*ARTICLES*

# Measuring the Impact of Digital Heritage Collections Using Google Scholar

*Ángel Borrego*

**ABSTRACT**

*This study aimed to measure the impact of digital heritage collections by analysing the citations received in scholarly outputs. Google Scholar was used to retrieve the scholarly outputs citing Memòria Digital de Catalunya (MDC), a cooperative, open-access repository containing digitized collections related to Catalonia and its heritage. The number of documents citing MDC has grown steadily since the creation of the repository in 2006. Most citing documents are scholarly outputs in the form of articles, proceedings and monographs, and academic theses and dissertations. Citing documents mainly pertain to the humanities and the social sciences and are in local languages. The most cited MDC collection contains digitized ancient Catalan periodicals. The study shows that Google Scholar is a suitable tool for providing evidence of the scholarly impact of digital heritage collections. Google Scholar indexes the full-text of documents, facilitating the retrieval of citations inserted in the text or in sections that are not the final list of references. It also indexes document types, such as theses and dissertations, which contain a significant share of the citations to digital heritage collections.*

## INTRODUCTION

In recent years, many libraries have been devoting a large amount of resources, in terms of staff, equipment and infrastructure, to digitalize their special collections and to make them available on the web. The European Union "has invested €265 million in research and innovation for advanced digitisation technologies, digital curation and innovative cultural projects."[1] In most cases, the purpose of these initiatives is twofold: to facilitate access to scholars, and to the wider public, to rare and important materials, and to enhance long-term preservation.

Despite the benefits and value derived from these initiatives, there are few examples of evaluation of their impact. Much of the existing evidence remains anecdotal and the culture of assessment "has not yet penetrated digitization and digital collection building activities to nearly the same extent as many other areas of research library activity."[2]

Most previous research aimed at assessing the results of digitalization projects has focused on issues such as interface design, usability or users' information behaviour, whereas "studies about the impact of digital collections have not been conspicuous in the field."[3] A meta-analysis of 41 evaluations of Europeana revealed that system-centered evaluations prevailed over user-centered evaluations and "only a marginal number of studies tried to assess the impact of Europeana on different stakeholders."[4] More recently, a survey conducted by LIBER's working group on Digital Humanities and Digital Cultural Heritage showed that "digital humanities work within libraries is currently undergoing limited evaluation," with over half of the respondents not conducting any

**Ángel Borrego** (borrego@ub.edu) is Associate Professor, Universitat de Barcelona (Spain).


https://doi.org/10.6017/ital.v39i2.12053

specific assessment.[5] The report recommended that research libraries should measure their achievements and impact, to not only make decisions, prove success, and support arguments for resources if required, but also to provide new ways for academics to value the library.

According to Shaw, comprehensive assessment of digital collections requires a combination of methodological approaches, including statistics and surveys, user and usability studies, and web-based analytics.[6] The latter would encompass citation analysis, a method that could be employed to assess the reach and impact of digitized collections in a similar fashion to its use as a metric of the impact of scholarly works. Unfortunately, citation information for digitized collections is hard to capture due to the lack of specific guidelines in standard citation formats for these materials, resulting in inconsistent citation practices.

The impact of digitalization projects is sometimes measured in terms of visiting statistics, which are used as a proxy to evaluate the effectiveness of the resources devoted to digitalizing heritage collections. Biswas and Marchesoni, at the Hunter Library, were among the first authors to employ web analytics to obtain usage data from digital collections.[7]

Although enlightening, usage figures alone do not illuminate the reasons for usage and its impact, since download statistics do not indicate whether users find digital collections useful for learning, teaching, research, or leisure purposes. In a different approach, Sinn conducted a bibliometric study aimed at determining the relationship between digital resources and historical research.[8] She analyzed references and figures in articles published in *American Historical Review* to observe how frequently and widely digital collections were used. She found that secondary materials were the most frequently employed digitized resource, with archival materials coming in second place. Digital archival materials were more frequently mentioned in figures than in citations, proving the difficulty in compiling citation information for digitized collections.

