Received: 30 March 2021 | Revised: 15 September 2021 | Accepted: 23 September 2021

DOI: 10.1002/asi.24583

RESEARCH ARTICLE

JASIST WILEY

Discovering emerging topics in textual corpora of galleries, libraries, archives, and museums institutions

Gustavo Candela | **Rafael C. Carrasco**

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain

Correspondence
Gustavo Candela, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Carretera Sant Vicent s/n, 03690 Sant Vicent del Raspeig, Alicante, Spain.
Email: gcandela@ua.es

Abstract

For some decades now, galleries, libraries, archives, and museums (GLAM) institutions have provided access to information resources in digital format. Although some datasets are openly available, they are often not used to their full potential. Recently, approaches such as the so-called *Labs* within GLAM institutions promote the reuse of digital collections in innovative and inspiring ways. In this article, we explore a straightforward computational procedure to identify emerging topics in periodical materials such as newspapers, bibliographies, and journals. The method is illustrated in three use cases based on public digital collections. This type of tools are expected to promote further usage by researchers of the digital collections.

1 | INTRODUCTION

Over the past decades, national libraries, universities, archives, and museums have been made available in digital format periodic publications such as journals, magazines, and newspapers. These collections open new opportunities for historians, journalists and researchers, and also for the general public.

In parallel, standards have been promoted to increase interoperability and facilitate the access to large digital collections. The use of open licenses, for example, facilitates the reuse of the collections in new contexts including computationally driven research.

GLAM institutions have published digital collections using traditional methods based on websites and Application Programming Interfaces (APIs) (Europeana, 2020b). In some cases, advanced technologies such as the semantic web and Linked Open Data (LOD) are employed (BnF Data, 2014; Romero et al., 2018). Often, experimental methods are based on collaborative edition such as transcription, tagging, and description (Biblioteca Nacional de España, 2019; British Library, 2015; Library of Congress, 2018). Some innovative methods, such as *collections as data*, aim to the publication of digital collections as datasets amenable to computational use (Padilla et al., 2019). Numerous papers and best practices have highlighted the importance of using open data to ensure reproducibility in computational research (Baillieul et al., 2017, 2018; Rule et al., 2018).

Text mining techniques such as information extraction, categorization, clustering, natural language processing (NLP), and topic modeling have become very popular in the research community to search for relationships among text documents (Buenaño-Fernández et al., 2020). For such purposes, visualization methods can facilitate the analysis of text by providing interactive charts and dashboards. When periodic literature is available, it is possible to explore trends and temporal variations in the content. For example, we may be interested in the early detection of *emerging topics*, that is, novelties in the content. Note that this is different from the identification of *trending topics* which are not necessarily new and can be in use for a longer period. For example, emerging subjects in scientific papers can help funding

bodies to identify promising areas of research where additional investment may be recommended.

The purpose of this paper is to illustrate how a computational method can easily identify emerging topics in corpora provided by GLAM institutions when the content has a publication time mark, as it is the case of periodical publications. Books could be also in principle amenable to this type of analysis, using for example, the publication date, although the amount of items should be large enough and, in case there is a wide coverage of topics, results can be more difficult to compare and interpret. We will therefore examine use cases providing reproducible notebooks on the journals and newspapers provided by British Library, Biblioteca Virtual Miguel de Cervantes (BVMC) and *dblp computer science bibliography*.

Our approach is based on a simple lexical analysis of the content, with no attempt to semantically cluster related concepts (for example, migration and refugee). It is, however, sometimes difficult to conclude if a new term should be considered as a new category or as the continuation of a pre-existing concept (for example, if *deep learning* and *artificial neural networks* should be categorized together).

The main contributions of this paper are the following: (a) A computational method to identify emerging topics in periodic publications; (b) a collection of notebooks based on newspapers and journals published by relevant GLAM institutions; and (c) a practical example of reuse of journals published as LOD.

The paper is organized as follows: After a brief description of the state of the art in Section 2, Section 3 introduces the framework to identify emerging topics in large corpus. Section 4 evaluates the framework by means of several relevant GLAM institutions and shows the results of their application. The paper concludes with an outline of the framework and future work.

2 | RELATED WORK

A number of GLAM institutions distribute digital collections which include periodical publications such as newspapers and journals. For instance, the Newspaper Navigator project of the Library of Congress allows users to browse over 1.5 million images through the Chronicling America[*] collection (Library of Congress, 2020). The British Library's EThOS service comprises metadata descriptions of hundreds of thousands of PhD theses awarded by UK Higher Education institutions (Heather Rosie, 2021). Europeana newspaper explores the headlines, articles, advertisements, and opinion pieces from European newspapers from 20 countries, dating from 1618 to 1996 (Europeana, 2018, 2020a). The Atlas of Digitized Newspapers and Metadata is an open-access guide to a selection of newspaper databases around the world (Beals & Bell, 2020). The Early Journal Content (EJC) on JSTOR—a digital library of academic content in many formats and disciplines—includes public domain journal articles published in the United States before 1923 and articles published in other countries before 1870 including metadata, n-grams, and full text for text mining purposes (JSTOR, 2017). A sample of GLAM institutions providing access to periodical datasets can be found in the appendix (Table A1).

When digital collections are published as LOD by using standard vocabularies to describe their content, the collections can be enriched with external repositories—such as Wikidata and GeoNames—to include contextual information (Romero et al., 2019). In particular, open publication allows researchers to create Jupyter Notebooks,[†] a type of publication which has become very popular in the research community and provides a web environment for transparent, collaborative, reproducible, and reusable data analyses. A notebook integrates detailed workflows, narrative text, and visualization of results. For instance, the GLAM Workbench (Sherratt, 2019) provides notebooks based on digital collections showing a collection of tools, tutorials, and examples. The GLAM Jupyter Notebooks explore the creation of machine actionable collections by means of datasets provided by several relevant GLAM institutions (Candela et al., 2020). The Library of Congress has published a collection of Jupyter Notebooks to query, download, and visualize cartographic material (Weinryb-Grohsgal, 2020). Recently, a new collection of Jupyter Notebooks has been published by the National Library of Scotland (National Library of Scotland, 2020). Additional Jupyter Notebooks are based on digitized newspaper data from the National Library of Estonia (Tinits, 2020).

So far, tools based on newspapers allow users to browse curated lists, and search the full text and metadata in a date range. In addition, NLP techniques, such as Named Entity Recognition (NER), have been applied to unstructured text in order to identify and classify named entities (places, persons, and locations). (Neudecker, 2016) Other examples are built upon machine learning techniques to detect shapes and objects, and search for similar images (Wevers & Lonij, 2017). Swiss and Luxembourgian newspapers have been explored in order to segment newspaper images and to classify detected segments according to a newspaper typology (Barman et al., 2020). A computational analysis of historical Hebrew newspapers has been performed to identify trends in the discourse (Segal et al., 2019).

