doi: http://dx.doi.org/10.23925/1980-7651.2018v21;p41-55

History of science: the problem of cataloging, knowledge indexing and information retrieval in the digital space

Carla Bromberg[•]

Abstract

Text-based and multimedia documents in and for history of science are displayed in libraries and ought to be organized to make knowledge and information on history of science accessible. The traditional approach to the organization of and access to knowledge and information was expressed by classification schemes primarily influenced by philosophical traditions, and then mostly based on the literary warrant principle. Within this context, the scholarly and scientific literature was seen as representing facts about knowledge and structures of knowledge. Cataloging and classification were essential to provide users access to information. Cataloging elements consist of bibliographic description, subject analysis and classification. Currently, within the digital environment, not only text-based documents, but documents of all sorts must be included, classified and organized in order to be browsed. In this paper I call the attention to some of the improvements and challenges that currently affect the relationship between catalogs, knowledge organization, classification and information retrieval. As an example I mention the catalog-interface that is being developed for the digital library of CESIMA-Brazil.

Keywords

History of Science; Document classification; Knowledge Classification; Catalogs; Bibliography

[•] Center Simão Mathias of Studies in History of Science (CESIMA), Pontifical Catholic University of São Paulo, Brazil. ✉ carlabromberg@gmail.com. The present paper derives from a presentation at symposium Doing History of Science in a Digital, Global, Networked Community: Tools and Services Linking Scholars and Scholarship, 25th International Congress of History of Science and Technology, Rio de Janeiro 22-29 July 2017.

Introduction

In 2013, the Focus section of volume 104 of journal *Isis* was devoted to discussing classification in the history of science. The essays dealt with the history of classification schemes, archival organization and the intersection of human and machine classification systems.[1] Later one, volume 107 was devoted to a discussion of archives for history of science, also including the process of digitization and information retrieval.[2] In these volumes, bibliographical classification, knowledge classification and archival topics appeared as relevant issues for the process of digitization and the establishment of digital libraries.

Text-based and multimedia documents in and for history of science are displayed in libraries and ought to be organized to make knowledge and information on history of science accessible. As is known, the traditional approach to the organization and access to knowledge and information was expressed by classification schemes primarily influenced by philosophical traditions, and then mostly based on the principle of literary warrant.[3] Mostly dealing with text-based documents, different types of bibliographies were developed, expressed by classification systems used in libraries and databases, including the Dewey Decimal Classification (DDC), Library of Congress Classification (LCC), Universal Decimal Classification (UDC) and the facet analysis approach. Within this context, the scholarly and scientific literature was seen as representing facts about knowledge and structures of knowledge.[4]

Bibliographies constitute the core of catalogs. And catalogs are the primary tool for organizing libraries,[5] since they afford models of organization of library resources that enhance accessibility and retrieval.[6]

In the beginning of the nineteenth century, the librarians in charge of cataloging and displaying documents were also responsible for describing the content of documents. However, by the end of the nineteenth century, due to the increasing number of documents, the responsibility for abstracting and indexing shifted from librarians to another group of experts.[7] According to Jesse Shera, one of the reasons ”for the breakdown of the libraries’

[1] Stephen P. Weldon, “Introduction,” *Isis* 104 (2013): 537-9.

[2] Focus: the History of Archives and the History of Science, *Isis* 107, no. 1 (2016): 74-120.

[3] The concept of literary warrant was formulated by the English librarian E. Wyndham Hulme (1859-1954) in a book published between 1911-12.

[4] Ana Maria Alfonso-Goldfarb, Silvia Waisse & Márcia H.M. Ferraz, “From Shelves to Cyberspace: Organization of Knowledge and the Complex Identity of History of Science,” *Isis* 104 (2013): 551-60.

[5] In the past, catalogues were primarily tools for organizing the shelving of physical items according to classification schemes; Manfred Kochen, *Principles of Information Retrieval* (Los Angeles: John Wiley and Sons, 1974).

[6] The pioneer Charles A. Cutter’s (1837-1903) greatest accomplishments were in the realm of libraries systematization and tools for access to recorded knowledge; Francis L. Miksa, “Charles Ammi Cutter: Nineteenth-century Systematizer of Libraries”(PhD Dissertation, The University of Chicago, 1974), on xv.

[7] Fred A. Tate, & James L. Wood, “Libraries and Abstracting and Indexing Services: A Study in Interdependency,” *Library Trends* 16, no. 3 (1968): 353-73, on 353.

machinery for providing content access [lays] in the rapid development of the journal as an important form of publication."[8]

It was the Abstracting and Indexing Services (A&I) which provided libraries and their users content analysis of serial, dissertation and patent literature. The faster dissemination of content analysis by A&I services forced libraries to speed up their processing of documents. Not only regarding the physical transfer of documents—either from libraries to other libraries or to other institutions, but specially by adopting the latest available electronic data processing procedures and equipment.[9] Machine readable databases were being developed, and the Library of Congress (USA) developed the MARC (Machine Readable Catalog) Project, which came to be the most significant development in the area.[10] Since then, libraries work with and rely on A&I services produced by mechanized databases.[11] Within this context, online catalogs developed since the 1960s—and especially starting in the 1980s—they represent to library users the most visible and widely used computer interface for information retrieval (IR).[12]

