Journal of Informetrics 11 (2017) 1080–1094

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Mapping science using Library of Congress Subject Headings

Fei Shu[a,*], Jesse David Dinneen[b], Banafsheh Asadi[a], Charles-Antoine Julien[a]

[a] School of Information Studies, McGill University, Canada
[b] School of Information Management, Victoria University of Wellington, New Zealand

ARTICLE INFO

Article history:
Received 16 March 2017
Received in revised form 30 August 2017
Accepted 31 August 2017

Keywords:
Map of science
LCSH
Gephi
Assignment
Citation

ABSTRACT

Maps of scientific knowledge are generally created by analyzing scientific literature including journal articles, conference proceedings, books, and monographs. Although citation analysis is the most popular method for generating maps of science from scientific journal articles and their citations, other relationships between scientific topics can be used to map science. This study offers a map of science generated from examining non-fiction book topics and their relationships as defined by Library of Congress Subject Heading (LCSH) co-assignments. The resulting map reveals which sub-disciplines of science must be learned together, showing that *Physics* and *Mathematics* are the central topics required to practice science, which is not revealed by previous studies. This novel LCSH-based science map reveals new relations between the major sub-disciplines of science to produce a more complete representation of scientific domains and how they interact.

1. Introduction

Maps of science are meant to visualize the structure and evolution of scientific inquiry (Börner, Theriault, & Boyack, 2015; Klavans & Boyack, 2015) by classifying science and relating the classes, which are generally derived from the analyses of elements of scientific literature such as authors, journals, disciplines, or other information (Klavans & Boyack, 2009). One of the challenges of mapping science is to create a valid placement of scientific domains and their relationships (Suominen & Toivanen, 2016): relatedness can be identified by examining, for example, expert judgements, citations, subject categories, topic modeling, or course descriptions. Each of these has limitations but the Library of Congress (LC) Subject Headings (LCSH), the most widespread knowledge organization in the world (Klavans & Boyack, 2009), has never been used to map science. A collection organized by LCSH topics offers a new way to map knowledge that reflects content published in non-fiction books. This differs from traditional citation-based maps of science that reflect how research disciplines collaborate to produce new knowledge, while an LCSH-based map has the potential to uniquely reveal which topics must be learned together as expressed by the topic co-assignments of non-fiction books. This study presents a map of science generated from LCSHs assigned to a representative non-fiction science book collection and investigates differences with previous maps of science.

1.1. Mapping science

The earliest maps of science were created using expert judgment and drawn by hand. Bernal (1939), Ellingham (1948) and Balaban and Klein (2006) presented maps that explicitly represent the hierarchical structure of science topics, but their

[*] Corresponding author at: School of Information Studies, McGill University, 3661 Peel, Montreal, QC H3A 1X1, Canada.
E-mail address: fei.shu@mail.mcgill.ca (F. Shu).

http://dx.doi.org/10.1016/j.joi.2017.08.008
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

Table 1
List of knowledge mapping approaches using citation analysis.

Mapping Approach	Unit of Analysis	Topics	Relationships	Studies
Direct citation	Journals	Controlled topics are assigned to journals	Citations between journals, articles or journal disciplinary categories create a relationship between their respective topics.	Bassecoulard and Zitt (1999), Leydesdorff (2015), Leydesdorff, Moya-Anegón and Nooy (2015), Leydesdorff and Rafols (2012)
	Papers	Controlled topics are assigned to individual articles		Boyack and Klavans (2014b); Pan, Zhang and Wang (2013); Waltman and Eck (2012)
	Categories	Controlled topics are assigned to journal disciplinary categories.		Leydesdorff, Carley and Rafols (2013), Leydesdorff and Rafols (2009), Zhang, Liu, Janssens, Liang, and Glänzel (2010)
Bibliographic coupling	Journals	Controlled topics are assigned to journals	Bibliographic coupling of articles or journals create relationships between their respective topics.	Boyack (2008)
	Papers	Controlled topics are assigned to individual articles		Boyack (2008)
co-citation	Papers	Controlled topics are assigned to individual articles	Co-cited articles or categories create relationships between their respective topics.	Boyack and Klavans (2014a), Klavans and Boyack (2007), Klavans and Boyack (2008), Small (1993, 1999), Small and Garfield (1985), Small and Griffith (1974)
	Categories	Controlled topics are assigned to journal disciplinary categories.		Moya-Anegón et al. (2007); Moya-Anegón et al. (2004)
Hybrid approach	Journals	Combination of different citation-based approaches including direct citation, co-citation, bibliographic coupling etc.		Boyack, Klavans and Börner (2005), Leydesdorff (1987), Gomez-Nunez, Vargas-Quesada, Moya-Anegon, Chinchilla-Rodriguez and Batagelj (2016), Tijssen, Raan, Heiser and Wachmann (1990)
	Papers			Boyack and Klavans (2010), Braam (1991a, 1991b), Janssens, Glänzel and De Moor (2008); Persson (2010), Tijssen et al. (1990)