Lately, words embeddings have attracted strong interest from the research community to perform tasks, such

as Named Entity Recognition, part-of-speech tagging, and text classification (Bakarov, 2018; Sung et al., 2020). Danish newspapers have been reused to identify how the use of language has evolved since the 18th century by means of experimental visualizations and word embeddings (KB Labs, 2016, 2018). Other approaches are based on more advanced models such as fastText and Bidirectional Encoder Representations from Transformers (BERT) (Hammou et al., 2020; Polignano et al., 2019).

The text mining technique topic modeling has become a popular procedure for clustering documents into semantic groups. Interactive visualization tools based on raw text data have been introduced by organizations such as DARIAH-DE that supports research in the humanities and cultural sciences with digital methods and procedures (DARIAH-DE, 2020).

There are existing web-based tools which assist in the discovery of trends in massive textual corpora: For example, Lansdall-Welfare and Cristianini employ n-grams and Zipf's law to identify the topics with the highest relevance. Also burst detection in time series has been addressed before to identify topics that grow in intensity for a period of time (Kleinberg, 2003). Detection of emerging topics has however a slightly different objective, as novelties may pass undetected before they become a trend or burst and, sometimes, early detection of such topics is requested. Past work implements term-selection or clustering techniques to select relevant terms. (Chandrakala et al., 2019) These methods employ standard techniques, such as stop-word removal and computation of TF-IDF frequencies to select terms before analyzing their temporal evolution. In this paper, we show how a straightforward and reproducible method, weighting the terms according to their novelty, provides good results without the need of a term-selection or training step.

Although many studies are focused on newspapers and journals to extract information from text corpora, to our best knowledge, the definition of a framework for the identification of emerging topics in datasets provided by GLAM institutions have not been addressed. In this sense, the combination of the definition of a framework and the publication of a collection of notebooks enabling the research reproducibility can help to increase the visibility of the datasets encouraging researchers to reuse them.

3 | A SIMPLE METHOD TO IDENTIFY EMERGING TOPICS

The procedure depicted in Figure 1 works in four steps:

1. Identification,
2. Access and retrieval,
3. Machine processing, and
4. Human revision and visualization.

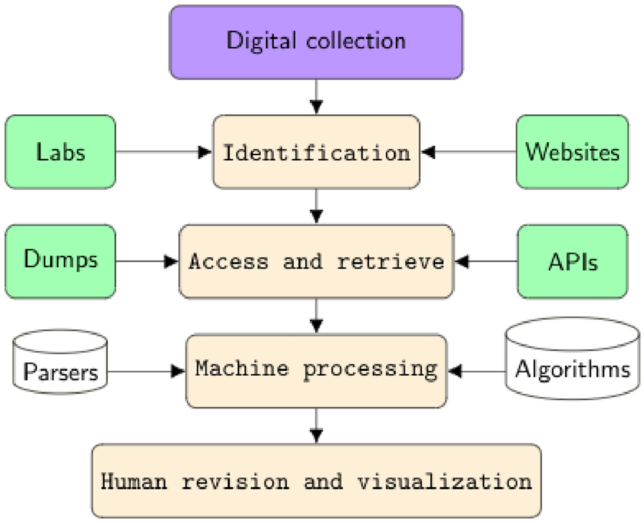

FIGURE 1 Framework employed to identify emerging topics in text corpora provided by GLAM institutions

Depending on the characteristics of the dataset such as license, language, format, and accessing method, the application of the framework may slightly vary for each institution. Each step may require different adjustments in order to adapt the framework to other digital collections published by GLAM institutions.

3.1 | Identification

The identification step consists of the selection of existing resources which support the identification of emerging trends. It requires the application of a set of evaluation criteria including, among others: (a) License to identify if the resources are available for open access or using copyright restrictions; (b) accessibility in order to measure the extent to which data including metadata and full text are available; (c) completeness to measure the extent to which data are of sufficient depth for the task (e.g., dates of publication are distributed over a long enough period); and (d) ease of use to assess whether the collections are available by means of simple APIs or dump files.

Moreover, the type of the resources —including newspapers and journals—, and its content and metadata are relevant aspects to take into consideration. In this sense, coverage, provenance, and transparency are crucial aspects to obtain optimum results. GLAM institutions have made a considerable effort in providing

computational access to the digital collections and documentation about them as well as making them available under open licenses to promote innovation and provide access to the audience including researchers and general users.

Information retrieval systems built upon websites allow users to explore the digital collections provided by GLAM institutions. Particular sections may be devoted to specific content such as thesis, maps, articles, journals, and newspapers. The content of journals and newspapers resources can be provided as PDF files and HTML, but also as text files that are more easily processed by computers. GLAM Labs have started to publish datasets ready for reuse in several formats, demonstrating examples of use and reuse (Mahey et al., 2019). In some cases, open-access data repositories such as Zenodo and *figshare* allow researchers to deposit datasets and research results.

Although many institutions publish their digital collections under open licenses, it is not infrequent to find images and resources with imprecise copyright statements (Schlosser, 2009). Some records may also use outdated models and vocabularies (Pitt Rivers Museum, 2020).

3.2 | Access and retrieve

Basic approaches for the access and retrieval of content focus on the extraction of text corpora and metadata by accessing the websites using a web crawler or a script. Advanced methods rely, however, on the use of APIs enabling the download of manageable slices of the data in several formats such as the result of a search, or the works of an author or a subject. Several models to build APIs are available including, among others, Open Archives Initiative Protocol for Metadata Harvesting(Open Archives Initiative, 2002) (OAI-PMH), International Image Interoperability Framework (IIIF Consortium, 2014) (IIIF), and SPARQL (World Wide Web Consortium, 2013). IIIF has become very popular among the community and many GLAM institutions are adopting it for the publication of their digital collections including images. SPARQL is a powerful language to query a repository, however, its use requires some knowledge of semantic web technologies.

Simpler approaches are based on the publication of data dumps as large files which often require a preprocessing step before their usage. For example, Chronicling America provides access to the complete set of images, texts, and OCR coordinates as compressed archive files.[*] Also the Bibliothèque Nationale du Luxembourg provides several newspapers datasets of increasing size (Bibliothèque nationale du Luxembourg, 2019).

3.3 | Machine processing

Uses a corpus of textual data as input and identifies.

The procedure to identify emerging topics uses a corpus of textual data as input and identifies incipient terms. We have tested that a straightforward approach may suffice for newspaper and journal collections and, in general, for publications issued at regular intervals.

Given a collection containing N documents $D = (d_1, d_2, ..., d_N)$ with publication dates $T = (t_1, t_2, ..., t_N)$, the *term frequency* matrix

$$\mathrm{TF} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1N} \\ c_{21} & c_{22} & \dots & c_{2N} \\ c_{M1} & c_{M2} & \dots & c_{MN} \end{bmatrix} \quad (1)$$

specifies the number c_{mn} of occurrences of term w_m in document d_n. The global frequency g_m in the collection of term t_m can be obtained by simply adding the content of every row, that is, $g_m = \sum_{n=1}^{N} c_{mn}$. If the addition is replaced by the count of nonzero values, one obtains the *document frequency* vector $\mathrm{DF} = (f_1, f_2, ..f_M)$, that is, the number f_m of documents containing w_m at least once.