Moving ahead thirty years, online catalogs (OPACs)[13] coexist in the digital space with other extremely high technical tools which: (1) can provide resources beyond libraries' collections and enable any type of document search; (2) provide query formulation by browsing and natural language, challenging the traditional catalog search approach—from query to browsing and Boolean operators; and finally (3) a digital environment that also provides full-text access to innumerous documents.[14]

Also the relationship between library and users is being challenged. In the Web 2.0 world there is an implicit demand for building an environment likely to encourage users' participation and collaboration.[15] In the specific case of history of science, Stephen Weldon, the editor of *Isis* bibliography, proposed a new tool that he called Isis Document Indexing Platform or "Bibliography 2.0," which draws on a networked-information-and

[8] Jesse H. Shera, *Documentation and the Organization of Knowledge* (Handen [CT]: Archon Books, 1966), 26.

[9] Tate & Wood, 368.

[10] It might represent an example of A&I effort performed by a library.

[11] Tate & Wood, 369.

[12] Sharon E. Clark & William H. Mischo, "Online Public Access catalog Retrieval Structures and Techniques: With Reference to Recent Developments in the United States and Great Britain," *Proceedings of the IATUL Conferences*, 1991, paper 20, 113-120, http://docs.lib.purdue.edu/iatul/1991/papers/20.

[13] OPAC was defined by Library of Congress in 1980 as an access tool and resource guide to the collections of a library or libraries, which contains interrelated sets of bibliographic data in a machine-readable form and which can be searched interactively on a terminal by users; Emily G. Fayen, *The Online Catalog: Improving Public Access to Library Materials* (New York: Knowledge Industry Publications, 1983).

[14] Tanja Mercun, & Maja Žumer, "New Generation of Catalogues for the New Generation of Users: A Comparison of Six Library Catalogues," *Program: Electronic Library & Information Systems* 42, no. 3 (July 2008): 243–61. Eric L. Morgan, "A 'Next -Generation' Library Catalog— Executive Summary (Part #1 of 5)," online posting, July 7, 2006, LITA Blog: Library Information Technology Association, http:// litablog.org/2006/07/07/a-next-generation-library-catalog. Marshall Breeding, Introduction to "Next Generation Library Catalogs," *Library Technology Reports* 43, no. 4 (July/Aug. 2007): 5–14.

[15] Marshall Breeding, "Next Generation Library Catalogs," *Library Technology Reports* 43, no. 4 (2007):5-42. The application of Web 2.0 principles on libraries has been framed as library 2.0. See Jack Maness, "Library 2.0 Theory: Web 2.0 and Its Implications for Libraries," *Webology* 3, no. 2 (2006), article 25. http://www.webology.org/2006/v3n2/a25.html.

communication structure. According to Weldon, by combining the interlinked data in the bibliography with communication and information networks in open Internet, it is possible to create a social network of people linked to each other by way of their publications. This interactive bibliography is becoming a social tool with the objects of people's scholarship at its core.[16]

While libraries are trying to improve their relationship with users through social media and tool improvement, one should not forget that librarians remain in charge of providing information on library collections, and that the structuring of catalogs and the definition of their nature remains a core function of librarianship.[17] Yet, the library catalog has been reformulated following the ever increasing changes in technology, and digital libraries find themselves in an environment undergoing constant change. Taking this into consideration, I call the attention in the present paper to some of the improvements and challenges that currently affect the relationship between catalogs, knowledge organization, classification and information retrieval. As an example I mention the catalog-interface that is being developed for the digital library of Center Simão Mathias of Studies in History of Science (CESIMA)-Brazil.

Generations of catalogs

As mentioned above, cataloging and classification are essential processes in libraries. The earliest online catalogs were very similar in structure and function to the card catalog system. Thus they were described as "automated circulation database query systems masqueraded as public access library catalogs."[18]According to Charles R. Hildreth, this first generation catalogs lacked subject access,[19] and keyword access to titles and subject headings. They only afforded one single mode of interaction with the system and little of online user assistance.

The second generation of catalogs was characterized by a connection between the library catalog and conventional online information retrieval; Hildreth described it as a bibliographic information retrieval system. In these catalogs, search was improved by means of conventional IR keywords and Boolean search approaches. However, according to Hildreth, this second generation of catalogs could not still perform a number of important tasks: they could not lead users from the found information to other related data. Then, to

[16] Stephen Weldon, "The *Isis* Bibliography from Its Origins to the Present Day: One Hundred Years of Evolution of a Classification System," *Circumscribere* 6 (2009): 26–46.

[17] On those who believe that also users can participate in the creation of a library resources see the literature regarding the term 'presumption.'

[18]Charles R. Hildreth, "Beyond Boolean: Designing the Next Generation of Online Catalogs," *Library Trends* 35, no. 4 (1987): 647-67, on 650. In this article I follow Hildreth's classification of three generations of catalogs. It is important to notice that the three different generations of catalogues might coexist. See also Iris Xie, *Interactive Information Retrieval in Digital Environments* (Hershey, [PA]: IGI Publishing, 2008), 29-52.