main disciplines and map placement differ (Klavans & Boyack, 2009). The exception to these early manual approaches was Small and Griffith (1974), who created the first citation-based map of science; co-citation analysis can express the extent to which disciplines cite each other to create new knowledge. Disciplines can be placed on a 2D map where, for example, proximity or edge thickness expresses higher co-citation rates.

Citation analysis is currently the dominant method used to generate data for knowledge maps. Table 1 presents existing approaches to generate citation-based map data. Maps created with direct citation analysis between journals or their disciplinary categories represent a broad, discipline-level structure within science that offers limited detail (Waltman & Eck, 2012), while finer questions are answered using maps derived from document-level analyses like direct citation or co-citation (Boyack & Klavans, 2014a). Hybrid approaches combining two or more different citation analysis are reported to be more accurate (Boyack & Klavans, 2010).

Table 1 presents the predominant knowledge map data generation approaches beyond which there are alternatives. For example, co-words analysis, regarded as an alternative to co-citation analysis, generates map data from the co-occurrence of words in titles, abstracts or keywords (Ding, Chowdhury, & Foo, 2001; Leydesdroff, 1989; Peters & van Raan, 1993a, 1993b; Rip & Courtial, 1984). Balaban and Klein (2006) mapped science as it was represented in undergraduate course pre-requisites at Texas A&M University, and Suominen and Toivanen (2016) developed a map of science using topic modeling to visualize latent patterns in texts retrieved from the Web of Science (WoS). Taken as a whole, the knowledge mapping literature shows a clear preference for citation-based maps, while acknowledging that other mapping approaches are necessary to provide a comprehensive understanding of the relative importance of knowledge disciplines and how they might be related.

1.2. Study objective

Citation-based maps use data provided by citation indexing databases (e.g., WoS, Scopus, or Google Scholar) that generally index journal articles while excluding other types of documents such as books.[1] The resulting maps of knowledge reflect how disciplines draw upon each-other to produce new scientific knowledge; however, a different story is likely to emerge if only books are considered. Non-fiction scientific books/monographs differ from scientific articles: books tend to cover broader

[1] Citations to books and other materials are now included is some citation databases (e.g., WoS and Scopus), but they are still rare in mapping science practice.

topics that already have a sufficient body of specific scientific literature required to fill the much longer book format. Books are clearly significant to scientific research; Larivière, Archambault, Gingras and Vignola-Gagnè (2006) show that they account for 13–32% of citations in the natural sciences and 50% in the social sciences and humanities. Boyack and Klavans (2014b) have also confirmed the importance of non-journal article documents by comparing direct citation maps that include or exclude sources other than journal articles (e.g., regional journals, conference papers, books and their chapters, etc.) from Scopus: they found that books are more frequently cited than journal articles and are cited by all knowledge disciplines.

The purpose of this study is to improve our understanding of scientific disciplines and their interactions by developing an alternative approach to mapping science by examining the LCSH topics assigned to non-fiction books. As the following section shows, using LCSH topics to map science adds to existing knowledge maps by expressing which topics must be learned in conjunction.

2. LCSH science map

LCSH is the standard list of possible book topics used by most large general libraries in North America, many special and smaller libraries, and many libraries in other countries (Chan & Salaba, 2016). LCSH is a controlled topical vocabulary maintained by the LC for the creation of subject access points and their assignment to bibliographic records. LCSH should not be confused with LC Classification (LCC), which is designed to produce an alphanumeric representation of the book's main topic as part of its *call number,* which then serves as the unique address for a physical object on a book shelf. Assuming LCC numbers are available, though this is not the case in many eBook collections, analyzing LC classification is a more direct and tractable computational task than analyzing LCSH.

LCSH topics are preferable to LCC numbers for mapping science because the semantic content offered by the LCC number, although valuable, is necessarily a subset of that provided by assigned LCSHs that describe *all* major topics discussed within a book. Contrary to LCC's single main topic, having multiple LCSHs per book permits the analysis of LCSH co-assignments, which can be used as a measure of the relative strength of the relationship between two LCSHs. Co-assigned LCSH relationships express the likelihood that existing knowledge about two topics will be read together in the same book, which differs from citation analysis since the latter expresses how frequently disciplines cite each-other's publications to create new knowledge.