The present study aimed to emulate Sinn's pioneering study on the use of citation analysis to measure the impact of digitized heritage collections, expanding our understanding of how digital heritage collections are used for academic and scholarly purposes. To do this we introduced two changes to Sinn's design related to the population of citing documents and the tool employed to retrieve the citations. First, instead of analyzing the references in a sample of journals in a given field to identify citations to digital collections, we retrieved all the citations to a specific digital collection that was used as a case study. Therefore, we did not measure how digital collections are cited in journals in a given discipline, but how scholarly outputs in different formats and disciplines cite a specific digital heritage collection. The collection used as a case study is *Memòria Digital de Catalunya* (MDC, http://mdc1.csuc.cat/en), a cooperative open-access repository containing digitized collections related to Catalonia and its heritage. The project is promoted by the universities of Catalonia and the Biblioteca de Catalunya, with the participation of other Catalan institutions.

Second, we used Google Scholar to retrieve the citations to MDC. The rationale for this choice was to obtain an accurate picture of the diversity of research outputs and disciplines in which digitized collections are used. Previous research has shown Google Scholar to be reliable and to have good

coverage of the diversity of disciplines, languages, and document types in the humanities and the social sciences, where usage of heritage collections is expected to be highest.[9]

The study aimed to address two questions:

1. To what extent are MDC collections being used in the creation of scholarly outputs?
2. Is Google Scholar a useful tool to measure the impact of digital heritage collections?

**METHODOLOGY**

On February 23, 2019, we searched Google Scholar for documents including a web reference to MDC in the full-text. The server hosting MDC has changed its URL several times since the inauguration of the service in 2006, so we used six queries to retrieve as many records as possible: *mdc.cbuc.cat*, *mdc.csuc.cat*, *mdc1.cbuc.cat*, *mdc2.cbuc.cat*, *mdc1.csuc.cat* and *mdc2.csuc.cat*. All queries, except for the last one, retrieved some records, giving a total of 366 results. For each record, we accessed the full-text of the document. At this stage, we removed 42 duplicates, i.e., copies of the same document hosted on two servers that Google Scholar had not been able to match. Additionally, two records were no longer available at the URL listed in Google Scholar and were removed from the analysis, leaving 322 citing documents.

In order to download the records, we used the "My library" feature in Google Scholar. This service allows users to export records in four formats: BibTeX, EndNote, RefMan, and CSV. Exported records had eight fields, although the level of completion was different for each field. "Authors" and "title" were provided for all the records, but the level of completion was lower for the "publication" (53 percent), "volume" (24 percent), "number" (33 percent), "pages" (44 percent), "year" (85 percent), and "publisher" (34 percent) fields. In order to analyse the evolution in the number of citations by year of publication, we manually retrieved the year of publication for 46 additional documents, thus covering 99 percent of the records.

For each of the 322 citing documents, we searched the full-text for the citation to MDC. However, 48 documents were behind a paywall and we were unable to access them. As a result, the population of citing documents for the second part of the analysis on cited MDC collections was reduced to 274 documents.

Most citing documents included a single reference to MDC, but, in some cases, the number of citations to MDC was higher, with an extreme case of an analysis of medical cartoons citing 323 resources in MDC. In order to analyse the results, when one document cited different resources in a single MDC collection, we counted this as a single citation. This was, for instance, the case for the medical cartoons example, since all references cited the same magazine. However, when a single document cited different MDC collections we counted as many citations as collections were cited. For instance, if a document cited a digitized resource in a parchment collection plus an article belonging to a collection of digitized magazines, we counted two citations. In total, the 274 citing documents contained 313 citations.

Citations were made at very different levels. In some cases, authors referred to the whole MDC website, citing the generic URL of the platform. In other cases, they cited a specific collection in MDC such as "Incunabula" or "Manuscripts." Some references cited a certain digitized magazine

whereas, in other cases, they referred to a specific issue or an individual article in a magazine. Additionally, many links to MDC were broken, forcing us to gather much information manually.