[19] "Subject access means providing information on what publications are about [...] it enables people to see which and how many publications that exist in a specific field of knowledge and which topics are contained in a particular national publishing output or a particular library collection." *IFLA Working Group on Guidelines for Subject Access in National Bibliographies*, Leipzig, 2011, 1-95, on 6.

optimize retrieval results in subject search more than one approach had to be employed in the overall search strategy. Finally, the system assumed that users knew what they were looking for, and that they could describe it in the language of the catalog database being searched.

In an article from 1988,[20] Hildreth delineated his idea of what a third generation of catalog should be. He was concerned with functional improvements that could make catalogs more user friendly - like use of natural language search and browse devices, but also with expanded coverage and scope and different aids, such as spelling corrections, synonyms and, automatic term conversion.

As is known, the functions outlined by Hildreth were not implemented in OPAC (online catalogs) systems, but were incorporated into Web search engines and online databases. By the early 1990s, the Internet became international in scope. Although still primarily used by academicians and businesspeople, the Internet grew, and by the end of the 1990s the Web was invented.[21] Regarding libraries, the focus on Internet/computer devices and technology aimed at developing linkages between library local systems. It was the most relevant topic for library automation, as with their systems linked, libraries could loan materials, cooperate in the development of collections and share resources.

Nevertheless, by the end of the 1990s the demands on library resources were changing as a function of the online access to information.[22]

Catalogs and classification

Cataloging elements consists of bibliographic description, subject analysis and classification. In library classification, documents are arranged and then sub-arranged based on disciplines and sub-disciplines, and relationships are established between documents.[23]

Catalogs are made of bibliographical records that are supposed to describe documents. Regardless of the fact that they might provide us some insights about documents (such as form, e.g., a book, or field, i.e., words used in the title) nothing in them leads us to a direct interpretation of the texts they describe. According to Birger Hjørland, the points of access offered in online catalogs (OPACs) are still author, title and subject, despite the fact that many other useful data are readily accessible. Even today, OPACs

[20] Charles R. Hildreth, C.R., "Online Library Catalogs as IR Systems: What Can We Learn from Research?" in *Future Trends in Information Science and Technology, Proceedings of the Silver Jubilee Conference of the City University's Department of Information Science,* ed. Penelope A. Yates-Mercer (London: Taylor Graham, 1988), 9-25.

[21]National Research Council, *Funding a Revolution: Government Support for Computing Research.* (Washington DC: The National Academies, 1999), 169-83, on 179.

[22] Chris Sugnet, "Standards: Where Are We Headed?," *Library Hi Tech* 4, no. 2 (1986):95-103, on 99-100.

[23] Stephen Paling, "Classification, Rhetoric, and the Classificatory Horizon," *Library Trends* 52, no. 3 (2004): 588-603, on 589-90.

mostly consist of records organized around the concept of main entry, which was relevant for card catalogs.[24]

It is important to notice that the particular choice of elements described in bibliographic records is temporally, theoretically, disciplinary and socially bounded. Let us think of the prominence of the identification of authorship.[25] It reflects a particular assertion about the ontological status of materials. Concepts such as date and place of publication and publisher were not relevant in ages prior to printing, when personal copying played a relevant role in the dissemination of texts.[26] Or let us think about the bibliographic description of manuscripts (either textual or musical) which are mainly done with the help of other professionals (codicologists, paleographers, musicologists). Manuscripts are normally described by their *incipit* or word-content (copy of the first verses) and one needs to be aware that catalogs can also be of part-holding of libraries collections. Catalogs can also describe manuscript fragments. Because of such particularities, bibliographical records are known to exhibit different degrees of accuracy (depending on the material being described, e.g. manuscripts) and depend on second tools such as elements of control (e.g., indexes) and vocabularies to fully accomplish their tasks.

Bibliographic records have another interesting characteristic. There are no hierarchical structures and no explicit relationships between statements (author's name, title, date, publisher, etc.) although there is an implicit assumption that all the information recorded on entries relates to the same thing.[27] It is important that such entries or statements (author, date, publisher, etc.) enable, especially in modern online catalogs, associative relationships between different materials and different interfaces.

In bibliographical history, bibliographical records are divided in two types of classifications: systematic bibliography and critical bibliography. The first refers to the classification of records (in the past, mostly books) according to some guiding principle, in contrast to the comparative and historical study of their markup. It is to the former classification that the catalog discussion mostly pertains.

Bibliographical lists can be understood as attempts to record particular literature pertaining to one person or compiler, or can cover all the literature on all subjects. The latter, or general bibliographies, can be exemplified in history by the works of authors such as Conrad Gesner (1516-1565), John Hartley (ca.1630-1699), Christoph Hendreich (1630-1702) and Robert Watt (1774-1819).[28] Until the twentieth century, many were the formats of

[24] "As for the means of providing subject access, LC Subject headings are still used as they have been for so long in manual catalogues," Birger Hjørland, "Arguments for 'the Bibliographical Paradigm': Some Thoughts Inspired by the New English Edition of the UDC," *Information Research* 12, no. 4 (2007) paper colis06.