The average date of the occurrences of term w_m is simply:

$$\mathfrak{T} = (\mathrm{TF} \times T)/\mathrm{DF}, \quad (2)$$

where the quotient represents element-wise division. Terms which are novel in the collection will show the highest value of $\mathfrak{T}$ and, therefore, ranking terms according to this value should reveal which ones are the most recent. User-defined lower threshold for the term and document frequencies will remove cases where the available information is not enough to draw significant conclusions—by default, we have selected $c_m \geq 10$ and $f_m \geq 3$. Stopwords and terms with a very high document frequency (by default, above $0.1N$) have been also removed. The pseudo-code of the procedure is shown below.

Required input: A list of documents.

$D = (d_1, ..., d_N)$ and a list of dates

$$T = (t_1, ..., t_N).$$

Optional input: thresholds $f_{\min}$, $g_{\min}$, $g_{\max}$, and a range of dates $(t_{\min}, t_{\max})$.

Output: A sorted list of emerging topics.

Procedure:

1. Remove d_n and t_n if $t_n > t_{\max}$ or $t_n < t_{\min}$ and reindex D and T accordingly.
2. Compute term frequency matrix $\mathrm{TF} = (c_{mn})$
3. Get document frequency vector $\mathrm{DF} = (f_m)$
4. Add TF rows for global frequency $\mathrm{G} = (g_m)$
5. Remove t_n from TF and DF (and reindex) if $f_m < f_{\min}$, $g_m < g_{\min}$, or $g_m > g_{\max}$
6. Compute $\mathfrak{T} = \mathrm{TF} \times T/\mathrm{DF}$
7. Return list of terms w_m sorted according to $\mathfrak{T}$

Terms in this approach can be single words or n-grams: An n-gram is a sequence of n consecutive words such as" deep learning." In our experiments, terms of length $n = 1$ (words) and $n = 2$ (bigrams) were extracted from the documents, although storing terms with higher values of n is also feasible.

3.4 | Human revision and visualization

Human revision of the results can improve the accuracy of this procedure, for example, discarding cases where an apparent novelty of a term is indeed a change of nomenclature.

Furthermore, the results can be transformed into interactive charts or graphical reports enabling researchers to understand, explain, and collect patterns from the data. Visualizations tools such as interactive charts and scorecards —visual interfaces to easily find, analyze, and explore information— can lead to new discoveries which, in turn, can foster the exploitation of legacy material using new techniques and methods.

4 | EVALUATION OF THE FRAMEWORK

This section introduces three use cases to assess the framework proposed in Section 3. We have selected the datasets according to the following criteria:

- They are available for open access,
- The content is provided as full text,
- Dates of publication are distributed over a long enough period, and
- Simple access to the contents is provided through an API or in the form of bulk data files.

Jupyter Notebooks have been used to combine documentation, data, charts, and code: The project is available in GitHub[*] as a collection of interactive notebooks executable in the cloud-based platform Binder.[†] The notebook collection has been assigned a Digital Object Identifier (DOI) with the data archiving tool Zenodo (Candela et al., 2021).

The collection of Jupyter Notebooks is based on Python since it is a popular language with a low entry barrier (Koerner et al., 2020; Raschka et al., 2020). Binder is a well-known platform in the research community and provides an easy-to-use cloud environment to execute and reproduce the results.

The notebooks employ some open-source tools for handling data in Python such as NumPy,[‡] Python Data Analysis Library[§] SciPy[¶] and Natural Language Toolkit[‖] Additional packages are used to create HTTP requests, and to retrieve and handle the results in a variety of formats such as JSON and CSV.

Regarding the optimization, we have identified the optimum results by setting the DF and TF thresholds to 3 and 20, respectively. Results are shown in Tables 3–10. The first column corresponds to the term identified as an emerging topic and the second column shows the average date in which the term appears in the documents. It is relevant to notice that the results may include several terms in the same year (e.g., the values 1994,4 and 1994,25 shown in Table 6).

4.1 | Doxa: A journal published as LOD at the BVMC

The catalog of the BVMC contains about 285,000 records and it was published as LOD in 2015. The repository was built using the Resource, Description, and Access (RDA) vocabulary to describe the items in the catalog (RDA Steering Committee and ALA Digital Reference, 2015). The LOD repository contains several types of materials such as videos, audios, images, books, journals, and maps, including metadata about the authors, dates, and subjects. The RDA vocabulary contains classes and properties to describe the resources that are linked by means of typed relationships. For instance, the whole-part relation is an association between a resource representing a part and a resource representing its corresponding whole that is used to describe journals. Figure 2 shows how the manifestations representing journals, volumes, and articles are linked by means of the property `wholePartManifestationRelationship` in the namespace rdam.[**]

Doxa. Cuadernos de Filosofía del Derecho is a periodical publication[††] issued every year since 1984 to promote the interaction between philosophers of law from Latin America and Latin Europe. The information regarding this publication has been included in the BVMC and it has been published as LOD in the repository, including metadata and text, being accessible by means of the public SPARQL endpoint.

In the examples below, the SPARQL API from data.cervantesvirtual.com/sparql was employed to retrieve the articles of the journal *Doxa—Filosofía del derecho*. Figure 3 shows the SPARQL query to retrieve the articles, including the PDF file containing the full text of each item. An overview of the results retrieved are shown in Table 1.

Although the LOD repository does not contain the full text, it can be easily retrieved through the URLs which are stored as Item entities in the RDA vocabulary using the property `rdai:identifierForTheItem`. The Tika library[‡‡] was used to extract the content of PDF files.

The metadata and full text retrieved are loaded as a Pandas DataFrame before the extraction of emerging topics takes place. In order to provide a snapshot of the application of the method, the analysis has been performed for three periods: 1984–1989 (164 issues, results shown in Table 1), the period 1990–1995 (224 issues), and the period 2010–2018 (261 issues).

The output was assessed by content managers of the library in charge of the curation of the journal. Identifying the most relevant topics such as concepts, people, and locations, as well as particular terms that were highlighted as emerging topics in specific years.

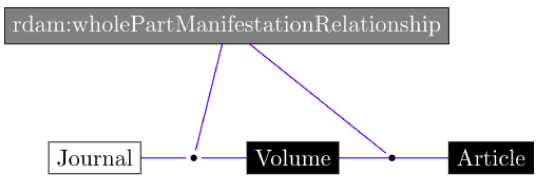

FIGURE 2 Relationships between the resources described as journal articles in RDA. Manifestations representing journals, volumes, and articles are linked by means of the property `rdam:wholePartManifestationRelationship`

In particular, the top 100 potential emerging topics obtained as a result per each period were manually assessed with an average success rate of about 50%. Tables 2–4 show the top 10 emerging topics that are automatically obtained by the framework for each period, while Tables 5–7 show the top 10 emerging topics obtained after the assessment by the content curators. In general, the overall assessment by the content managers was positive, and they provided useful feedback and comments about the emerging topics obtained for each period.

