Int J Digit Libr (2018) 19:151–161
https://doi.org/10.1007/s00799-017-0210-1

The references of references: a method to enrich humanities library catalogs with citation data

Giovanni Colavizza[1] · **Matteo Romanello**[1] · **Frédéric Kaplan**[1]

Received: 30 September 2016 / Revised: 22 February 2017 / Accepted: 28 February 2017 / Published online: 8 March 2017

Abstract The advent of large-scale citation indexes has greatly impacted the retrieval of scientific information in several domains of research. The humanities have largely remained outside of this shift, despite their increasing reliance on digital means for information seeking. Given that publications in the humanities have a longer than average life-span, mainly due to the importance of monographs for the field, this article proposes to use domain-specific reference monographs to bootstrap the enrichment of library catalogs with citation data. Reference monographs are works considered to be of particular importance in a research library setting, and likely to possess characteristic citation patterns. The article shows how to select a corpus of reference monographs, and proposes a pipeline to extract the network of publications they refer to. Results using a set of reference monographs in the domain of the history of Venice show that only 7% of extracted citations are made to publications already within the initial seed. Furthermore, the resulting citation network suggests the presence of a core set of works in the domain, cited more frequently than average.

Keywords Digital libraries · Bibliometrics · Citation extraction · Information retrieval · History of Venice

✉ Giovanni Colavizza
giovanni.colavizza@epfl.ch

Matteo Romanello
matteo.romanello@epfl.ch

Frédéric Kaplan
frederic.kaplan@epfl.ch

[1] Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne, Station 14, 1015 Lausanne, Switzerland

1 Introduction

The humanities are the Cinderella of sciences with respect to citation-driven information retrieval. The lack of citation indexes not only prevents the quantitative analysis of the field's communication practices [2], but hinders the daily work of researchers, for whom the manual look-up of reference lists (reference chaining) is still the most important way to learn the state of the art on a topic of interest [4]. Mainly for this reason, research libraries in the humanities, which are oftentimes able to build collections responding to most needs of scholars in the humanities [10], still devote entire sections to *reference works* in specific domains. Reference works are deemed of importance within a domain of study and can be identified using library resources such as classification and shelving strategies. Their selection is usually done by librarians and domain experts, with the purpose of accelerating the retrieval of relevant literature by users.

There are several reasons for this state of affairs, yet the lack of citation data is the best known problem, lamented several times over [8,15,27]. For these and other reasons, the use of citations as a means to evaluate research in the humanities has also been questioned [28], with alternatives being proposed [7,17]. Coverage of services such as Web of Science and Scopus is still far from satisfying, albeit improving over time [19,30], both for journals [20] and monographs [36]. It is worth stressing that monographs are especially important, as the practice in the humanities still favours them over other kinds of publications in order to get recognition within the field, despite variations in citation patterns among different disciplines [12,33]. The humanities have also been found to identify core works at a slower peace than other sciences, in part due to longer times required for citation accumulation, entailing a longer life-span of publications [21].

The very presence of a core set of landmark, highly cited works in specific domains of the humanities, has been questioned, or at least problematized, in several studies. In a seminal work on the history of technology, McCain [18] struggled to identify even a small set of works emerging via citation analysis. Some have attributed this apparent absence to the lack of systematic information retrieval practices in the humanities [3]. Others have pointed out how the diversity of fields of inquiry, even within the same discipline, is simply too broad to allow for a set of works to be considered as a shared core [29]. As examples, Heinzkill [9] found that over 40% of the monographs cited from a set of articles in English and American literature fall outside of the field as individuated by library classification. Weingart [31], on the other hand, confirmed at least the presence of clear, if only minimally overlapping boundaries in the co-citation network of the history and philosophy of science. Lastly, the very notion of core works has been regarded as elusive and artificial [14]. Yet, if they existed, core works could help improve core collections within research libraries. In an attempt to pursue this approach, Nolen and Richardson [22] also highlight that "the combination of these research habits, the diversity of their topics, and the controversial aspect of attempting to define a 'core collection' provide very real barriers to identifying, selecting, and acquiring stand-out publications for a library collection in the humanities". The same authors suggest that a possible reason for this lack of success in identifying core works for humanities' disciplines lies in the focus on the disciplinary macro level, but they are still unable to find core works at the more granular levels of the domain or sub-domain of study. Despite such difficulties, some authors have managed to investigate citation patterns in the humanities using mixed strategies, in the direction taken by the present article. As an example, Hammarfelt [6] used citations to monographs coupled with library classification to investigate the shifts in the intellectual base of literature studies, showing an increase in interdisciplinary citing over time.

The automatic extraction of citations from scholarly publications is instead a more mature area of research. Recent developments include fully fledged architectures to extract and use citation data, embedded within digital library systems [34]. Several citation extraction services exist, such as ParsCit [5], BILBO [11], GROBID [16] and FreeCite.[1] Citing in the humanities is a less standardized practice than in other sciences. More specifically, reference lists at the end of a publication are optional, as citations are commonly made in footnotes. Furthermore, humanists developed elaborated practices for the abbreviation and encoding of references, which also entail making a variety of usages of formatting features such as italics or variations in type module. Lastly, it is common in the humanities to refer to both primary and secondary sources. In general, a primary source is the documentary evidence used to support a claim and a secondary source is a scholarly publication [32]. Unfortunately, these characteristics of citing in the humanities make it difficult to reuse existing services out-of-the-box, as it will become clear in what follows.

At the heart of the present article is a distinction between:

- *Reference monographs* the part of the reference works in a library and for a specific domain, which contain scholarly contributions in prose. Reference monographs, as reference works, are deemed of importance within a domain of study, and can be identified in a similar way. Yet if a set of reference works can in theory include scholarly publications, but also dictionaries and serial publications, bibliographies, edition of sources, and any other published material of importance, reference monographs consist only of the subset of these works which contain scholarly contributions in prose. As a result, reference monographs include for example critical editions of sources, but exclude historical dictionaries and bibliographies.
- *Core works* a small group of works which are considered foundational for a given domain of study, because they are highly cited by the existing literature and especially by reference monographs. Core works can be of any nature, not only scholarly contributions in prose.