In order to compare the coverage of Google Scholar with that of Scopus, we also searched Scopus using the "Reference website" option in the advanced search. This option retrieves the URL of a website of a cited reference. We used the same six queries previously employed in Google Scholar. We only found twelve records when searching for *mdc.cbuc.cat*, whereas the other five queries did not retrieve any results.

## RESULTS

### *Citing documents: chronological evolution, document types, disciplines, languages, and coverage of Google Scholar compared to Scopus*

The number of documents citing MDC has grown steadily since its creation in 2006, reaching a maximum of 61 documents published in 2016 citing the repository (see figure 1).

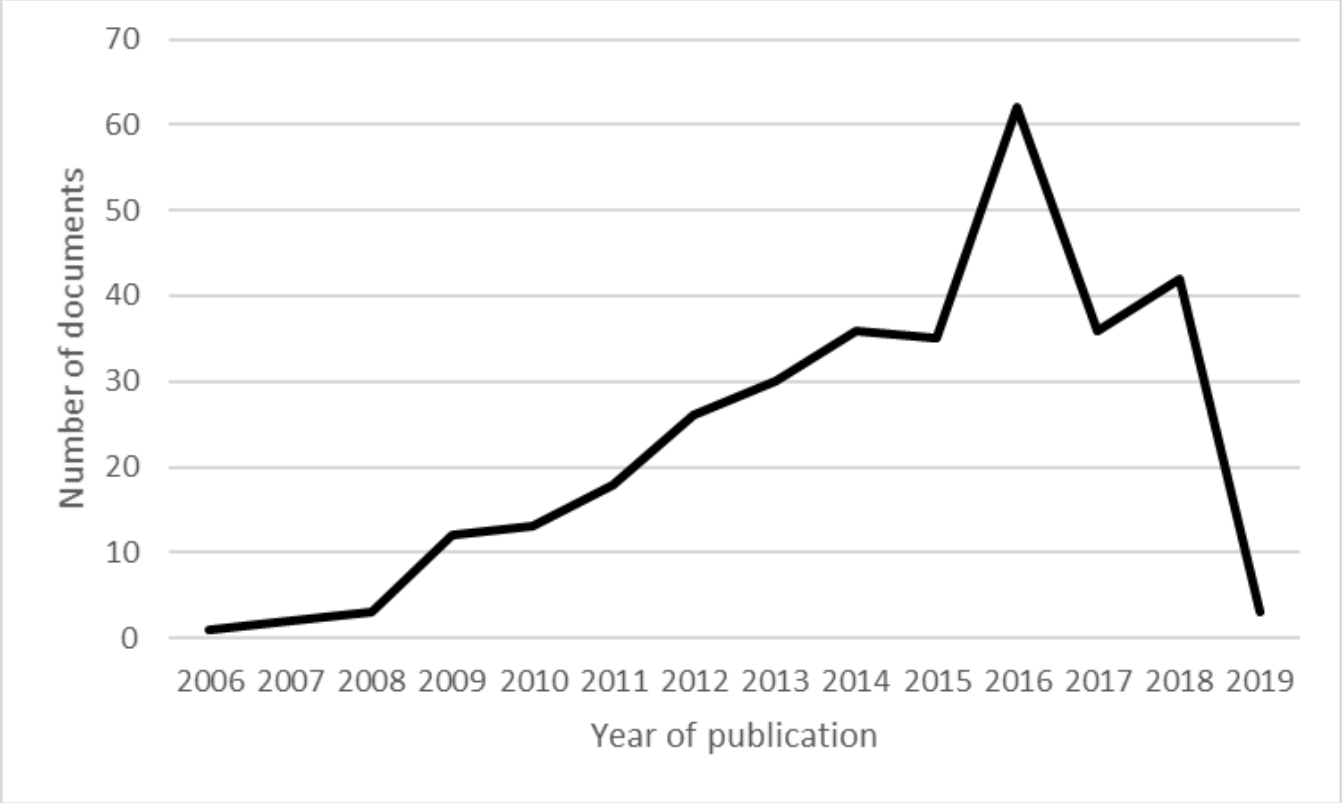



**Figure 1.** Documents citing MDC by year of publication, January 2006–February 2019 (n = 319).

Documents citing MDC can be broadly classified into two large categories: published scholarly outputs, such as journal articles, proceedings, or monographs (68 percent of the citing documents), and academic theses and dissertations (25 percent), with the remaining 8 percent of the citing documents including reports, open educational resources (OERs), course syllabuses, conferences, etc. (see table 1).