[25] Michel Foucault, "What Is an Author," in *The Foucault Reader, an Introduction to Foucault's Thought*, ed. Paul Rabinow (Harmondsworth: Peregrine Books, 1984), 51-75.

[26] Michael T. Clanchy, *From Memory to Written Record: England 1066-1307* (Oxford: Blackwell, 1993); Henry J. Chaytor, *From Script to Print: An Introduction to Medieval Literature* (Cambridge: Cambridge University Press, 1945).

[27] David Shotton, "CiTO, the Citation Typing Ontology," *Journal of Biomedical Semantics* 1 (Suppl. 1): S6, 2010. http://dx.doi.org/10.1186/2041-1480-1-S1-S6.

[28] Warden Boyd Rayward, *Systematic Bibliography in England: 1850-1895* (Urbana-Champaign: University of Illinois Graduate School of Library Science, 1967), 1-54, on 3.

bibliographical lists, and many scholars devoted much attention to organize the literature that was being published and accumulated in associations, universities and academies.

Upon reflecting about the disorder of the literature on social sciences in Belgium, Paul Otlet (1868-1944) began to develop what was bibliographically needed to introduce in these sciences the order, rigor and cumulativeness he admired in the natural sciences. According to Warden B. Rayward, "[Otlet] thought to divorce the book's content from its author and his/her authorial intentions enabling to extract from books their contribution to knowledge."[29] The information could then be entered in cards, which could be arranged to reflect the subject relationships involved. Otlet envisaged a system built of separate cards which allowed ''all the manipulations of classification and continuous interfiling,'' which also demanded a classification or ''very detailed synoptic outline of knowledge'' that could be used both as basis for arranging the cards in a catalog and "for organizing collaborative work among scholars in the compilation of the catalogue."[30] In 1892, when he wrote ''Something about Bibliography,'' no such classification seemed to be available; he only purchased a copy of Melvil Dewey's Decimal Classification in 1895.

Melvil Dewey (1851-1931) introduced the Amherst shelf classification in 1876 and planned to give to books numbers indicating not their physical location in the shelving system, but their relative location in a classification scheme. It was called, therefore, a 'movable' or 'relative' shelf classification system. The classification was a hierarchical enumeration of knowledge expressed by a hierarchy of decimal numbers, and was named Dewey Decimal Classification (DDC). Otlet reworked Dewey's system with some of his colleagues at the International Office of Bibliography (OIB) in the attempt to create a large card catalog. This catalog was called *Répertoire Bibliographique Universel* (RBU) and was arranged in a classified order by a highly elaborated version of DDC that became known as *Classification Décimale Universelle* (UDC).

In turn, Charles Cutter (1837-1903), going beyond Dewey's initial scheme, turned his attention to classification as another viable approach to subject access. He developed an enumeration of knowledge classification for the arrangement of books which became a model for its comprehensiveness relative to physical libraries; documents with the same subject were shelved together.[31] The system classification designed for the arrangement of library shelves allowed choosing, from the variety of subjects found in the document, the one that best described a given document.[32]

[29] Warden Boyd Rayward, "The Origins of Information Science and the International Institute of Bibliography/International Federation for Information and Documentation (FID)," *Journal of the American Society for Information Science* 48, no. 4 (1997): 289-300, on 291.

[30] Paul Otlet, "Something about Bibliography (1892)," in *The International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*, ed. & trans. Warden B. Rayward (Amsterdam: Elsevier, 1990), 18-9.

[31] Thomas M. Dousa, "Categories in Charles A. Cutter's Systems of Subject Cataloging and Bibliographical Classification," in *Proceedings from North American Symposium on Knowledge Organization*, University of California, Los Angeles, 2015, vol. 5, 1-21.

[32] Different types of classification schemes exist and have different approaches toward vocabulary and subject headings. Since the publication of DDC in 1876, several general classification systems were developed. S.R. Ranganathan divided these systems into five species: enumerative, almost enumerative, almost-faceted, rigidly-faceted, and freely faceted. This categorization of species depends on the extent of enumeration or provisions for

Although a succession of cataloging codes appeared after Cutter's time, some codes are acknowledged to have been less successful than others.[33] However, all of them aimed at arranging physical documents into subject groups using pre-coordinated classifying schemes which relied on hierarchies, categories or facets.[34]

Catalogs and metadata

Cataloging is also how librarians create data about data, or metadata. Despite the lack of consensus in the definition of this term, within library studies metadata is explained as the description of all types of information resources, either in electronic or print format.[35] Metadata is currently used to indicate all the information relating to a document as a record that makes up the various types of bibliographic repertoires. Considering activity in time, metadata ordering refers to the order assigned to bibliographic references and their access index, catalog cards according to their headings, access points to databases of any type and terms that make up the hierarchies of online resources.[36]

Metadata is fundamentally descriptive and as such it has to show how one can find a resource (discovery), how a resource can be distinguished from similar resources (identification), how to select what is needed (selection), how to obtain a copy of the resource (acquisition) and how to bring together all the versions of a work (collocation). It also includes evaluation (which can be subjective or expressed by content ratings), linkage and usability.