Assuming the existence of a set of *reference monographs* for a given domain in the humanities, this article proposes to use it as a means to enrich library catalogs with citation data. Reference monographs can be identified by their physical location in the library (e.g. reference shelves), by catalog subject headings, and by scholarly bibliographies. These are, in fact, the only methods available to scholars to find relevant literature, excluding the process of reference chaining. Yet, to the best of our knowledge, no previous attempt has been made to use reference monographs as a seed for both citation analysis and the enrichment of library catalogs with citation data. The approach should therefore help focus the efforts of research libraries to index a subset of monographs which are likely to be of interest for the community. It will also help expanding reference works with core works previously not part of the reference set. In fact it is worth noting that such reference monographs, as identified by library resources, do not necessarily need to overlap with core works in any given domain. The question is to what extent they can span the literature of the domain of interest and help improve the information retrieval capacities of the library.

In order to demonstrate the viability of the proposed approach, this article presents:

- a full pipeline for the selection of reference monographs, the extraction of citations and their look-up in the library

[1] http://freecite.library.brown.edu/.

catalog. The pipeline accounts for the variability of citation practices, is mostly language-independent, and detects citations to primary sources, monographs or journal articles (secondary sources).
– A case study to validate whether a set of reference monographs, individuated using library resources, is in fact an effective hub pointing to the relevant literature within a domain of interest. An important question is also whether a set of core works, cited significantly above the average, emerges from the literature of the domain. The case study considers the domain of the history of Venice, and the collection of the Humanities Library at the University of Venice.

To avoid confusion due to the specific use of the term *reference* in the present article, the term *citation* is always used to identify a reference made from a citing work to a cited work, irrespective of the number of mentions (or in-text citations) given. The term *reference* is still used to indicate the text containing a citation. This article is organized as follows: Sect. 2 details our approach, Sect. 3 presents the results of the extraction pipeline applied to the case study, Sect. 4 investigates the effectiveness of reference monographs to span the given domain, and Sect. 5 concludes the paper.

2 Approach

The main steps of the proposed approach are illustrated in Fig. 1. Given a specific domain of interest in the humanities, the corpus of reference monographs is first selected, then digitized and OCRed. The manual annotation of a subset of citations is then followed by their automatic extraction over the whole corpus, using standard supervised machine learning techniques. Next, the look-up module finds matches into the library catalog and connects paired resources: the citing to the cited monograph. Eventually, the look-up module can be also used to evaluate the proportion of citations given to monographs already within the corpus itself, which we define as the cohesiveness of the corpus. It is worth noting that this approach is iterative: given a first seed of reference monographs, more of them can be found through the extracted citations from previous iterations.

The case study on the historiography on Venice considers only modern books (published from the year 1850).

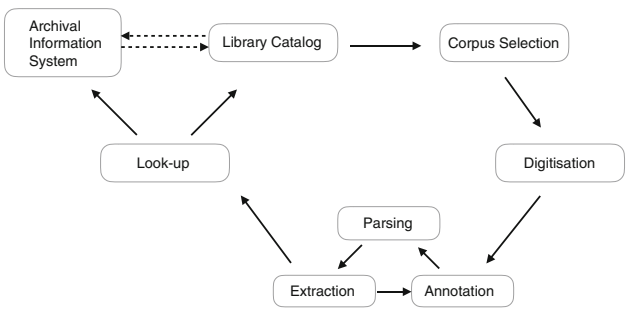

Fig. 1 The proposed pipeline for the enrichment of library catalogs with domain-specific citation data

2.1 Corpus selection and acquisition

Libraries can provide a first means to identify a set of reference monographs of interest within a specific domain of study, even more so if they specialize in this domain. For the purpose of this research, the Italian library catalog was used in order to extract:

1. all the resources on reference shelves marked as "History of Venice", from now on defined as (rapid) consultation monographs;
2. all the resources under subject history of Venice (e.g. Dewey code 945.31);
3. additional resources found by keyword search over the title (e.g. using words such as "Venice" in multiple languages) and manually selecting what is relevant, or by means if scholarly bibliographies (e.g. Zordan [35] repertory).

The outcome is a set of 1904 monographs. The number of monographs with a list of references is 836 (of which 201 are in consultation, defined as cat. 1), or 44% of the total. The monographs with reference lists are also equally distributed over time, as shown in Fig. 2. Of these, 700 (183 in cat. 1) have structured lists of references, as opposed to end notes. This last subset of 700 monographs with structured reference lists has been used to extract citations. The distribution of the

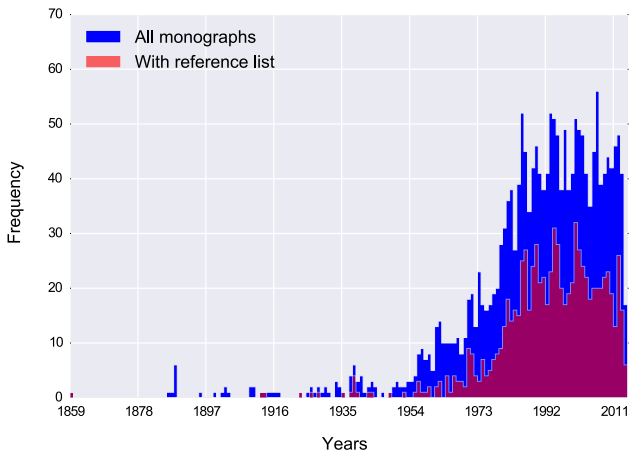

Fig. 2 Number of monographs in the corpus per year (*blue*/*black*), over the monographs with a reference list (*red*/*grey*): reference lists are uniformly distributed over time (color figure online)