Additional tests were performed to compare the performance for different lengths of the time period. For example, when the intervals 2010–2019 and 2015–2019 were employed to obtain 20 potential emerging topics, the manual assessment found a success rate of 80% and 70%, respectively. Longer periods did not, however, lead to significant improvements.

4.2 | UK doctoral thesis metadata from EThOS: The British library

The British Library provides the bibliographic metadata for all UK doctoral theses listed in EThOS, the UK's national thesis service (Heather Rosie, 2021). The data in this collection comprise PhDs awarded by the UK Higher Education institutions described since 1787.

The metadata is provided as a CSV file including several fields such as the title of the thesis, the name of the author, the abstract (long abstracts have been truncated), the year or the institution responsible for the thesis.

The CSV file containing the metadata is loaded as a Pandas DataFrame before the extraction of emerging topics takes place. The analysis has been performed for the period 2015–2021 including 112,776 PhDs. The top first 10 potential emerging topics obtained as a result are shown in

```
PREFIX rdam: <http://rdaregistry.info/Elements/m/>
PREFIX rda:<http://www.rdaregistry.info/>
PREFIX rdai:<http://rdaregistry.info/Elements/i/>

SELECT ?num ?numTitle ?article ?articleTitle ?date ?noteEdition
?carrierCharacteristic ?pdf
WHERE { VALUES ?date {
    <http://data.cervantesvirtual.com/date/1984>
    <http://data.cervantesvirtual.com/date/1985>
    <http://data.cervantesvirtual.com/date/1986>
    <http://data.cervantesvirtual.com/date/1987>
    <http://data.cervantesvirtual.com/date/1988>
    <http://data.cervantesvirtual.com/date/1989>
}
?num rdam:wholePartManifestationRelationship
<http://data.cervantesvirtual.com/manifestation/237680> .
?num rdam:title ?numTitle .
?num rdam:dateOfPublication ?date .
?article rdam:wholePartManifestationRelationship ?num .
?article rdam:title ?articleTitle .
?article rdam:exemplarOfManifestation ?item .
?article rdam:noteOnEditionStatement ?noteEdition .
?item rdai:identifierForTheItem ?pdf .
?item rdai:itemSpecificCarrierCharacteristic ?carrierCharacteristic; }
```

FIGURE 3 A SPARQL query retrieving the articles included in the journal Doxa from the Biblioteca Virtual Miguel de Cervantes (BVMC) LOD repository. The instruction `VALUES` is used to specify the dates of publication of the articles

TABLE 1 Overview of the results retrieved in the SPARQL query in Figure 3

num	numTitle	Article	articleTitle	Date	noteEdition	carrierCharacteristic	pdf
.../manifestation/325787	Doxa. Cuadernos de Filosofía del derecho. Núm. 1, 1984	.../manifestation/678172	Presentación [Doxa, núm. 1 (1984)]	.../date/1984	Edición digital a partir de Doxa: Cuadernos de Filosofía del derecho, núm. 1 (1984), pp.7–9	Pdf	http://www.cervantesvirtual.com/descargaPdf/presentacion-doxa-num-1-1984/
.../manifestation/325787	Doxa. Cuadernos de Filosofía del derecho. Núm. 1, 1984	.../manifestation/678181	Aulis Aarnio (Helsinki) / traducción de Juan Ruiz Manero	.../date/1984	Edición digital a partir de Doxa: Cuadernos de Filosofía del derecho, núm. 1 (1984), pp.11–14	Pdf	http://www.cervantesvirtual.com/descargaPdf/aulis-aarnio-helsinki/
.../manifestation/325787	Doxa. Cuadernos de Filosofía del derecho. Núm. 1, 1984	.../manifestation/678191	Robert Alexy (Gottingan) / traducción de Ernesto Garzón Valdés	.../date/1984	Edición digital a partir de Doxa: Cuadernos de Filosofía del derecho, núm. 1 (1984), pp.15–17	Pdf	http://www.cervantesvirtual.com/descargaPdf/robert-alexy-gottingan/
.../manifestation/325787	Doxa. Cuadernos de Filosofía del derecho. Núm. 1, 1984	.../manifestation/678199	Joaquín Almoguera (Madrid)	.../date/1984	Edición digital a partir de Doxa: Cuadernos de Filosofía del derecho, núm. 1 (1984), pp.19–23	Pdf	http://www.cervantesvirtual.com/descargaPdf/joaquin-almoguera-madrid/

Note: Dots stand for the common prefix data.cervantesvirtual.com

TABLE 2 Top 10 emerging topics obtained as a result of the framework based on the journal Doxa (1984–1989)

Term	Av. date	DF	TF
Maximizador	1989	5	29
Sig	1989	7	25
Esperada	1989	5	36
Derrida	1989	3	49
Deliberativa	1989	3	22
Principio mayoría	1989	4	20
Postmoderna	1989	4	43
Postmodernidad	1989	7	31
Boaventura Sousa	1989	3	24
Utilidad esperada	1989	4	33

TABLE 3 Top 10 emerging topics obtained as a result of the framework based on the journal Doxa (1990–1995)

Term	Av. date	DF	TF
Neutralidad estatal	1995	3	23
RDA	1995	3	90
[a]DOXA	1995	22	22
Kymlicka	1994,4	5	39
Obligación política	1994,25	4	37
Reciprocidad	1994,23	13	25
Orden estatal	1994,2	5	30
Propiedad privada	1994,17	12	101
Comunitarios	1994,1	11	74
Geiger	1994	6	37

TABLE 4 Top 10 emerging topics obtained as a result of the framework based on the journal Doxa (2010–2018)

Term	Av. date	DF	TF
Paco	2017	12	64
Laporta Liborio	2017	17	51
HIERRO LAPORTA	2017	11	23
HIERRO Francisco	2017	8	24
Derecho solo	2016,9	20	30
Judicial activism	2016,83	6	21
Ideal imperio	2016,78	9	23
Inscribe	2016,73	15	21
Felipe González	2016,71	7	25
José Luis	2016,67	12	24

Table 8. Figures 4–7 show the document frequency according to the period and obtained as a result using as a source the metadata collection EThOS. The top

TABLE 5 Top 10 emerging topics obtained as a result of the assessment by the content curators based on the journal Doxa (1984–1989)

Term	Av. date	DF	TF
Maximizador	1989	5	29
Esperada	1989	5	36
Derrida	1989	3	49
Deliberativa	1989	3	22
Principio mayoría	1989	4	20
Postmoderna	1989	4	43
Postmodernidad	1989	7	31
Utilidad esperada	1989	4	33
Cooperar	1988,83	6	22
Cooperativo	1988,75	4	40

TABLE 6 Top 10 emerging topics obtained as a result of the assessment by the content curators based on the journal Doxa (1990–1995)

Term	Av. date	DF	TF
Neutralidad estatal	1995	3	23
Kymlicka	1994,4	5	39
Obligación política	1994,25	4	37
Orden estatal	1994,2	5	30
Propiedad privada	1994,17	12	101
Comunitarios	1994,1	11	74
Geiger	1994	6	37
Cultura política	1994	11	28
Liberalismo político	1994	6	47
Maximización Riqueza	1994	3	32

10 terms emerging topics obtained as a result correspond to PhDs subjects in Medicine and Health, History and Archaeology, Business and Administrative Studies, Biological Sciences, and Engineering and Technology.