One of the advances in the digital world was the connection librarians established between library catalogs and digital repositories mostly by transforming metadata in

synthesis with the help of main schedules, special and common isolates. LCC and the early editions of the DDC are examples of enumerative classification systems. The Colon Classification (CC) from its fourth edition (1952) onward is an example of freely faceted classification. One might reasonably assume that the future belongs to faceted systems, which are analytico-synthetic in nature, see Aida Slavic, "Classification Revisited: A Web of Knowledge," in *Innovation in Information Retrieval: Perspectives for Theory and Practice*, ed. Allen Foster, & Pauline Rafferty (London: Facet, 2011), 23-48.

[33] Cutter's work was followed by the American Library Association (ALA) in 1908, 1941 and 1949, the Anglo-American Cataloguing Rules (AACR) in 1967, a second edition of AACR2 in 1978, and a revision of the AACR2R issued in 1988 which was based on ISBDs.

[34] In 1908 the first results of an agreement on standardization were achieved with the publication of *Catalog Rules: Author and title Entries.* A number of publications during the 1930s, 1940s until the 1950s Cataloging Rules and principles and in 1961 the Paris Principles. The latter was a major event in cataloging history and constitutes an important fact in the history of FRBR. William Denton, "FRBR and the History of Cataloguing" in *Understanding FRBR: What It Is and How It will Affect our Retrieval*, ed. Arlene G. Taylor (Westport, [CT]: Libraries Unlimited, 2007), 35-57; Barbara Tillet, *What is FRBR? A Conceptual Model for the Bibliographic Universe* (Washington DC: Library of Congress, 2003).

[35] Currently, according to W3C we have a restrictive definition: metadata is machine understandable information for the Web; Priscilla Caplan, *Metadata Fundamentals for All Librarians* (Chicago: American Library Association, 2003), 2.

[36] Cristina D. Ortega, "Ordering Documents: Underpinnings and Relations with Bibliographic Classification," in *Knowledge Organization and Cultural Diversity*, ed. José A. C. Guimarães, & Vera Dodebei (Pernambuco: ISKO Brasil; UFPE, 2017), 726-55, on 732.

software such as DSpace from Dublin Core to MARC.[37] Metadata is normally processed and stored in local databases, but to be exchanged it needs to be transported (allowing export-import). Local systems use formats as transport syntaxes, being MARC[38] the most common syntax used by librarians. According to Karen Calhoun, MARC transformed cooperative cataloging, thus providing libraries a new plane on which to build resource sharing, references and local library systems.[39]

Bibliographical metadata is important because it allows accessing and linking (i.e., Open URL exploitable by link resolvers),[40] search representations of the key information of a bibliographical item (automatic extraction of key-terms from the article content) and interoperability and application of different services to bibliographical information.[41]

During the 2004 meeting of the Working Group on Metadata Schemes held at IFLA[42] Conference attention was paid to the relationship between traditional card catalogs and the Resource Description Framework (RDF), which is the standard for creating the Semantic Web. BIBFRAME was another community effort led by the Library of Congress to start the transition from MARC 21 to RDF.[43]

According to Emmanuelle Bermès, from the French National Library,

> "[…] it may be time for libraries to start moving beyond the deeply buried data silos that are today's library catalogues towards freeing

[37] Amy L. Allen, Deborah E. Kulczak, & Mary A. Gilbertson, "From Digital Repositories to the Library Catalogue: Two Workflows for Transforming Metadata," *Journal of Digital Media Management* 6, no. 1 (2017): 95-117.

[38] MARC - Machine Readable Cataloging - was developed by the Library of Congress (LC) in the 1960s to enable computer production of catalog cards that would be distributed through the Cataloging Distribution Service. It is important to notice that MARC is a composite of a structure for machine-readable records (defined in ANSI/NISO Standard Z39.2) and a set of encoding rules documented in MARC21 and other LC publications. It constitutes what can be represented in machine-readable form by specific formats.

[39] Karen Calhoun, "Pre-print Being a Librarian: Metadata and Metadata Specialists in the Twenty-first Century," in *Metadata and Digital Collections: A Festschrift in Honor of Thomas P. Turn*er, ed. Elaine Westbrooks, & Keith Jenkins. (Lanham [MD]: Scarecrow Press, 2005), 12.

[40] A link resolver is a product that, figuratively speaking, sits between a 'source,' where a user begins the search process, and a 'target,' i.e., where the user goes next. The 'source' might be an abstracting and indexing database, an A-Z list of journal titles, an OPAC, or a footnote in an electronic full-text article. If the source system can build a standards-compliant Open URL containing metadata (typically bibliographic citation information) about the 'object' the user is interested in, then a link resolver like SFX can generate a list (or menu) of relevant targets. Targets might be the electronic full text of the cited article (perhaps available from a provider other than the source where the user found the citation), a document delivery request form or a web-based service that will automatically reformat a bibliographic citation according to a specified style manual, suitable for pasting into the user's bibliography. https://www.carli.illinois.edu/products-services/link-resolver-sfx

[41] Patrice Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," in *Research and Advanced Technology for Digital Libraries*, ed. Maristella Agosti et al. (Berlin: Springer, 2009), 473-4.