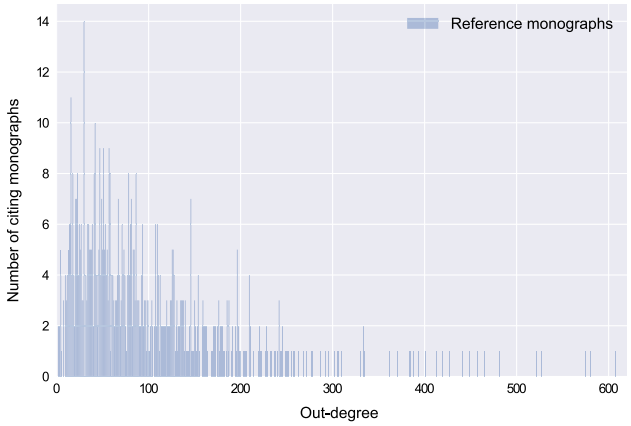

Fig. 3 Histogram of the out-degrees for the 700 citing monographs

number of citations made by these 700 monographs is given in Fig. 3. Values are reasonably between 20–30 and 300; more extensive reference lists are rare. Ideally, citations should be considered from the whole corpus, yet extracting them from footnotes is considerably more challenging than just using reference lists. At the same time, the absence of a systematic bias with respect to topic, period or publisher influencing the presence or absence of reference lists, allows us to proceed with this subset of the original corpus.

The second step in the pipeline is the classification of reference styles and the manual annotation of a subset of citations for each individuated class. As mentioned earlier, reference styles can be quite varied in the humanities, and change over time, author and publisher. Yet they convey important information for citation extraction. A *reference style* is a specific combination of elements in a reference, such as author and title, encoded in a predefined way (e.g. using quotation marks for the title). Styles can be grouped in *classes* and *families*. For example:

> De Virine, Theodore Low. Notable Printers of Italy during the Fifteenth Century. New York: The Grolier Club, 1910.

is a reference presenting the author's surname and then name separated by comma, title, place of publication, publisher and date. The punctuation and capital letters in use are particularly relevant. A different class stems from the elimination of at maximum one element, or one change in encoding. E.g. removing the publisher would create a new class of the same family. A different family is identified by at least two removals or additions of elements, and/or sensible changes in the encoding of the same information. For example:

> De Virine, T. L. Notable Printers of Italy during the Fifteenth Century. The Grolier Club, 1910.

would create a different class in a separate family as the author's name is now abbreviated and the publisher has been dropped. Classes and their families used as a feature for parsing have improved results in a sensible way, since the citations from a specific publication all belong to a unique class/family combination.

In total, 33 classes and 6 families were individuated.

Manual annotation was done over a randomly selected subset of citations for each class.[2] Annotations are divided into two categories: *generic* and *specific*. A generic annotation distinguishes the completeness of a reference (if full or abbreviated) and the type of object referred to (if a monograph or a contribution, such as a journal article). There can be 6 generic types of annotation: monograph, contribution, or article, all either full or partial. Specific annotations identify instead the components of generic categories. Examples of specific annotation tags are: "author", "title", "publisher".

Approximately 27% of the 700 monographs have been annotated, 2 pages of references each on average. As a consequence, circa 3.8% of all available pages with references have been annotated, for a total of 49,580 annotations, of which 8646 are generic (i.e. citations) and 40,934 specific (i.e. their components).

2.2 Citation parsing and extraction

The next component of the pipeline is a parser and citation extraction module, which performs two tasks:

1. *Citation parsing* given a text stream of lists of references, parse the text to assign the most likely specific tag to each token.
2. *Citation extraction and categorization* given a stream of tokens with specific tags, decide where a reference begins and ends, and assign a generic category to the citation ("monograph", "abbreviated reference" and "contribution").

Both parsers use Conditional Random Fields with the same set of features—except for specific tags resulting from task 1 that are used in task 2—a technique commonly adopted in similar settings, introduced by Lafferty et al. [13]. The order of the tasks has been determined empirically to maximize performance on a subset of specific tags (crucially author, title and year of publication), which are the most relevant for the look-up module. 8051 annotated references were used for training and testing, for a total of 122,612 tokens, or circa 15 tokens per reference, plus 35,124 negative tokens (outside of references).

[2] Using the Brat annotation environment available at http://brat.nlplab.org/. Cf. Stenetorp et al. [26].

2.3 Catalog look-up

Extracted citations need to be disambiguated in order to be used in the catalog. This task is performed by a look-up system that tries to match the components of the extracted citation against a library catalog. Given the nature of the data at hand, such look-up system needs to: have a good coverage of the domain; have the ability to work with a limited set of metadata fields as input; and have a high degree of tolerance for OCR errors.

The implemented solution attempts to match the metadata fields of the extracted citation against the bibliographic records contained in the catalog of the Italian National Library Service (SBN), which at the time of writing contains almost 16 million entries. This catalog provides a good coverage of the publications cited by reference monographs, which can be easily explained in light of the focus on the history of Venice.

A full dump of the SBN catalog was used, thanks to an ongoing collaboration with the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU), which owns the SBN catalog, and is responsible for its maintenance and updates.

The data are loaded onto an instance of ElasticSearch as it constitutes an efficient solution to search through such a large dataset. The catalog dump is in a JSON format derived from MARC21, the format in which the catalog records are originally stored. Each publication in the catalog is described according to the Italian national guidelines,[3] by means of the following metadata fields (see Fig. 4):

- the record's unique identifier ("codiceIdentificativo")
- the main author ("autorePrincipale")
- the title ("titolo")
- the indication of publisher, place and year of publication ("pubblicazione")
- the collection to which the publication belongs ("collezione")
- the type of document ("tipo")
- the level of the bibliographic description, e.g. monograph, journal, etc. ("livello")
- the document's physical description ("descrizioneFisica")
- the country of publication ("paesePubblicazione")
- the language(s) of the document ("linguaPubblicazione")
- the normalized form of entities mentioned in the record ("nomi")
- other identifiers for the publication, e.g. its ISBN number ("numeri")
- subject classification according to SBN's classification system ("soggetti)

[3] The last version is detailed in SBN [25].


```
{
  "codiceIdentificativo":"IT\ICCU\LO1\1162930",
  "autorePrincipale":"Infelise, Mario",
  "titolo":"L'editoria veneziana nel '700 /
   Mario Infelise",
  "pubblicazione":"Milano : F. Angeli, 2005",
  "collezione":"Saggi di storia ; 6",
  "tipo":"Testo a stampa",
  "livello":"monografia",
  "descrizioneFisica":"426 p. ; 22 cm.",
  "paesePubblicazione":"IT",
  "linguaPubblicazione":"ita",
  "nomi":[
    "Infelise, Mario"
  ],
  "numeri":[
    {
      "ISBN":"88-464-1967-7"
    }
  ],
  "soggetti":[
    "EDITORIA - Venezia - Sec. 18."
  ],
  "dewey":[
    {
      "cdd":"070.5094531",
      "ed":"20",
      "dec":"Editoria. Venezia"
    }
  ],
  "localizzazioni":[
    {
      "isil":"MN0120"
    },
    {
      "isil":"TO0529"
    }
  ]
}
```


Fig. 4 An example of SBN catalog record

- subject classification according to Dewey's decimal system ("dewey")
- identifiers of the libraries where a copy of the publication can be found ("localizzazioni").