4.3 | The DBLP computer science bibliography

The *DBLP computer science bibliography* provides open bibliographic information on major computer science journals and proceedings since 1993.[*] Inspired by the BibTeX format—a management tool for formatting lists of references— the project provides a dataset in XML format including the publication records. (Ley, 2009)

The metadata is provided as an XML file including an extensive list of publication records described with fields such

TABLE 7 Top 10 emerging topics obtained as a result of the assessment by the content curators based on the journal Doxa (2010–2018)

Term	Av. date	DF	TF
HIERRO LAPORTA	2017	11	23
Judicial activism	2016,83	6	21
Felipe González	2016,71	7	25
RUIZ-GIMÉNEZ	2016,57	7	48
LAPORTA HIERRO	2016,54	13	22
Ley visión	2016,5	26	32
Interamericana	2016,43	7	20
Valor autonomía	2016,33	9	21
Predecibilidad	2016,31	13	38
Postpositivista derecho	2016,2	9	22

TABLE 8 Top first 10 emerging topics obtained as a result based on the UK Doctoral Thesis Metadata from EThOS (2015–2021)

Term	Av. date	DF	TF
scRNA-seq	2019,55	12	20
Indus	2019,33	9	21
Section B	2019,31	26	27
Policy uncertainty	2019,11	10	20
Seagrass	2019,1	10	26
Learning analytics	2019,08	12	36
RBPs	2019,08	13	33
Mid-air	2019,07	14	30
Dure	2019,04	23	26
CER	2019	12	20

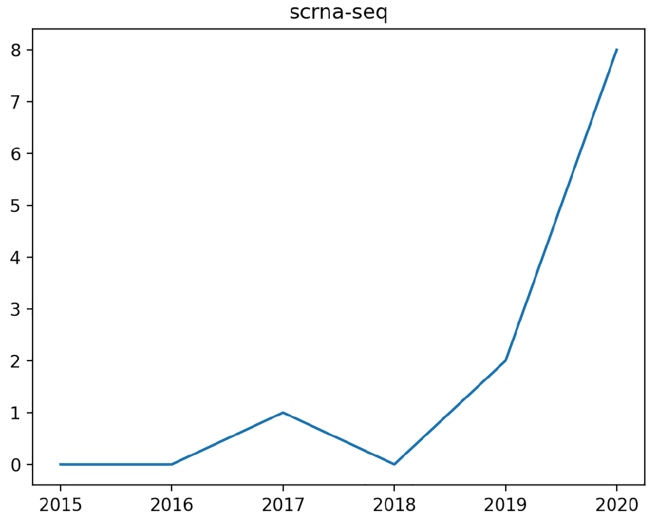

FIGURE 4 DF of the term *scRNA-seq* in EThOS

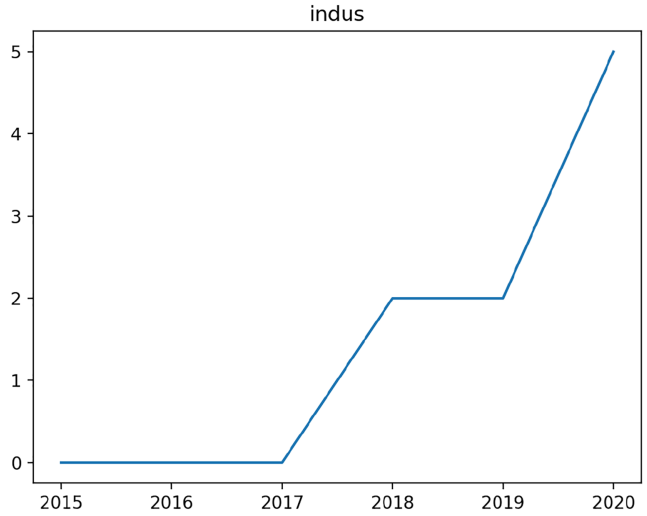

FIGURE 5 DF of the term *Indus* in EThOS

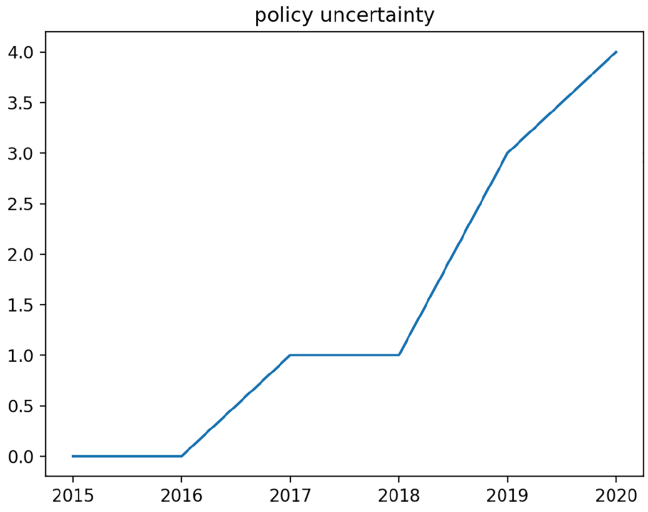

FIGURE 6 DF of the term *policy uncertainty* in EThOS

as the type of document (e.g., article, PhD thesis or conference paper), the title, the name of the authors, the year, the URL, or the journal in which the article was published (Figure 8).

First, a CSV file including only the title, year, and type of content (e.g., article) is extracted from the original XML file which is publicly available as a downloadable dataset (Carrasco & Candela, 2021).