[42] IFLA- International Federation of Library Associations and Institutions.

[43] Library of Congress, *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services* (Washington; DC: Library of Congress, 2012). http://www.loc.gov/bibframe/pdf/marcldreport-11-21-2012.

> bibliographic data from the confinements of the catalogue and making it open, available and reusable as part of the global 'Web of Data'."[44]

Within the new digital space, the rules for cataloging (at least in the Anglo-American framework, AACR2) are being replaced by Resource Description and Access (RDA) specially designed for the digital environment. RDA defines the elements required for description and access and gives instructions for the formulation of the data to be record in each element. RDA data can be encoded, stored and transmitted using existing technologies and databases (as MARC records used to do in traditional library catalogs). According to Barbara Tillet, from the Library of Congress, RDA is intended to enable the creation of well-formed metadata about resources that can be used in any environment, whether a card-based catalog, an online catalog or a web-based interactive resource discovery system.[45]

According to Martin Malmsten, although library catalogs contain enormous amounts of structured, high-quality data, this data is generally not made available to Semantic Web applications.[46] According to him, even though bibliographic exchange has been a reality for decades, exchange of authority information and links between records are still not widely implemented.[47] In contrast, the Semantic Web[48] is by definition built upon linking of information.

The idea of the Semantic Web is that data are presented in machine-readable format, and linked together in such a way that the links can be followed and explored by humans and machines alike.[49]

As defined by the W3C consortium:

> "The Semantic Web is a web of data [...] the collection of semantic web technologies (RDF, OWL, SKOS, SPARQL, etc.) provides an environment where applications can query that data, draw inferences using vocabularies, etc. Linked data lies at the heart of what the semantic web is all about: large scale integration of, and reasoning on, data on the Web (W3C, 2010)."[50]

[44] Emanuelle Bermès, "Enabling Your Catalogue for the Semantic Web," in *Catalogue 2.0: The Future of the Library Catalogue*, ed. Sally Chambers (London: Facet Publishing, 2013), 117-42, on 117.

[45] Barbara B. Tillett, *RDA: Resource Description and Access*. Series of webcasts. https://www.loc.gov/catdir/cpso/rdawebcasts.html.

[46] The standard way of making bibliographic data available is still through search-retrieve protocols such as Z39.50. Though this makes single bibliographic records retrievable, it does not provide a way to directly address them and reveals little or nothing about links between records; see Martin Malmsten, "Making a Library Catalogue Part of the Semantic Web," *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2008, 146-152, on 146.

[47] Malmsten, "Making a Library."

[48] Tim Berners-Lee, James Hendler, & Ora Lassila, "The Semantic Web," *Scientific American* 284 (2001): 34-43; Berners-Lee, *Linked Data*. http://www.w3.org/DesignIssues/LinkedData.html.

[49] Bethan Ruddock, "Linked Data and the LOCAH Project," *Business Information Review* 28, no. 2 (2011):105-111.

[50] W3C.Linked Data, 2010. http://www.w3.org/standards/semanticweb/data. The term Semantic Web is used to refer to the technologies and standards used for structuring and linking of data by providing a proper description of concepts, terms and their associations within a given domain, such as RDF framework and

Therefore, one of the types of data being enthusiastically added to the Linked Data landscape is bibliographic data, for library catalogs to provide data that can lend itself to meaningful links and relationships.[51] Smith et al. provide a list of linked data for libraries (LD4L) and for production projects (LD4P) as a survey of current library linked data implementation including the major libraries in the world.[52] They explain how linked data can increase efficiency in the cataloging workflow and n how authority control[53] has been the area of linked data transition that has caused the most concern.

It seems to be a consensus that librarians are the staff with the specialized and advanced knowledge needed in the transition to a web semantic and linked data. According to Karen Calhoun, to support intellectual access to this zone [Semantic Web] librarians should engage in "distributed metadata management" rather than in the traditional "bibliographic control."[54] According to Jane Greenberg, the Semantic Web needs librarians and information professionals (who according to her have been working together for a while) because of their expertise with standards and bibliographic control, as well as their experience as information custodians for the last one hundred years.[55]

Digital content and semantic relationships, classification and information retrieval

Catalog records serve as retrieval aid for a library and its collections.

Older generations of catalogs allowed users to search and find their objects on shelves. Today catalogs allow relevant works not only to be found, but to be instantly read and copied.

Developing Digital Humanities and mass book digitization projects made possible for digitized materials to become viewable in a page-by-page manner and to be content searchable.

As is known, historians rely on many types of resources and bibliographical databases for finding and searching their documents. Currently, the research work of a

schema, SKOS, SPARQL (which is a RDF query language), N3, N-triplets (a format for storing and transmitting data), Turtle and OWL.

[51] "In the Linked Data, not just documents but also data elements and links between these data elements exist. Linked Data extends the current web that consists of documents and the computer-meaningless links between documents," Martin Dostal, *Text-Mining with Linked data* (PhD thesis, University of West Bohemia, 2014), 14.