3 Citation extraction results

This section expands the discussion of the pipeline and presents full-pipeline results (i.e. for extraction, parsing, look-up) on the corpus of 700 monographs on the history of Venice, with a structured reference list.

3.1 Citation parsing and extraction

The parsing and extraction module comprises two supervised models. The first one performs the following: given a stream of text likely to contain a list of references, it initially tags

every token with specific tags. A second model then parses the text again in order to attribute generic and begin-end tags at the same time. Eventually, all individuated citations for each monograph are exported to the look-up module. All the implementation has been done in Python, using the CRFSuite [23]. The set of features includes:[4]

- The token as is, the lowercase token, its position in the line, its shape and type according to a set of predefined classes (e.g. for shape: "UUDDDD" for "AD1900" meaning two uppercase characters and four digits. For classes, in this case we would have "AllUpperDigits", "InitUpper").
- Suffixes and Prefixes from 1 to 4 characters included.
- A set of indicator features: for example, if the token contains two digits, if four digits, if it could be an abbreviation or contain Roman numbers, etc.
- The reference style category (unique combination of class and family).
- The specific token tag, only for model 2.

A validation set of 25% of annotated citations on which all final evaluations are based, has been initially put aside and never used for training. On the remaining 75%, a set of cross-validation experiments have been performed, in order to find the best parameters and combinations of training approaches. These changes have been tested: (1) reducing the number of features by removing the token and its lowercase version, plus all suffixes and prefixes; (2) removing citations to primary sources; (3) training separate models for each of the 6 families of reference styles; (4) splitting the training data in different sizes (sets of citations to parse contiguously); and (5) changing the order of the parsing tasks. Test 2 yielded positive improvements and was kept, test 4 gave a window of slices of text containing 5 citations as optimal for splitting annotated pages for training. Tests 1 and 5 were negative as they slightly reduced performance. Eventually, test 3 produced over-fitted models, or models that were not able to generalize properly on test data, probably due of the lack of balanced annotated data for every family. Nevertheless, removing the reference style category as a feature in the models, or using just families, lead to a slight downgrade of performance too. These details of the ablation analysis are omitted for brevity.

Once the tasks were defined, the best configuration of CRF parameters was detected. Using a quasi-Newton gradient descent method (L-BFGS), there are two main parameters: c1 for L1 and c2 for L2 regularizations, respectively. Good parameters were found to be:

- Model 1, c1: 0.0289; c2: 0.0546.
- Model 2, c1: 1.53; c2: 0.002.

[4] The full list of features is available upon request.

Table 1 Extraction results for task 1: parsing

Class	Precision	Recall	F1-score	Support
0) null	0.679	0.553	0.609	9033
1) pagination	0.900	0.905	0.902	811
2) publisher	0.780	0.688	0.731	1029
3) author	0.847	0.862	0.855	5464
4) title	0.839	0.911	0.873	18,834
5) pub. nbr-yr	0.772	0.835	0.802	466
6) pub. place	0.860	0.873	0.867	1729
7) year	0.882	0.880	0.881	1744
Avg/total	0.805	0.812	0.806	39,110

Table 2 Extraction results for task 2: extraction and classification

Class	Precision	Recall	F1-score	Support
0) out	0.936	0.958	0.947	4815
1) begin monograph	0.846	0.903	0.873	1349
2) in monograph	0.841	0.911	0.874	15,683
3) end monograph	0.862	0.894	0.878	1352
4) begin contribution	0.812	0.759	0.785	523
5) in contribution	0.892	0.802	0.845	10,930
6) end contribution	0.823	0.820	0.822	523
7) begin abbreviated	0.418	0.266	0.325	192
8) in abbreviated	0.418	0.362	0.388	1963
9) end abbreviated	0.325	0.193	0.242	192
Avg/total	0.841	0.845	0.842	37,522

Intuitively, model 2 benefits from sparse regularization much more than model 1. The result is a set of 181,699 citations, 8632 of which were part of the golden set and 173,067 were newly parsed and extracted.

A fivefold validation over the whole dataset gives a flat and weighted F1-score of 0.77 and 0.85 for task 1 and 2 respectively, while validation scores on the validation set are summarized in Tables 1 and 2, which should be read along with confusion matrices in Fig. 5.

For model 1, the main source of errors are null tokens (without tag). Several initially present tags have been removed due to them being either under-represented or too varied to be properly captured by the model. This explains the parser's difficulty in properly fitting the null tag, which ended-up being a repository of variability. Model 2 instead behaves consistently with the availability of data, meaning that abbreviated references are not as well captured as monographs and contributions. It is nevertheless important to note that begin tags mostly get mistaken for other begin tags, and the same for inside and end tags, all of which are minor errors.

Separate models trained to detect citations to primary sources, whose description and evaluation is beyond the scope of this article, were preliminarily used to avoid mixing citations to primary and secondary sources.

