf.
59. Binder. URL https://mybinder.org/.
60. Jarlbrink J and Snickars P. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *J Documentation* 2017; 73(6): 1228–1243. DOI:10.1108/JD-09-2016-0106. URL https://doi.org/10.1108/JD-09-2016-0106.
61. van Strien D, Beelen K, Ardanuy MC et al. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*. pp. 484–496. DOI:10.5220/0009169004840496. URL https://doi.org/10.5220/0009169004840496.
62. Poncelas A, Aboomar M, Buts J et al. A tool for facilitating ocr postediting in historical documents, 2020. 2004.11471.

63. Colutto S, Kahle P, Hackl G et al. Transkribus. A platform for automated text recognition and searching of historical documents. In *15th International Conference on eScience, eScience 2019, San Diego, CA, USA, September 24-27, 2019*. pp. 463–466. DOI:10.1109/eScience.2019.00060. URL https://doi.org/10.1109/eScience.2019.00060.
64. IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional Requirements for Bibliographic Records, 1998. URL https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.
65. RDA Steering Committee. RDA Registry, 2014. URL http://www.rdaregistry.info/.
66. Temnikova I, Baumgartner Jr WA, Hailu ND et al. Sublanguage corpus analysis toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 1714–1718. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/675_Paper.pdf.