[52] MacKenzie Smith et al., *BIBFLOW: A Roadmap for Library Linked data Transition* (California: University of California Library Davis; Zepheira Inc., 2017), 1-63. BIBFLOW is an Institute for Museum and Library Services (IMLS) funded multi-year project of the University of California Davis Library and Zepheira Corporation.

[53] " [Authority control] is the work that deals with the formulation and recording of authorized heading forms in catalogue records" [such] "that names and other headings that are access points to records are given one and only one conventional form," Robert L. Maxwell, *Maxwell's Guide to Authority Work* (Chicago: American Library Association, 2002),1.

[54] Karen Calhoun, *The Changing Nature of the Catalog and its Integration with other Discovery Tools: Final Report* (Washington DC: Library of Congress, 2006).

[55] Jane Greenberg, "Advancing Semantic Web via Library Functions," in *Knitting the Semantic Web.* ed. Jane Greenberg, & Eva Méndez (New York: Routledge, 2007), 203-26.

historian can be shortened to allow preliminary work to be done from the historian's own desk in their institution or home. Appreciation of the full content of many documents is afforded by computational methods applied to digitized materials that are text searchable.[56]

That a digital document can be text searchable means that is now technically feasible for a computer to identify its content. Machines were taught to extract factual data and meaning, they mine the text.[57] In practice, text mining seeks to identify words or phrases that explain possible underlying structures and relationships in the data uncovered through distribution analysis, association rules, or different clustering approaches.[58]

Most text mining systems allow users to browse.[59] It is assumed that the browsing distributions in a text mining system also work as a more efficient retrieval system. Traditional document retrieval systems allow users to ask for all documents containing certain terms or concepts, providing them an entire set of matching documents with little or no information on its content and internal structure. In turn, browsing distribution in a text mining system enable users to investigate the content of a document.[60] The system also provides visual tools, such as graphics that allow users identify patterns, navigate and appraise data, while using visualization software that supports RDF and XML data interchange approaches.[61]

Therefore, text mining can be used to classify documents based on their content.[62] In the specific case of the digital library of CESIMA,[63] a text mining method called SOBEK is being used.[64] SOBEK looks for frequent terms and their relationship in a text, thus enabling the identification of relevant terms and representing them in a graphical way.[65] SOBEK can be employed for the analysis of any type of text, and it can also analyze a collection of texts, even when they are presented in different formats. By comparing the list of terms extracted from a collection of texts with the terms extracted from one single text within that collection a database can be constructed, and new relationships can be inferred from their analysis. Such automatic generated databases can be edited, and databases can also be created if

[56] Abstracting and indexing of library resources during the twentieth century were strongly modeled on the physical sciences, which means brevity and reduction; Matthew B. Gilmore, & Donald O. Case, "Historians, Books, Computers and the Library," *Library Trends* 40, no. 4 (1992): 667-86, on 676.

[57]Mining is how search engines operate to allow discovering content and making connections; Ronen Feldman, & James Sanger, *The Text Mining Handbook* (Cambridge: Cambridge University Press, 2007).

[58] Ibid.

[59] Although text mining systems can generate a huge number of patterns and relationships, they also have a browsing interface that supports queries, thus affording refined tools for the shaping, constraining, pruning and filtering of result-set data.

[60] It does it by sorting the document according to the distribution of any node in a concept hierarchy. When documents are analyzed in this way, users can identify a specific subgroup within the search; Feldman & Sanger, 177-80.

[61] Ibid., 194.

[62] Indeed, it has been recognized as a promising tool in classifying fields of science; see: Patrick Glenisson et al., "Combining Full Text and Bibliometric Information in Mapping Scientific Disciplines," *Information Processing and Management* 41 (2005): 1548-72.

[63] CESIMA Digital is scheduled to be launched on mid-2018; for updates see http://www.pucsp.br/pos/cesima/

[64] SOBEK is a text-mining tool developed by Grupo de Pesquisa em Tecnologia Aplicada à Educação (GTech.Edu), Federal University of Rio Grande do Sul (UFRG), Brazil.

[65] Eliseo Reategui et al., "Sobek: A Text Mining Tool for Educational Applications," in *Proceedings of the International Conference on Data Mining* (Las Vegas: IEEE, 2011), 59-64.

needed.[66] At this point, SOBEK has been used to generate graphs about specific items hold at CESIMA digital library.

Such graphs are being added to the CESIMA library catalog. The catalog has been organized through the use of open source digital library software DSpace. As is known, DSpace includes many fields for description of the documents (title, author, etc.) and also allows customizing new fields. The graphs generated by the application of SOBEK to the collections resulted in new fields in the DSpace platform.[67] Such incorporation allows users to get a prior analysis of the document's content (if the graph shows the internal relationships between terms and concepts). Graphs are provided showing the relationships between document and collection, document and field or between document and subject heading and so on. Tasks and relationships can be defined depending on the choice of a particular combination of techniques to apply in a particular situation, as a function of the nature of the data mining task and the nature of the available data.

Final remarks

Libraries have been restructuring their services and rethinking the roles of the catalog to position themselves on the web, which has become the main source of information and documents.