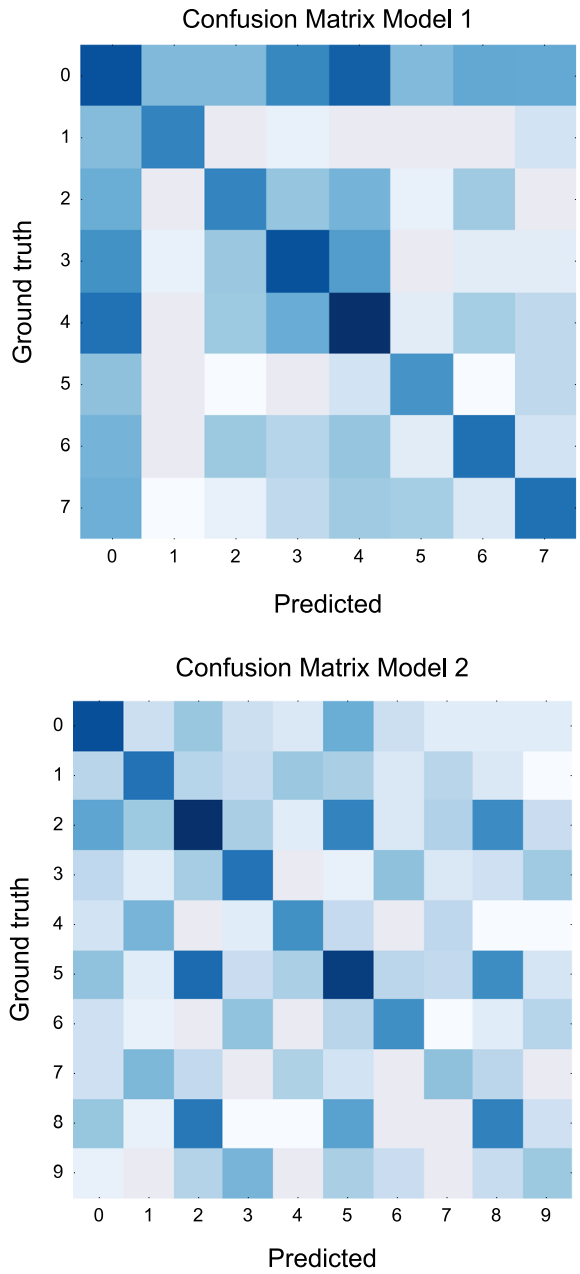

Fig. 5 Confusion matrices for models 1 and 2. Identifiers are the same as in Tables 1 and 2, respectively, for models 1 and 2. A darker square means more matches within the bin. For example, in matrix 1, the *null* class is the most problematic as both other classes are wrongly classified as *null*, and vice-versa *null* tokens are mistaken to be of another class. In matrix 2, errors are consistent with expectations, e.g. *in monograph* mistaken for *in contribution* or *in abbreviated*

3.2 Catalog look-up

3.2.1 Method

The catalog look-up attempts to match the metadata fields of the extracted citation against the bibliographic records contained in the SBN catalog.

The look-up is performed in two steps: (1) retrieval of disambiguation candidates and (2) comparison of the input citation with each candidate.

The first step consists of retrieving a list of possible disambiguation candidates by searching through the catalog. By doing so, the search space is reduced in order to avoid comparing the input citation with a high number of totally unrelated catalog records. Several experiments allowed to find the ideal settings to strike a good balance between efficiency and accuracy (and especially recall). The best results are obtained when searching for candidate records whose title contains the first two (content) words of the title in the citation, and then pruning the list, sorted by title similarity, to return a maximum of 2000 candidates. In fact, using the full cited title to search the catalog leads to an extremely low level of recall.

The second step of the look-up consists of comparing each retrieved candidate with the input citation in order to compute a global similarity score (ranging from 0 to 1): only the candidates with score above a certain threshold are then considered as matches and returned. The metadata fields that are considered for comparison are: author, title, publisher, year and place of publication. For each of these fields, the similarity between candidate and input citation is calculated using fuzzy matching algorithms. Before being compared, each field is pre-processed in order to recompose hyphenated words and remove punctuation signs as well as stopwords.

3.2.2 Evaluation

A gold standard corpus consisting of 2000 randomly sampled citations is used in order to evaluate the accuracy of the look-up: 500 manually annotated citations, and the remaining 1500 automatically extracted using the approach described in the previous section. Two annotators then disambiguate each citation by assigning the corresponding bibliographic identifier (BID) in the SBN catalog.[5]

From the original set of 2000, four cases were removed since the bibliographic record which corresponds to the BID assigned by the annotators was not found within our dump of the SBN catalog. Eighty-three other citations that did not have a title were also removed, as this is the minimum requirement for a citation to be looked up. The final set of 1903 citations is used for the evaluation.

Before discussing more in detail the evaluation results, the question of what constitutes a *correct match* (for evaluation purposes) needs to be briefly addressed as it is less trivial than it may seem. Provided that the extracted citation includes the year of publication, it is possible to try to match the citation

[5] The annotators use the online search interface of the library catalog to retrieve the BIDs. The search interface is available at http://www.sbn.it/opacsbn/opac/iccu/free.jsp.

Table 3 Results of the evaluation of the citation look-up module

Class	Precision	Recall	F1-score	Support
Man. annotated	0.820	0.911	0.863	493
Autom. extracted	0.771	0.901	0.831	1410
Total	0.784	0.904	0.840	1903

with the record of the very same edition in the SBN catalog. As a result, a citation linked to a different edition than the referred one is considered as a wrong match. Reprints of a publication constitute the only exceptions to this rule (i.e. the BIDs of both the original version and the reprint are considered as a correct match). Moreover, in those cases where the citation points to a work in several volumes, the correct BID is the one of the record that describes the work as a whole, not the BIDs of individual volumes .

The results of the evaluation are presented in Table 3. Among the disambiguation candidates returned by the look-up, only those with a global similarity score above 0.4 are retained, as this threshold value proved to yield the best results. The candidate with the highest score is then chosen as the predicted match. For each citation, the manually assigned identifier—that may or may not be there—is compared with the identifier returned by the look-up.

3.2.3 Discussion

As shown in Table 3, the accuracy of the look-up does not vary substantially between the citations that were manually annotated and those automatically extracted. While the level of recall of the look-up is overall satisfactory (0.904), its precision (0.784) could be improved.