**Table 1.** Typologies of Documents Citing MDC (n = 322)

| | | n (%) |
|---|---|---|
| Scholarly outputs | Journal articles | 167 (52%) |
| | Proceedings | 22 (7%) |
| | Book chapters | 20 (6%) |
| | Books | 9 (3%) |
| Theses and Dissertations | Undergraduate | 32 (10%) |
| | Postgraduate | 25 (8%) |
| | PhD | 22 (7%) |
| Other | Academic essays, conferences, datasets, open educational resources (OERs), reports, syllabuses, etc. | 25 (8%) |

More than half of the documents citing MDC were journal articles. These articles had been published in 126 different journals. In order to determine the disciplines of the journals citing MDC, we searched for them in *Ulrich's Periodicals Directory*. Figure 2 shows the subject of the 148 articles published in journals indexed in *Ulrich's Periodicals Directory* that we were able to retrieve; the remaining 19 articles had been published in journals not indexed in *Ulrich's Periodicals Directory*.

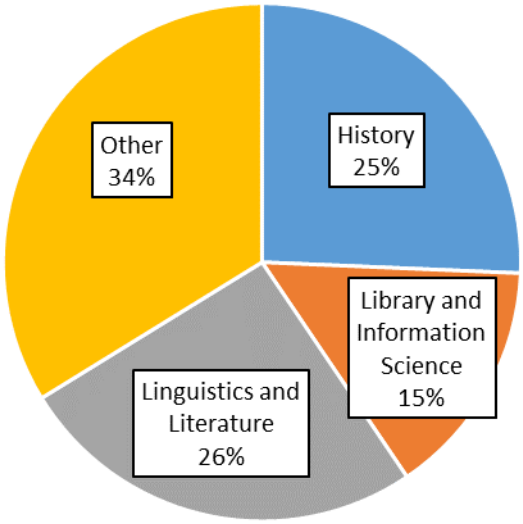



**Figure 2.** Disciplines of the Articles Citing MDC (n = 148)

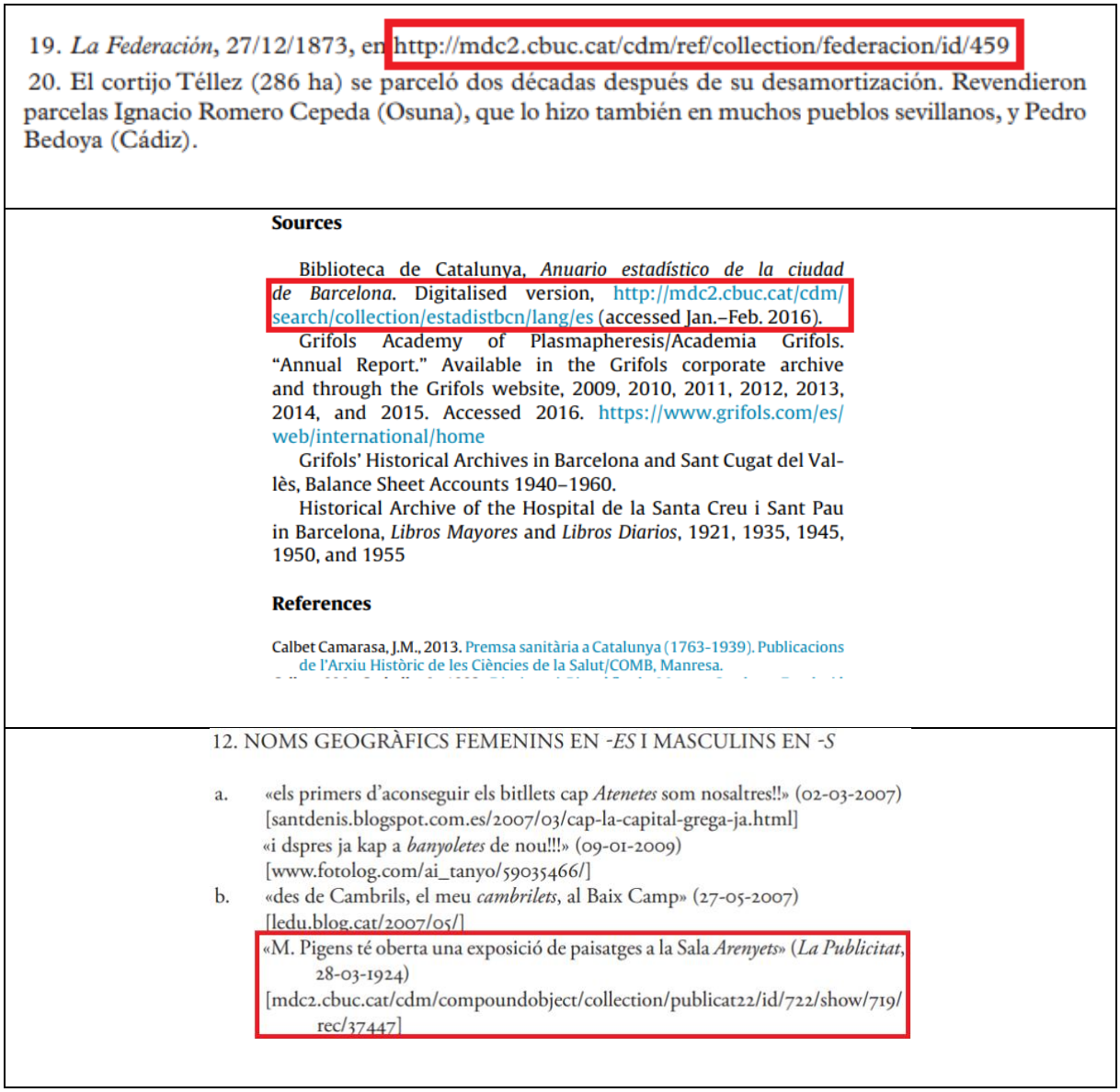
As shown in figure 2, half of the articles citing MDC were published in two humanities disciplines: linguistics and literature (26 percent) and history (25 percent). Additionally, a third of the articles (34 percent) had been published in journals in different fields in the social sciences and humanities, including education (6 percent), music (4 percent), art (3 percent), religion (2 percent), anthropology (1 percent), philosophy (1 percent), and political science (1 percent), among others. There were also a few cases of citing articles published in journals in the

experimental, health, and natural sciences (3 percent), usually in studies employing a historical approach to these disciplines. In all cases, the authors of the articles had used MDC as a primary source for their research. Finally, 15 percent of the articles citing MDC had been published in library and information science journals. In most cases, these articles had been written by librarians who described the features of MDC to a professional audience in their field.

Consistent with the high number of outputs in the humanities and the social sciences, most citing documents were in local languages, i.e., 48 percent had a title in Catalan and 35 percent in Spanish. The remaining outputs were in English (11 percent) or in other languages (7 percent).

When we looked at the coverage of the citing journals in Scopus, we observed that 37 (29 percent) of the 126 citing journals were indexed by the database. These 37 journals had published 49 articles citing MDC. These results were inconsistent with those obtained when searching Scopus for MDC references, when we were able to retrieve just twelve records. We followed up these discrepancies, i.e., the articles published in journals indexed in Scopus whose MDC citations were retrieved in Google Scholar but not in Scopus, and in most cases, the reason for the discrepancy was that MDC citations were not in the list of references at the end of the article. The references to MDC were inserted in the text or located in footnotes, annexes, lists of websites, etc. In some cases, MDC was cited in document types not indexed by Scopus, such as book reviews.