2.1. LCSH process and standardization

LCSHs are assigned to books by LC-trained, Master's level professionals who possess qualifications or demonstrable knowledge of a relevant subject or specialist knowledge (e.g., Biology, Arts History, Law). The books' electronic records are then shared by LC and downloaded with little or no change by member libraries throughout the world. This means that a given book can generally be assumed to have an identical or highly similar set of LCSHs in any LCSH-organized collections. Fig. 1 illustrates the process of LCSH assignment to standard bibliographic records. It shows that publishers participating in LC's Cataloguing-In-Publication (CIP) Office —as mandated by legal deposit laws in the US — submit galleys of their books that are processed through LC cataloging channels, which include the maintenance of the publicly accessible LCSH authority records that define the thesaurus. This produces the books' CIP information, including their appropriate LCSHs, which are then returned to the publisher for inclusion in the book itself (Chan & Salaba, 2016). The electronic record of the CIP metadata is shared in MARC[2] format from LC's Program for Cooperative Cataloging (PCC). Aggregators such as OCLC's WorldCat[3] then distribute existing bibliographic records collected from LC and member libraries, and they in turn can download existing standard bibliographic records. By storing and aggregating bibliographic records from member libraries throughout the world, OCLC currently offers over 10.5 million[4] standard bibliographic records of published works housed in libraries throughout the world.

Individual libraries can and sometimes do create original cataloging records (e.g., for unique, non-published works) or modify existing ones (e.g., to further specify a topic of local significance); however, the mass availability of relatively inexpensive downloadable records has created a systematic process of *copy cataloguing* whereby

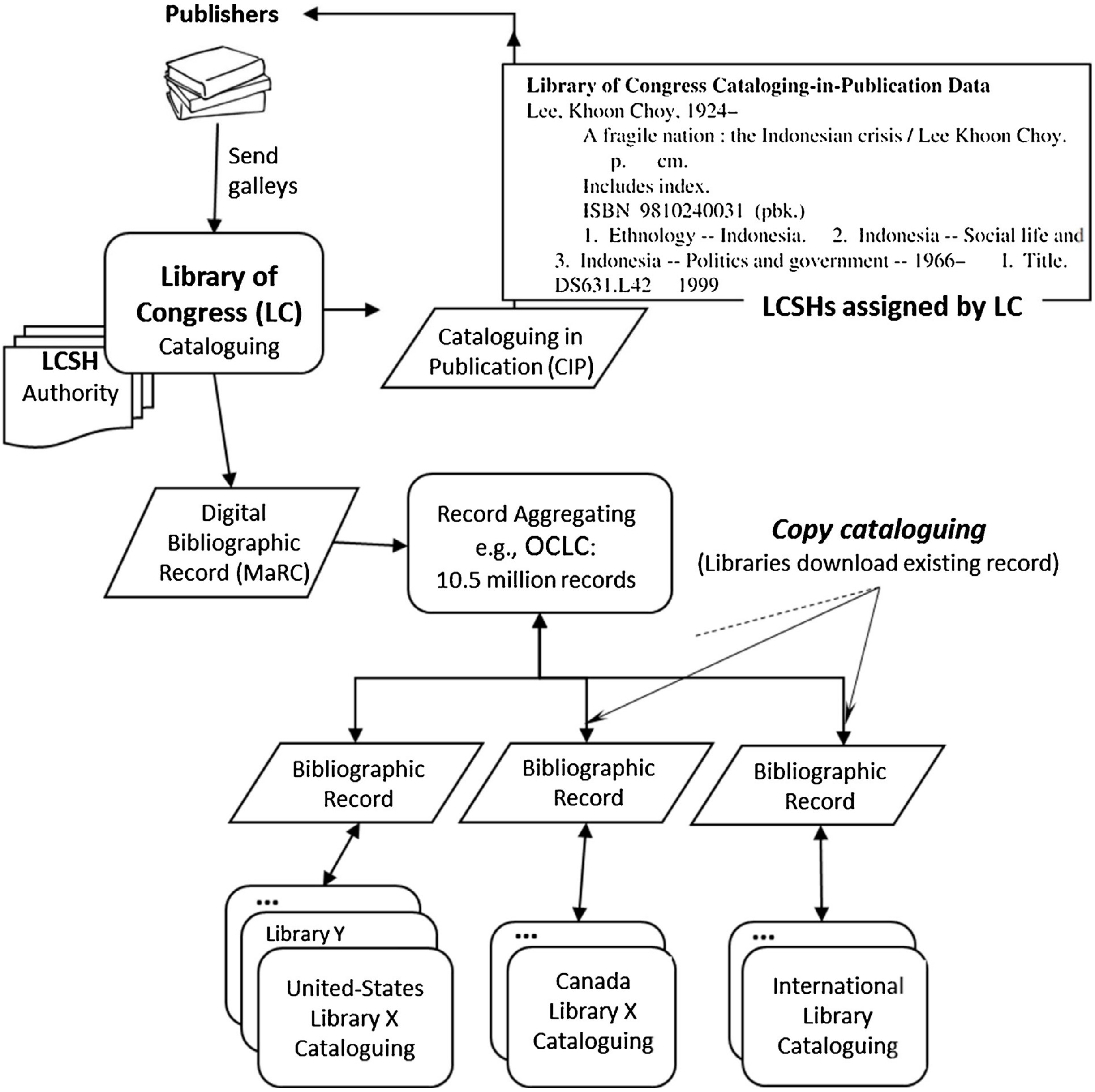
> "most local cataloging departments stay as close as possible to strict [downloaded record] copy cataloging because doing so has been found to bring about a large increase in the productivity of cataloging staff" (Chan & Salaba, 2016, p. 72).