A possible improvement of the overall accuracy concerns the matching of author names. Currently, the similarity between author names is computed by comparing the author names as they appear in the reference (mostly abbreviated) with the author field of the catalog records (where the names are always expanded). In order to boost those records that are more likely to be correct, it is possible to introduce an intermediate step where the abbreviated name form is looked up in a database of author names, and eventually replaced with the expanded form. In other words, taking into account first names for the comparison is expected to increase the capability of the look-up module of assigning a higher rank to the correct disambiguation candidate.

Finally, the evaluation results need to be interpreted in light of the following considerations. Firstly, although only the candidate with highest similarity returned by the look-up is considered, it must be noted that in a number of cases—approximately one-third of the false positives—the correct match is contained in the list of candidates with similarity score above the pruning threshold ($n = 0.4$). This detail is important given that the results of the look-up module will be eventually fed back into a correction/editing interface. Secondly, approximately 10% of the citations in our groundtruth refer to publications not contained in the SBN catalog, such as publications in foreign languages or early printed books.

4 Using the references of references to find the most relevant literature in the field

This section investigates two main characteristics of the given corpus: its *cohesiveness*, or the proportion of citations extracted from the corpus which refer to the corpus itself (endogenous citations), and its *connectedness*, or the dimension of the giant component in the co-citation network (considered both on the endogenous and exogenous citations, or citations to monographs outside of the corpus). Eventually, the section considers the feasibility of covering most of the relevant literature in a given domain, using reference monographs, and the presence of a set of core works in the spanned literature.

4.1 Cohesiveness and connectedness of the selected corpus

The proposed approach rests on an important assumption, which needs testing: reference monographs in the corpus act as a hub pointing to most, or at least a considerable part of the relevant literature within the domain. If the selected corpus is not spanning sufficiently outside of itself, at the same time being well-connected internally (presenting a dominant giant component), then it might not be effective in connecting different research areas within the same domain of study. This assumption is not directly supported by (non-abundant) previous work, which in general highlights great variability in citation patterns among different disciplines in the humanities. Co-citation structures by domain and by research themes can both be found (see e.g. Ahlgren et al. [1], Weingart [31]), but the proportion of citations to monographs and journal articles is quite varied in different domains [12]. The agreement over the absence of a set of core works in the humanities would also not encourage this method (see e.g. Nolen and Richardson [22]). To be sure, the literature on the topic is still underdeveloped, and thus the assumption should be tested empirically over larger datasets and across different domains, as this article attempts for historiography.

The look-up module was used in order to look citations up within the corpus itself. The adapted look-up module has been manually evaluated on a small set of 500 extracted citations, resulting in a precision score of nearly 1.00 and a recall score above 0.95. High-quality results are made possible by the fact that the catalog records of the monographs in the

Table 4 Citation span of the elected corpus: most citations are made to the outside

Dataset	Proportion	Matched (extracted) refs.
Consultation, cat. 1	0.0802	1861 (21,337)
Without cat. 1	0.0669	5398 (75,270)
All set	0.0699	7259 (96,607)

corpus have been adapted to the task. As a reminder, the extracted citations of 700 (37%) reference monographs (of which 183, or 9.7%, are in consultation, cat. 1) have been matched against the whole corpus of 1904 (100%) selected monographs. For this purpose, only the 96,607 extracted citations to monographs in full details are considered, over the total of 181,699.

The *cohesiveness* of the corpus is defined as the proportion of extracted citations which refer to monographs inside of the corpus itself, over the total number of citations. Results are summarized in Table 4. Overall, only 7% of the extracted citations are made to monographs already within the corpus, slightly more for monographs on reference shelves (8%).

The *connectedness* of the corpus is equivalent to the proportion of monographs from the corpus which are in the giant component of the co-citation network resulting from the look-up procedure. A more flexible definition would consider all the monographs which are in the k-weighted giant component, or the giant component made of all the edges with weight equal or greater than k. The weight of an edge is given by the number of times two monographs are cited together (that is, appear in the same reference lists). For the dataset under consideration, the giant component is well individuated and comprises circa 59% of the corpus. The coverage drops less than proportionally to 32.5% using only the 183 monographs in consultation.

These results suggest that most of the selected corpus could be useful as a source of citations pointing to the relevant literature in the domain, given that most of these citations point outside the corpus. Furthermore, a substantial part of the corpus of reference monographs is tightly connected into the giant component of the co-citation network, suggesting the viability to span the domain.

4.2 Spanning the domain

Connectedness can also be investigated at the level of the full set of extracted citations, after look-up. The resulting co-citation network comprises all the 37,626 monographs which scored sufficiently during look-up, or that are citing other monographs, and 6,030,398 edges among them. This co-citation network is not filtered by a minimum weight of edges, where the weight corresponds to the number of times two monographs have been cited together. In this network, the giant component comprises 37,359, or 99.3% of the nodes. Yet, all the monographs out of the giant component are simply part of the 700 citing monographs which are never cited, and thus the giant component effectively spans all cited monographs.

The size of the co-citation network produced using a seed of just 700 monographs suggests at the same time that it is possible to span a wide range of works using reference monographs and that it might be difficult to find a set of core works. Yet, a simple look at the directed citation network connecting citing with cited works highlights a strong concentration of citations. Note that this network is not bipartite, as a citing work (part of the 700) could also be cited in turn. This network naturally contains the same number of nodes as the co-citation network (37,626), but fewer edges (71,650, from the extracted 96,607). Every edge in this setting is a direct citation from a monograph to another monograph. The discrepancy from the extracted citations to the disambiguated citations is due to filtering out low confidence disambiguations and errors such as self-citations or multiple citations to the same work.

Two facts are worth noting:

1. First of all, the distribution of in-degrees is highly skewed. Figure 6 plots this distribution both globally and the reference monographs only, highlighting the skewness in log scale as well. In-degrees, in this settings, are the number of individual citations to a given work. A total of 243 works (or the 0.6% of the total) are cited 20 or more times, and 27,109 works (72% of the total) are cited only once.
2. The same applies to the set of the 700 citing reference monographs, which presents a division between a small group of works which are highly cited, and a larger set of barely, if ever cited works. More specifically, 37 monographs which received 20 or more citations (2% of the 1904 reference monographs), and providing a proportional 6.2% of given citations (4429/71,650), receive 33% of citations given to reference monographs (1280/3933). Notably, these 37 monographs are not more likely to be stored in easily accessible reference shelves (27% of them are in cat. 1, versus a proportion of 26% for the whole set of 700 parsed monographs). At the same time, 1375 reference monographs are never cited (72% of the total).