Figure 3 shows three examples of these discrepancies, i.e., MDC references cited in articles indexed in Scopus but not retrievable when searching the database. In the first example, the reference to MDC was in a footnote but not in the final reference list. In the second example, the reference to MDC was included in a list of "sources" placed before the reference list that starts at the bottom of the image. Finally, in the third example, the reference to MDC was included in an annex, but not in the list of references. In all three cases, the articles were indexed in Scopus, but the records did not include the references to MDC. Conversely, all three documents were retrieved in Google Scholar since it had indexed the full-text.



19. *La Federación*, 27/12/1873, en http://mdc2.cbuc.cat/cdm/ref/collection/federacion/id/459

20. El cortijo Téllez (286 ha) se parceló dos décadas después de su desamortización. Revendieron parcelas Ignacio Romero Cepeda (Osuna), que lo hizo también en muchos pueblos sevillanos, y Pedro Bedoya (Cádiz).

**Sources**

Biblioteca de Catalunya, *Anuario estadístico de la ciudad de Barcelona*. Digitalised version, http://mdc2.cbuc.cat/cdm/search/collection/estadistbcn/lang/es (accessed Jan.–Feb. 2016).

Grifols Academy of Plasmapheresis/Academia Grifols. "Annual Report." Available in the Grifols corporate archive and through the Grifols website, 2009, 2010, 2011, 2012, 2013, 2014, and 2015. Accessed 2016. https://www.grifols.com/es/web/international/home

Grifols' Historical Archives in Barcelona and Sant Cugat del Vallès, Balance Sheet Accounts 1940–1960.

Historical Archive of the Hospital de la Santa Creu i Sant Pau in Barcelona, *Libros Mayores* and *Libros Diarios*, 1921, 1935, 1945, 1950, and 1955

**References**

Calbet Camarasa, J.M., 2013. Premsa sanitària a Catalunya (1763-1939). Publicacions de l'Arxiu Històric de les Ciències de la Salut/COMB, Manresa.

12. NOMS GEOGRÀFICS FEMENINS EN *-ES* I MASCULINS EN *-S*

a. «els primers d'aconseguir els bitllets cap *Atenetes* som nosaltres!!» (02-03-2007) [santdenis.blogspot.com.es/2007/03/cap-la-capital-grega-ja.html]
«i dspres ja kap a *banyoletes* de nou!!!» (09-01-2009) [www.fotolog.com/ai_tanyo/59035466/]

b. «des de Cambrils, el meu *cambrilets*, al Baix Camp» (27-05-2007) [ledu.blog.cat/2007/05/]
«M. Pigens té oberta una exposició de paisatges a la Sala *Arenyets*» (*La Publicitat*, 28-03-1924) [mdc2.cbuc.cat/cdm/compoundobject/collection/publicat22/id/722/show/719/rec/37447]

**Figure 3.** Examples of MDC citations in articles indexed in Scopus not retrievable in the database. (Top: Antonio López Estudillo, "Especialización olivarera, cambios institucionales y desigualdad agraria en la Alta Campiña de Córdoba (siglos XVIII-XX)," *Historia Agraria* 73 (December 2017): 185–220, https://doi.org/10.26882/HistAgrar.073E07l; middle: Paloma Fernández Pérez and Ferran Sabaté Casellas, "Entrepreneurship and management in the therapeutic revolution: The modernisation of laboratories and hospitals in Barcelona, 1880–1960," *Investigaciones de Historia Económica – Economic History Research* 15, no. 2 (June 2019): 91–101, https://doi.org/10.1016/j.ihe.2017.09.001; bottom: Maria-Rosa Lloret, "La sufixació apreciativa del català: creacions lèxiques i implicacions morfològiques," *Caplletra: Revista Internacional de Filologia* 58 (2015): 55–89, https://doi.org/10.7203/caplletra.58.7137.)

In other cases, although the citations to MDC were in the final list of references, Scopus failed to retrieve them. Figure 4 shows an example of a complete reference in an original document (including the URL pointing to MDC), whereas the Scopus record, at the bottom of the image, just

included the name of the author and the title of the magazine. The MDC reference, therefore, was not retrieved when searching Scopus.