The CSV file containing the metadata is loaded as a Pandas DataFrame before the extraction of emerging topics takes place. The analysis has been performed for two periods: 2000–2010 (570,116 articles) and 2010–2021 (1,548,362 articles). The top first 10 potential emerging topics obtained as a result per each period are shown in Tables 9 and 10. Figures 8–11 show the document frequency according to the period

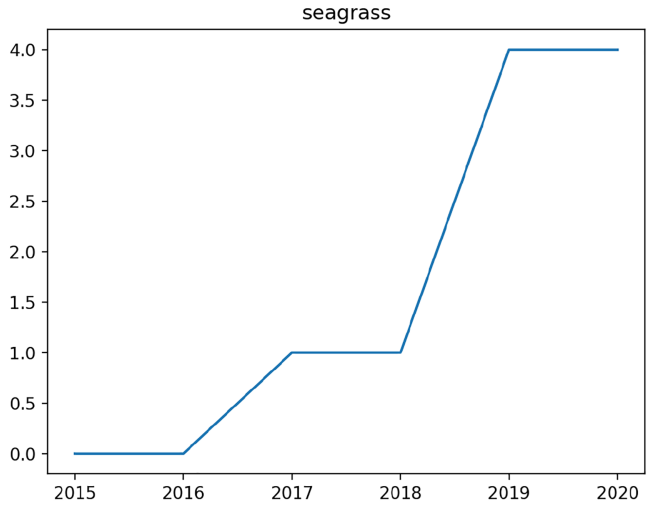

FIGURE 7 DF of the term *seagrass* in EThOS

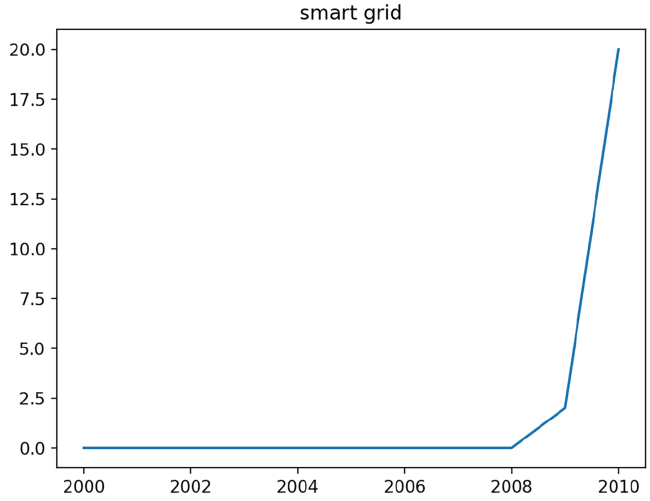

FIGURE 8 DF of the term *Smart Grid* in *dblp*

TABLE 9 Top first 10 emerging topics obtained as a result based on *dblp* (2000–2010)

Term	Av. date	DF	TF
Alk	2010	28	28
Alk paper	2010	28	28
Smart grid	2009,91	22	23
TerraSAR-X	2009,8	41	41
Cone metric	2009,71	21	21
Cloud computing	2009,63	142	142
G expansion	2009,62	26	26
LTE	2009,61	74	78
Internet things	2009,6	25	25
Two-way relay	2009,54	41	41

TABLE 10 Top first 10 emerging topics obtained as a result based on *dblp* (2010–2021)

Term	Av. date	DF	TF
COVID-19 patients	2020,39	36	36
COVID-19 infection	2020,34	29	29
COVID-19 lockdown	2020,34	29	29
Detection COVID-19	2020,34	44	44
RIS-aided	2020,33	21	21
COVID-19 detection	2020,33	52	52
Offline reinforcement	2020,32	31	31
COVID-19 diagnosis	2020,32	31	31
Prediction COVID-19	2020,32	22	22
Model COVID-19	2020,31	29	29

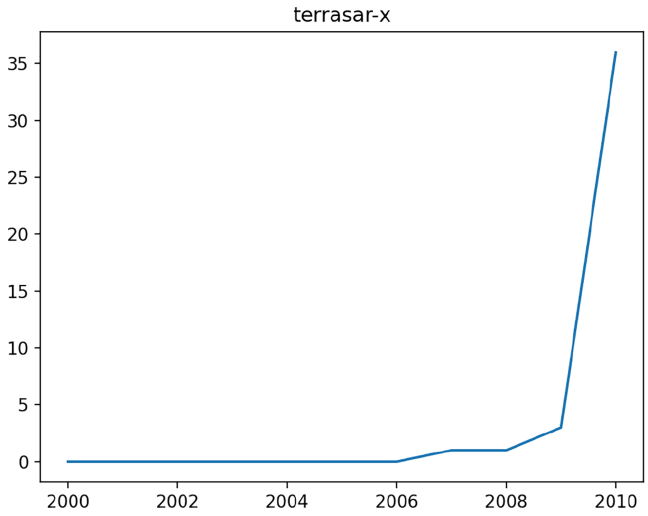

FIGURE 9 DF of the term *TerraSAR-X* in *dblp*

2000–2010 and obtained as a result using as a source the metadata from *dblp*.

4.4 | Discussion

The framework described in Section 3 can be optimized by selecting adequate values of the DF and TF thresholds. For collections containing a significant amount of text —such as Doxa and EThOS—, $f_{\min} = 3$, and $g_{\min} = 20$ proved to work satisfactorily. For shorter texts, such as titles provided by *dblp*, additional tests were done by using a lower value for $g_{\min}$ (e.g., 10). The results were similar, however, there were some differences which indicated that further configuration and optimization are needed to achieve the optimal performance.

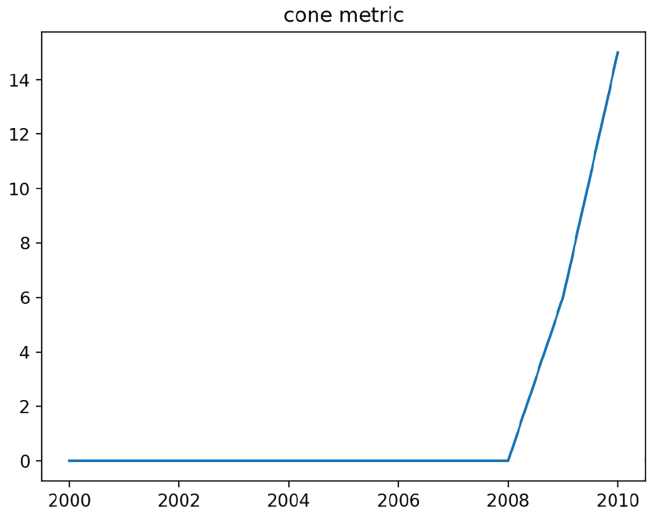

FIGURE 10 DF of the term *cone metric* in *dblp*

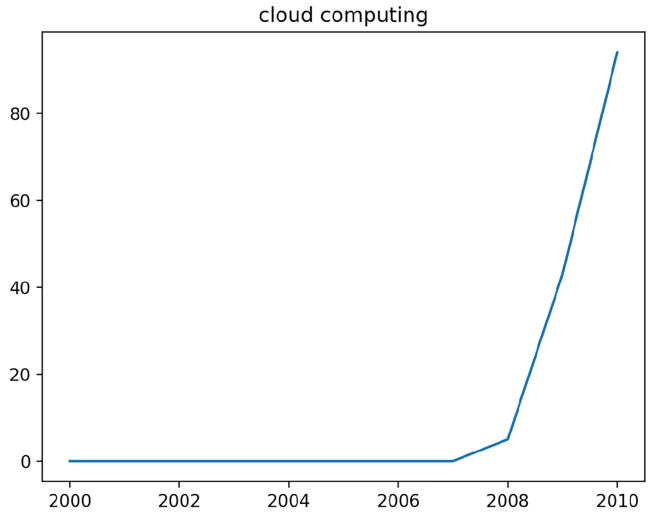

FIGURE 11 DF of the term *Cloud Computing* in *dblp*

TABLE 11 Top five 10 emerging topics obtained as a result based on Library of Congress' Chronicling America dataset (Bourbon News, 1913–1920)