We might conclude that the domain of the history of Venice is ultimately difficult to bound "bibliometrically", as the wide number of works cited from a relatively small seed suggests. Yet the field exists, as almost all cited monographs end up in the giant co-citation component. At the same time, skewed citations patterns emerge, indicating the possible existence of a core set of works in the domain, whose investigation will demand further analysis. Being a reference monograph indeed entails being more likely to be part of the most cited

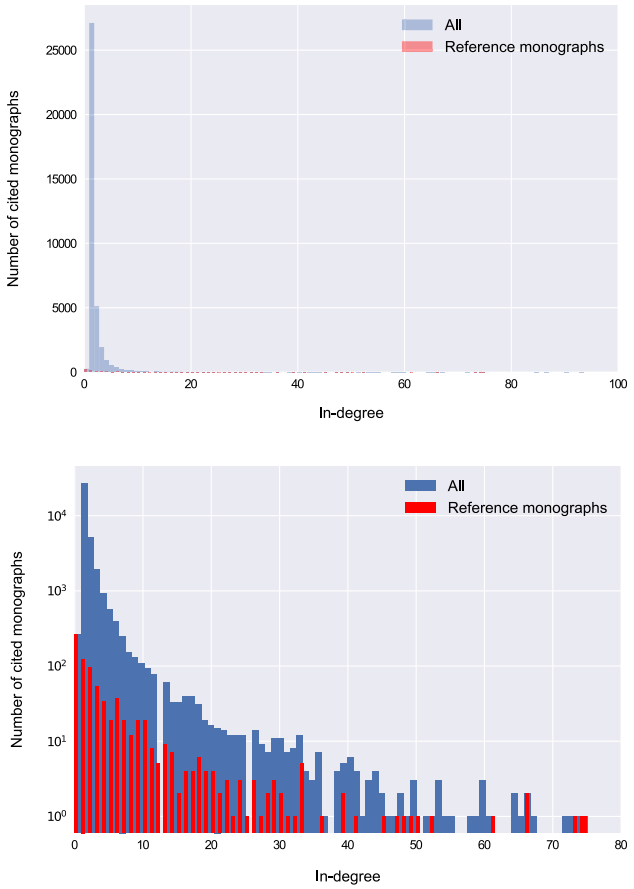

Fig. 6 Distribution of in-degrees for the global extracted corpus, including reference monographs (*blue/dark grey*) and only for reference monographs (*green/light grey*). The second plot is a zoom-in to the first part of the distribution, with a log scale on the y axis (color figure online)

group of works, but this chance remains very low (2.5% vs 0.5%). At the same time, more than 70% of reference monographs used as initial seed are never cited. To be sure, part of the reference monographs might be useful to the users of the library besides their importance in citation patterns. More problematic is the absence of most of the individuated highly cited works from the set of reference works, a discovery which could improve the identification of reference works by libraries in the future.

These results support the proposed approach, since they confirm that it is indeed possible to span a large portion of the literature in a domain to some extent and given a limited seed. In so doing, patterns of citations emerge, highlighting works that have been considered more important or durable within the community and were not, to a large extent, part of the initial reference works. No more than 2 iterations of the approach sketched in Fig. 1 should be necessary in order for the library to individuate most of the core works of a given domain, in order to integrate them in the reference works group, and for the citation network to span most of the literature of interest in the domain. Existing literature is consistent in picturing the humanities as a fragmented and holistic set of disciplines, as indeed has been shown here for the historiography on Venice. At the same time, provided sufficiently focused, but certainly not big citation data, important citation patterns do emerge, helping in identifying a core set of works and spanning a relevant amount of literature in the domain. In principle, albeit necessitating further study, the proposed approach should therefore be amenable for use in all domains of the humanities.

5 Conclusions

This article proposes to use reference monographs in specific domains in the humanities, in order to enrich library catalogs with citation data and improve the retrieval of the most relevant literature in so doing. A distinction is introduced between *reference monographs* and *core works*. The former is a set of monographs deemed of importance within a domain of study, which can be retrieved using library resources such as classification and shelving strategies; the latter is a group of works (usually monographs as well) which are cited by reference monographs more frequently than most of the rest of the cited literature. The main goal of the proposed approach is to allow users to rapidly find the most relevant publications on a topic of interest within the domain (the core works), and help libraries integrate their collection of reference monographs. The contribution in this respect is twofold.

Firstly, the article presents a robust pipeline for the individuation of reference monographs in a library, and the extraction and disambiguation of citations contained within it. The pipeline is evaluated on a corpus of monographs from the domain of the history of Venice. This system constitutes the first necessary step towards the envisaged enrichment of library catalogs, and effectively allows for enriching the library catalog with relations based on citations. The acquisition of citation data in the humanities is costly, and thus the proposed method based on a focused selection of reference works provides an effective trade-off.

Secondly, the article investigates the citation structure of the same corpus, and the literature it spans, in order to assess how effective it may be in serving as a hub to access the literature of the domain, and individuate its core works. Only 7% of the citations made from the reference monographs of the history of Venice are to monographs already within the corpus, suggesting that a wide span over the literature might be achieved from a limited set of selected monographs. At the same time, a small set of works, both within and outside of the reference corpus, appears to be cited much more frequently than average, suggesting the presence of a core set of works within the domain, and of a vast number of rarely cited works. Lastly, the collection of reference monographs is found to be more likely to contain core works, but only marginally

so, suggesting that the proposed approach could positively inform an improvement in the access policy of the library.

6 Data availability

The complete dataset resulting from this work is made publicly available under an MIT licence [24]. The dataset comes with a Python notebook replicating the figures of the article, as well as an export of the directed network of citations.