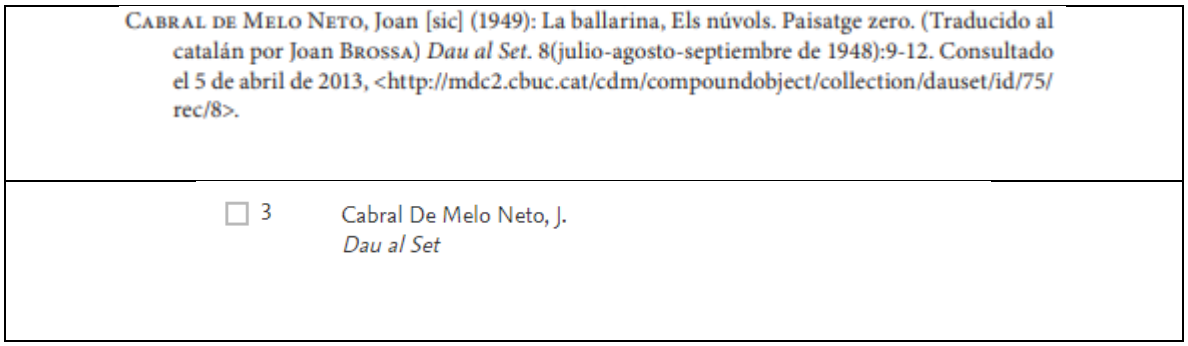



**Figure 4.** Differences between a reference in the original document and the record in Scopus. (Source of the example: Ramon Farrés, "La recepción del poeta catalán Joan Brossa en Brasil," *Meta: journal des traducteurs* 60, no 1 (April 2015): 158–72, https://doi.org/10.7202/1032404ar.)

Leaving aside scholarly outputs published in the form of journal articles, proceedings or monographs, the second category of MDC-citing documents (25 percent) was that of dissertations and theses. This included dissertations at all academic levels: undergraduate, postgraduate and PhD. In nearly all cases, dissertations were hosted in University institutional repositories and belonged to disciplines similar to those recorded for scholarly outputs, such as language and literature or history and, more generally, to the humanities and the social sciences.

Finally, up to 75 citing documents (8 percent) were classified in other typologies. As in the case of journal articles, this section included several examples of reports and other outputs prepared by librarians explaining the features of MDC to a professional audience.

***Cited collections in MDC***

The population of 274 citing documents available in open access sources included 313 references to MDC. One-fifth of the citations did not refer to any specific collection or document, but to the whole MDC portal (see table 2).

**Table 2.** Cited Sources in MDC (n = 313).

| | **n (%)** |
|---|---|
| Digital Memory of Catalonia (MDC) | 67 (21%) |
| Archive of Ancient Catalan Periodicals (ARCA) | 111 (35%) |
| Other collections | 135 (43%) |

The most cited collection in MDC was ARCA, a collection of digitized ancient Catalan periodicals, which received more than one third of the citations. These citations, however, were not strictly comparable, since four citations referred to the whole ARCA collection, 33 citations referred to a specific magazine, six citations referred to a specific issue, and 68 citations referred to a specific article. Thirty different digitized magazines were cited.

Finally, 135 citations referred to 49 different collections. Just three collections received more than ten citations each: a collection of digitized posters from the Spanish Second Republic and Civil War (15 citations), a photographic collection of the Hiking Club of Catalonia (11 citations), and a collection of manuscripts from the Biblioteca de Catalunya (11 citations).

## DISCUSSION AND CONCLUSIONS

The results of the study provide evidence of the academic impact of MDC collections among scholars and students. Academics make use of the digitized resources hosted by MDC and find them useful to build their scholarship, as evidenced by the citations made in their scholarly outputs. However, its impact goes beyond academic publications such as journal articles, proceedings, and monographs, including a significant number of dissertations and theses citing MDC resources.