Term	Av. date	DF	TF
It saves you	1920	36	36
By phone for	1920	27	27
You must say	1920	28	28
DOLLARS Cash in	1920	13	22
Win style	1920	28	28

The application of the methodology described here did not always lead to satisfactory results. For example, preliminary tests were performed on the Chronicling America newspapers collection* (Bourbon News section) provided by the Library of Congress. An example to reproduce the results is included in the Jupyter Notebook collection. The output included noisy terms due to the low accuracy of the published transcription, which consists of unsupervised OCR of documents that often include unusual text styles and small fonts (see Table 11). Although this method is sensitive to OCR quality, it will not reveal most common errors as emerging terms are not the most frequent ones. One could, however, employ a probabilistic language model to detect mistakes in the transcription, as characters recognized incorrectly usually appear as content that deviates from the standard morphemes in the target language.

The last step of the methodology (human revision) is particularly relevant. Issues solved include:

- Ambiguous acronyms (such as CER, which according to EThOS could be expanded to Conditioned Emotion Response or Corporate Environmental Responsibility).
- Structural contents such as section titles (e.g., Section B).

Although the overall assessment by the content managers was positive, selecting the right period to achieve good performance is a challenging task due to several reasons including the quality of the content, the number of years included in the collection or the number of results to assess. In this way, the knowledge of the content curators contributes significant value to the assessment process.

During the evaluation of the results on the Doxa journal, the curators identified a number of potential impacts of the technique: (a) The enhancement of the catalog with suggestions for search keywords based on the identified emerging topics; (b) visualizations based on the emerging topics of a particular journal depicting the dynamic behavior of terms over time, as hint for researchers; and (c) the identification of relevant items, such as popular works and topics, connected to a specific author, by analyzing author-oriented collections such as the *Anales Galdosianos* journal.*

5 | CONCLUSIONS

The framework described in Section 3 provides a simple method to identify emerging topics in newspapers, bibliographies, and journals. The method has been applied to three datasets provided by relevant GLAM institutions. The overall assessment by the content curators regarding the journal Doxa was positive obtaining a success rate of 80%, which was obtained by analyzing a period of 10 years.

The framework can be optimized by means of several parameters such as document and term frequency

thresholds. The adoption of this procedure by other institutions may require an adjustment of the configuration in order to obtain optimal results.

The differences of the datasets reused in our approach and the results obtained are useful to promote the reuse of the digital collections based on computational-driven research within GLAM institutions. In addition, OCR quality can pose a challenge when reusing text corpora.

Future work to be explored includes the normalization of the number of articles per year and the exploitation by means of visualization techniques. The definition of a semantic vocabulary to describe the model as well as the outputs will be explored. We plan to explore datasets, such as the Current Research Information Systems[†] digital repository, to check their applicability to support funding bodies in identifying promising areas of research.

ACKNOWLEDGMENTS

The authors thank the personnel from BVMC for their participation in the evaluation of the results for the journal Doxa.

ORCID

Gustavo Candela https://orcid.org/0000-0001-6122-0777

ENDNOTES

[*] https://chroniclingamerica.loc.gov/

[†] Project Jupyter, jupyter.org.

[*] https://chroniclingamerica.loc.gov/about/api/

[*] https://github.com/hibernator11/notebook-emerging-topics-corpora

[†] https://mybinder.org/

[‡] numpy.org

[§] Pandas, pandas.pydata.org.

[¶] www.scipy.org

[‖] NLTK, nltk.org.

[**] https://www.rdaregistry.info/Elements/m/

[††] https://doxa.ua.es

[‡‡] https://pypi.org/project/tika/

[*] https://dblp.org/

[*] https://chroniclingamerica.loc.gov/ocr/

[*] http://www.cervantesvirtual.com/portales/anales_galdosianos/

[†] https://dspacecris.eurocris.org/

REFERENCES

Baillieul, B., John, O. H., Larry, M. F. M., Jose, S. H., Sheila, G. S., Grenier, G., Forster, B., Michael, F. Z., Keaton, J., McCormick, D., & L. K. Moore. (2017). The First IEEE Workshop on the Future of Research Curation and Research Reproducibility, https://open.bu.edu/handle/2144/39028

Baillieul, J., Grenier, G., & Setti, G. (2018). Reflections on the future of research curation and research reproducibility [point of view]. *Proceedings of the IEEE*, *106*(5), 779–783.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. http://arxiv.org/abs/1801.09536

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S. A., & Kaplan, F. (2020). Combining visual and textual features for semantic segmentation of historical newspapers. https://arxiv.org/abs/2002.06144

Beals, M. & Bell, E. (2020, May). The atlas of digitised newspapers and metadata: Reports from oceanic exchanges.

Biblioteca Nacional de España. (2019). Comunidad BNE. https://bnelab.bne.es/en/tool/bne-community/

Bibliothèque nationale du Luxembourg. (2019). Historical newspapers. https://data.bnl.lu/data/historical-newspapers/

BnF Data (2014, August). 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, IFLA Information Technology Section; IFLA Semantic Web Special Interest Group; Bibliothèque nationale de France. http://ifla2014-satdata.bnf.fr/program.html

British Library. (2015) LibCrowds—Crowdsourcing projects from the British Library. https://www.libcrowds.com/

Buenaño-Fernández, D., González, M., Gil, D., & Luján-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access*, *8*, 35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983

Candela, G. & Carrasco, R. C. (2021, March). Notebook-emerging-topics-corpora: release1.4. https://doi.org/10.5281/zenodo.4637654

Candela, G., Sáez, M. D., Esteban, M. P. E., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. *Journal of Information Science*, *0*(0), 0165551520950246. https://doi.org/10.1177/0165551520950246

Carrasco, R. C. & Candela, G. dblp XML dataset as CSV for Python Data Analysis Library, March 2021. https://doi.org/10.5281/zenodo.4637205. https://dblp.uni-trier.de/xml/README.txt.

Chandrakala, D., Sumathi, S., Saran Kumar, A., & Sathish, J. (2019). Text clustering using PSO based dynamic adaptive SOM for detecting emergent trends. *IJIIT*, *15*(3), 64–78. https://doi.org/10.4018/IJIIT.2019070104

DARIAH-DE. Topics explorer (2020). https://dariah-de.github.io/TopicsExplorer/

Europeana. (2018). Europeana IIIF APIs. https://pro.europeana.eu/page/iiif

Europeana. (2020a). Newspapers. https://www.europeana.eu/es/collections/topic/18-newspapers

Europeana. (2020b). Issue 16: Newspapers. https://pro.europeana.eu/page/issue-16-newspapers

Hammou, B. A., Lahcen, A. A., & Mouline, S. (2020). Towards a real-time processing framework based on improved distributed recurrent neural network variants with fasttext for social big data analytics. *Information Processing & Management*, *57*(1), 102122. https://doi.org/10.1016/j.ipm.2019.102122

Heather Rosie. (2021, February). UK doctoral thesis metadata from EThOS. https://doi.org/10.23636/1344.