Acknowledgements We thank Martina Babetto and Silvia Ferronato for the digitization and annotation of the dataset. The Library of the Ca' Foscari University of Venice willingly collaborated with bibliographical resources and logistics support. The Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU) shared its catalog metadata with us. We thank both for their support. Finally, we thank our anonymous reviewers for their helpful comments. This project is funded by the Swiss National Fund under Division II, project number 205121_159961. Colavizza also benefits from a separate Swiss National Fund grant, number P1ELP2_168489.

References

1. Ahlgren, P., Pagin, P., Persson, O., Svedberg, M.: Bibliometric analysis of two subdomains in philosophy: free will and sorites. Scientometrics **103**, 47–73 (2015)
2. Ardanuy, J.: Sixty years of citation analysis studies in the humanities (1951–2010). J. Am. Soc. Inf. Sci. Technol. **64**(8), 1751–1755 (2013)
3. Barrett, A.: The information-seeking habits of graduate student researchers in the humanities. J. Acad. Librariansh. **31**(4), 324–331 (2005)
4. Buchanan, G., Cunningham, S.J., Blandford, A., Rimmer, J., Warwick, C.: Information seeking by humanities scholars. In: International Conference on Theory and Practice of Digital Libraries. Springer, pp. 218–229 (2005)
5. Councill, I.G., Giles, C.L., Kan, M.Y.: ParsCit: an open-source CRF Reference String Parsing Package. In: LREC (2008)
6. Hammarfelt, B.: Interdisciplinarity and the intellectual base of literature studies: citation analysis of highly cited monographs. Scientometrics **86**(3), 705–725 (2011)
7. Hammarfelt, B.: Using altmetrics for assessing research impact in the humanities. Scientometrics **101**(2), 1419–1430 (2014)
8. Heinzkill, R.: Characteristics of references in selected scholarly English literary journals. Libr. Q. **50**(3), 352–365 (1980)
9. Heinzkill, R.: References in scholarly English and American literary journals thirty years later: a citation study. Coll. Res. Libr. **68**(2), 141–154 (2007)
10. Kellsey, C., Knievel, J.: Overlap between humanities faculty citation and library monograph collections, 2004–2009. Coll. Res. Libr. **73**(6), 569–583 (2012)
11. Kim, Y.M., Bellot, P., Faath, E., Dacos, M.: Automatic annotation of bibliographical references in digital humanities books, articles and blogs. In: Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing, ACM, pp. 41–48 (2011)
12. Knievel, J.E., Kellsey, C.: Citation analysis for collection development: a comparative study of eight humanities fields. Libr. Q. Inf. Community Policy **75**(2), 142–168 (2005)
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML, pp. 282–289 (2001)
14. LindholmRomantschuk, Y., Warner, J.: The role of monographs in scholarly communication: an empirical study of philosophy, sociology and economics. J. Doc. **52**(4), 389–404 (1996)
15. Linmans, A.J.M.: Why with bibliometrics the Humanities does not need to be the weakest link: indicators for research evaluation based on citations, library holdings, and productivity measures. Scientometrics **83**(2), 337–354 (2009)
16. Lopez, P.: GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Research and Advanced Technology for Digital Libraries, Springer, pp. 473–474 (2009)
17. Marchi, M.D., Lorenzetti, E.: Measuring the impact of scholarly journals in the humanities field. Scientometrics **106**(1), 253–261 (2015)
18. McCain, K.W.: Citation patterns in the history of technology. Libr. Inf. Sci. Res. **9**, 41–59 (1987)
19. Mingers, J., Leydesdorff, L.: A review of theory and practice in scientometrics. Eur. J. Oper. Res. **246**(1), 1–19 (2015)
20. Mongeon, P., Paul-Hus, A.: The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics **106**(1), 213–228 (2015)
21. Nederhof, A.J.: Bibliometric monitoring of research performance in the social sciences and the humanities: a review. Scientometrics **66**(1), 81–100 (2006)
22. Nolen, D.S., Richardson, H.A.: The search for landmark works in English literary studies: a citation analysis. J. Acad. Libr. **42**(4), 453–458 (2016)
23. Okazaki N (2007) CRFsuite: a fast implementation of Conditional Random Fields (CRFs). www.chokkan.org/software/crfsuite
24. Romanello, M., Colavizza, G.: dhlab-epfl/LinkedBooks Monographs: LinkedBooksMonographs (version 1.0) (2017). doi:10.5281/zenodo.266889
25. SBN, G.: Reicat—GuidaSBN (2016). http://norme.iccu.sbn.it/index.php?title=Reicat&oldid=3034. Last Accessed 9 Jan 2017
26. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based Tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, EACL '12, pp. 102–107 (2012)
27. Sula, C.A., Miller, M.: Citations, contexts, and humanistic discourse: toward automatic extraction and classification. Lit. Linguist. Comput. **29**(3), 452–464 (2014)
28. Thelwall, M., Delgado, M.M.: Arts and humanities research evaluation: no metrics please just data. J. Doc. **71**(4), 817–833 (2015)
29. Thompson, J.W.: The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. Libri **52**(3), 121–136 (2002)
30. Waltman, L.: A review of the literature on citation impact indicators. J. Informetr. **10**(2), 365–391 (2016)
31. Weingart, S.B.: Finding the history and philosophy of science. Erkenntnis **80**(1), 201–213 (2015)
32. Wiberley, Jr S.E.: Humanities literatures and their users. In: Encyclopedia of Library and Information Sciences, pp. 2197–2204 (2010)
33. Williams, P., Stevenson, I., Nicholas, D., Watkinson, A., Rowlands, I.: The role and future of the monograph in arts and humanities research. Aslib Proc. **61**(1), 67–82 (2009)
34. Wu, J., Williams, K., Chen, H.H., Khabsa, M., Caragea, C., Ororbia, A., Jordan, D., Giles, C.L.: Citeseerx: Ai in a digital library search engine. In: Innovative Applications of AI Conference (2014)
35. Zordan, G.: Repertorio di storiografia veneziana: testi e studi. Il Poligrafo, Padova (1998)
36. Zuccala, A., Guns, R., Cornacchia, R., Bod, R.: Can we rank scholarly book publishers? A bibliometric experiment with the field of history. J. Assoc. Inf. Sci. Technol. **66**(7), 1333–1347 (2014)