Professional scholars accessing MDC collections online may save time and money compared to the travel costs required to consult the resources in situ. However, in the case of students, it is possible that in many cases they would be unable to access the collections had they not been digitized. When considering this type of impact, it should be borne in mind that the actual number of academic essays, dissertations, and theses citing MDC is presumably higher than the number recorded in this study. While all PhD theses in Catalonia are posted online, only bachelor and master dissertations with high grades are archived online in institutional repositories and, therefore, could be retrieved in our study.

As expected, given the characteristics of the digitized collections, most citations come from academic documents in local languages in the fields of humanities and social sciences. Most cited resources are located in a collection of digitized ancient Catalan periodicals, although citations are spread among a wide range of collections. In addition to the use of MDC by scholars and students, there is also a significant proportion of professional publications written by librarians citing MDC.

The results of the study show that Google Scholar is possibly the most suitable tool for conducting citation studies on the impact of digital heritage collections. Given the characteristics of the resources contained in these collections, most citations are retrieved from local journals, which are frequently not indexed in large citation indexes such as Web of Science or Scopus. Even when they are indexed, our results show that Scopus frequently fails to properly index citations to digital heritage collections, since most citations are not included in the final list of references but are inserted in the text, in footnotes, or in lists of resources employed. Conversely, Google Scholar indexes the full-text of documents and, therefore, allows retrieval of these citations. These results are consistent with those found in previous research that showed that digital archive materials are more frequently mentioned in figures than in citations and described how citation information for digitized collections is hard to capture due to the lack of specific guidelines for citing digitized collection materials, resulting in inconsistent citation practices.[10]

In addition, as mentioned previously, a significant proportion of citations to digital heritage collections originates from documents indexed by Google Scholar, but not by Scopus, especially theses and dissertations, but also reports, educational resources, etc.

Although our study provides evidence of the scholarly impact of MDC, it also suffers from some limitations. Our results do not reflect any other kind of impact, such as on learning, teaching, or leisure. Similarly, our study did not consider how MDC contributes to the long-term preservation of heritage collections. Further research could explore these issues using participatory research methods, including surveys and interviews with users.

**ENDNOTES**

[1] "EU Member States Sign Up to Cooperate on Digitising Cultural Heritage," European Commission, last updated April 24, 2020, https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-digitising-cultural-heritage.

[2] Emily Frieda Shaw, "Making Digitization Count: Assessing the Value and Impact of Cultural Heritage Digitization," *Archiving 2016 Final Program and Proceedings* (Springfield, VA: Society for Imaging Science and Technology, 2016), 197, https://doi.org/10.2352/issn.2168-3204.2016.1.0.197.

[3] Donghee Sinn, "Impact of Digital Archival Collections on Historical Research," *Journal of the American Society for Information Science and Technology* 63, no. 8 (August 2012): 1521, https://doi.org/10.1002/asi.22650.

[4] Vivien Petras and Juliane Stiller, "A Decade of Evaluating Europeana—Constructs, Contexts, Methods & Criteria," in *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice*, eds. Jaap Kamps et al. (Cham: Springer, 2017), 241, https://doi.org/10.1007/978-3-319-67008-9_19.

[5] LIBER, "Europe's Digital Humanities Landscape: A Report from LIBER's Digital Humanities & Digital Cultural Heritage Working Group," (2017), 28, https://doi.org/10.5281/zenodo.3247286.

[6] Shaw, "Making Digitization Count," 198.

[7] Paromita Biswas and Joel Marchesoni, "Analyzing Digital Collections Entrances: What Gets Used and Why It Matters," *Information Technology and Libraries* 35, no. 4 (2016): 19–34, https://doi.org/10.6017/ital.v35i4.9446.

[8] Sinn, "Impact of Digital Archival Collections," 1525.

[9] Alberto Martín-Martín et al., "Google Scholar, Web of Science, and Scopus: A Systematic Comparison of Citations in 252 Subject Categories," *Journal of Informetrics* 12, no 4 (November 2018): 1160–77, https://doi.org/10.1016/j.joi.2018.09.002.

[10] Sinn, "Impact of Digital Archival Collections," 1533**;** Shaw, "Making Digitization Count," 199.