For many centuries, catalogs were responsible for providing patterns of distribution and use of resources. Historically, the discovery and location processes were tied to each other in catalog; until today, the goal of cataloging is to make library collections findable and discoverable so they can be used.

In the present study I addressed some points relative to cataloging practice:

1. The restrictions inherent to older bibliographic classification systems. As is known, bibliographic classification systems are inherently hierarchical systems, therefore, unidimensional,[68] contrary to the network system currently available in the Web.

2. The fact that catalog entries could only be classed under one number, and numbers were not primarily assigned for the purpose of content nor for subject searching, but to park a book. As a result, the classification systems still in use are not meant to suit machine searching.[69]

[66]Ibid., 60-1.

[67]For more about the digital library project and the use of SOBEK, see Jose L. Goldfarb, & Odécio Souza, "Possibilidades de Sintaxe para eScience: A Utilização de uma Ferramenta de Mineração de Textos," in *Anais do 15 Seminário Nacional de História da Ciência e da Tecnologia* (Florianópolis: UFSC, 2016), 1-14.

[68] Jean-Claude Gardin, "Elements d'un modèle pour la description des lexiques documentaires," *Bulletin des Bibliotèques de France*, no. 5 (1966): 171-82.

[69] Janet S. Hill, "Online Classification Number Access: Some Practical Considerations," *Journal of Academic Librarianship* 10, no. 1 (1984): 17-22.

3. Catalogs normally lack descriptions of document contents (despite the assigning of subject headings in online catalogs and because keyword searching has limitations).[70]

4. Bibliographical frameworks are treated as ahistorical, while they must be contextualized. Bibliographical schemes and classifications need to be updated, they must follow theoretical frameworks and epistemologies.[71] And last, new trends in technology demand the sharing of metadata.

In the history of cataloging, the priority of improvement was not given to the development of semantic relationships found within the knowledge structures involved in catalog practice, but to standardization and internationalization of the record data.[72]

Nevertheless, as seen in the case of CESIMA library, it is possible to add value, structure and semantic information to catalog bibliographic records.[73] With text mining technology, semantic comparison of document content can be done with that of library metadata records which have been indexed with that concept to find the best matches. As is known, library collections have been organized using knowledge organization systems (KOS) based on literary warrant and accepted scientific and disciplinary schemes involving documentary classification, thesauri, vocabularies and subject-heading systems, which need to be constantly updated. The semantic web, with open contents and digital objects, is mapped with metadata, and their relationships have to be examined so as to match such semantic infrastructures. Semantic relationships need to be built on theoretical backgrounds.[74]

Since catalogs are navigating a new digital space, any library collection faces the pragmatic issue of how to make its metadata available to other collections and external searching software. It cannot be denied that there is an increasing partnership between the academic community, libraries and the information industry (such as Google, Microsoft, etc.) and their digital content and associated metadata should be key commodities.[75]

[70] Carol A. Manfred, & Judith Herschman, "Online Subject Access-Enhancing the Library Catalog," *Journal of Academic librarianship* 9, no.3 (1983): 148-55.

[71] Birger Hjørland, "Semantics and Knowledge Organization, "*Annual Review of Information Science and Technology* 41(2007): 367-405; Birger Hjørland, & Nissen K. Pedersen, "*Journal of Documentation* 61, no. 5 (2005): 582-97.

[72] Currently, through the Functional Requirements for Bibliographic Records (FRBR), based on the most relevant principles of Anglo-American cataloging, in turn grounded on the great cataloging traditions of the world. See Virgil L.P. Blake, "Forging the Anglo-American Cataloging Alliance: Descriptive Cataloging 1830-1908," *Cataloging & Classification Quarterly* 35, no. 1/2 (2002): 3-22; IFLA Meeting of Experts on an International Cataloguing Code, "Statement of International Cataloguing Principles," http://www.nl.go.kr/icc/down/070412_2.pdf.

[73] Machine readable forms of classification schedules can also be incorporated into online catalogs to provide additional keyword access points and a browsing structure for users; Ray R. Larson, "Experiments in Automatic Library of Congress Classification," *Journal of the American Society for information Science* 43, no. 2 (1992): 130-48, on 131.

[74]As Hjørland explains relations in and among KOS' structures are not a priori in spite of this claim in the literature; Birger Hjørland, "Are Relations in Thesauri "Context-free, Definitional, and True in All Possible Worlds"? *Journal of the Association for Information Science and Technology* 66, no. 7 (2015): 1367-73.

[75] Paul Miller, "Coming Together around Library 2.0: A Focus for Discussion and a Call to Arms," *D-Lib Magazine* 12, no. 4 (2006): doi: 10.1045/april2006-miller.

The convergence between library metadata and linked data should be based on the library interests (constructing vocabularies, describing properties of resources, identifying resources, exchanging and aggregating metadata) that are driving the development of semantic web technologies.[76] Catalogs are not simply inventory lists. Library catalogs are formal structures for organizing knowledge.

[76] Rachel Heery, "Metadata Futures: Steps Toward Semantic Interoperability," in *Metadata in Practice*, ed. D.I. Hillman & E.L. Westbrooks, (Chicago: American Library Association, 2004), 257-71.