IIIF Consortium. (2014). International image interoperability framework, https://iiif.io/

JSTOR. (2017). Data for research—Sample datasets. https://www.jstor.org/dfr/about/sample-datasets

KB Labs. (2016). Smurf. http://labs.statsbiblioteket.dk/smurf/

KB Labs. (2018). Word2Vec. http://labs.statsbiblioteket.dk/dsc/

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*, 373–397.

Koerner, L., Caswell, T. A., Allan, D. B., & Campbell, S. I. (2020). A python instrument control and data acquisition suite for reproducible research. *IEEE Transactions on Instrumentation and Measurement*, *69*(4), 1698–1707. https://doi.org/10.1109/TIM.2019.2914711

Ley, M. (2009). DBLP—Some lessons learned. *Proceedings of the VLDB Endowment*, *2*(2), 1493–1500. https://doi.org/10.14778/1687553.1687577 http://www.vldb.org/pvldb/vol2/vldb09-98.pdf

Library of Congress. (2018). By the people. https://crowd.loc.gov/

Library of Congress. (2020). Newspaper navigator. https://labs.loc.gov/work/experiments/newspaper-navigator/

Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Derven, C., Dobreva-McPherson, M., Gasser, K., Chambers, S., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S.-C., & Wilms, L. (2019, September). Open a GLAM lab. International GLAM Lab Community.

National Library of Scotland. (2020). Jupyter notebooks. https://data.nls.uk/tools/jupyter-notebooks/.

Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC* 2016, Portorož, Slovenia, May 23–28, 2016, pp. 4348–4352, http://www.lrec-conf.org/proceedings/lrec2016/summaries/110.html

Open Archives Initiative. (2002). Open archives initiative protocol for metadata harvesting, https://www.openarchives.org/pmh/

Padilla, T., Allen, L., Frost, H., Potvin, S., Roke, E. R., & Varner, S. (2019, May). Final report—Always already computational: collections as data. https://doi.org/10.5281/zenodo.3152935

Pitt Rivers Museum. (2020). Terms of use for Pitt Rivers Museum Database of Photography Collections. https://www.prm.ox.ac.uk/databaseterms.html

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In Bernardi, R, Navigli, R, & Semeraro, G., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, . http://ceur-ws.org/Vol-2481/paper57.pdf

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. https://arxiv.org/abs/2002.04803

RDA Steering Committee and ALA Digital Reference. (2015). RDA registry. https://www.rdaregistry.info/

Romero, G. C., Esteban, M. P. E., Carrasco, R. C., & Such, M. M. (2018). Migration of a library catalogue into RDA linked open data. *Semantic Web*, *9*(4), 481–491. https://doi.org/10.3233/SW-170274

Romero, G. C., Esteban, M. P. E., Carrasco, R. C., & Such, M. M. (2019). A linked open data framework to enhance the discoverability and impact of culture heritage. *Journal of Information Science*, *45*(6). https://doi.org/10.1177/0165551518812658

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2018). Ten simple rules for reproducible research in Jupyter notebooks. http://arxiv.org/abs/1810.08055

Schlosser, M. (2009). Unless otherwise indicated: A survey of copyright statements on digital library collections. *College & Research Libraries*, *70*(4), 371–385. https://doi.org/10.5860/0700371, https://crl.acrl.org/index.php/crl/article/view/16022

Segal, Z., Soffer, O., Greidinger, N., Rusinek, S., & Silber-Varod, V. (2019). Computational analysis of historical Hebrew newspapers: Proof of concept. *Zutot*, *17*, 97–110. https://doi.org/10.1163/18750214-12171087

Sherratt, T. (2019, November). GLAM-workbench/getting-started. https://doi.org/10.5281/zenodo.3549636

Sung, Y., Jang, S., Jeong, Y.-S., & Park, J. J. (2020). Malware classification algorithm using advanced Word2vec-based bi-LSTM for ground control stations. *Computer Communications*, *153*, 342–348. https://doi.org/10.1016/j.comcom.2020.02.005

Tinits, P. (2020, July) GLAMlab toolkit to access and analyse texts at the National Library of Estonia. https://doi.org/10.5281/zenodo.3953795

Weinryb-Grohsgal, L. (2020). LC maps for robots. https://blogs.loc.gov/thesignal/2020/05/lc-maps-for-robots/

Wevers, M. & Lonij, J. (2017). SIAMESE. http://lab.kb.nl/tool/siamese

World Wide Web Consortium (2013, March). SPARQL 1.1 query language. https://www.w3.org/TR/sparql11-query/

How to cite this article: Candela, G., & Carrasco, R. C. (2022). Discovering emerging topics in textual corpora of galleries, libraries, archives, and museums institutions. *Journal of the Association for Information Science and Technology*, *73*(6), 820–833. https://doi.org/10.1002/asi.24583

APPENDIX A.: DATASETS AND GLAM INSTITUTIONS

The GLAM institutions provided in Table A1 are examples of organizations providing datasets that include newspapers and journals in several languages and formats. The datasets contain metadata and full text ready for reuse.

TABLE A1 A selection of GLAM institutions providing datasets that include newspapers and journals

Institution	Dataset collection	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 Nacional de Chile	Memoria Chilena	http://www.memoriachilena.gob.cl/
Bibliothèque nationale du Luxembourg	Historical newspapers	https://data.bnl.lu/
Biblioteca virtual Miguel de Cervantes	BVMC labs	http://data.cervantesvirtual.com/blog/labs
British Library	UK doctoral thesis metadata from EThOS	https://doi.org/10.23636/1344
Dblp computer science bibliography	dblp.xml	https://dblp.org/xml/
Det Kgl. Bibliotek	KB Labs	https://labs.kb.dk/
Europeana	Europeana IIIF APIs	https://pro.europeana.eu/page/iiif
Impact	Tools and resources	https://www.digitisation.eu/tools-resources/
Library of Congress	Chronicling America	https://chroniclingamerica.loc.gov/about/api/
National Library of Australia	Trove	https://trove.nla.gov.au/newspaper
National Library of Estonia	Workshop	https://github.com/peeter-t2/RR_GLAMlab_pilot
National Library of Netherlands	KB Lab	https://lab.kb.nl/
National Library of Scotland	Data Foundry	https://data.nls.uk/
Österreichische Nationalbibliothek	ONB Labs	https://labs.onb.ac.at/en/
Repositorio del Patrimonio cultural de México	Mexicana	https://mexicana.cultura.gob.mx/
Staatsbibliothek zu Berlin	SBB Labs	https://lab.sbb.berlin/?lang=en
The Atlas	Atlas of digitized newspapers and metadata	https://www.digitisednewspapers.net/