This trend towards LCSH standardization across collections is accentuated by eBook subscription batch uploads, which can contain thousands of records with standard LCSH assignments that are generally used as-is by subscribing libraries. Additionally, licenced international libraries have developed multi-lingual LCSH versions that organize some of the world's largest non-English academic library collections; for example, a French-Canadian version of LCSH is maintained by Quebec

[2] https://www.loc.gov/marc/.

[3] https://www.worldcat.org/.

[4] https://www.oclc.org/support/services/worldcat-org/faq.en.html.



Library of Congress Cataloging-in-Publication Data
Lee, Khoon Choy, 1924–
A fragile nation : the Indonesian crisis / Lee Khoon Choy.
p. cm.
Includes index.
ISBN 9810240031 (pbk.)
1. Ethnology -- Indonesia. 2. Indonesia -- Social life and
3. Indonesia -- Politics and government -- 1966– I. Title.
DS631.L42 1999

Publishers
Send galleys
Library of Congress (LC) Cataloging
LCSH Authority
Cataloguing in Publication (CIP)
LCSHs assigned by LC
Digital Bibliographic Record (MaRC)
Record Aggregating e.g., OCLC: 10.5 million records
Copy cataloguing
(Libraries download existing record)
Bibliographic Record
Bibliographic Record
Bibliographic Record
...
Library Y
United-States Library X Cataloguing
...
Canada Library X Cataloguing
...
International Library Cataloguing

Fig. 1. LCSH assignment and copy cataloging process of bibliographic record standardization.

City's Laval University,[5] who in turn licences their *vedettes-matière* to other francophone university libraries (e.g., University of Montreal) that organize their collections using LCSH.

Given the widespread distribution of standard LCSH assignments across libraries caused by copy cataloging, we can reasonably expect a single book to have similar LCSHs across collections. This supports the generalizability and comparison of results between LCSH-organized collections since, in the case of books published in the US, the original LCSHs assigned by LC are likely to be systematically downloaded by every international library who has purchased the book or subscribes to the eBook. This avoids the inconsistencies in LCSH assignment that may arise across human indexers.

The question of topical indexing quality, relevance, or utility is an important one but beyond the scope of this study. We are assuming an adequate quality to be useful given that LCSH is the dominant topical vocabulary in North-American, and it is widely used by English language academic libraries across the world. LCSH is a case study of a working, evolving,

[5] http://rvmweb-v2.bibl.ulaval.ca/.

large topical information organization system containing over 240 K topics,[6] and is currently used by hundreds of thousands university students, faculty, and staff. Data describing an LCSH-organized collection therefore have the evidential potential to describe users' expectations about topical indexing structures and to improve topic and searching tools.

A collection's LCSH structure is the result of this topic vocabulary's organic growth and evolution for 120 years, and its application is performed using formal guidelines developed over that time; for example, assigned LCSHs should represent the major topics of the work in order of predominance. Major topics are those covered by approximately 20% of the work, a maximum of six topics is generally appropriate, and no more than ten may be assigned (Library of Congress, 2016). Therefore, no matter the knowledge discipline, works having more than six assigned topics would be anomalous. Unfortunately, beyond the ordering of their predominance, it is not possible to determine the relative importance of any single LCSH for a given book, and they are assumed to be equivalent *ceteris paribus* for the purposes of this study.

This section has shown that broad international standardization of bibliographic records and their assigned LCSHs is supported by practicality and cost reduction pressures; there is generally no need to recreate an existing record and local modifications are rare given the sheer bulk of current eBook batch subscriptions. Given its relative consistency across books and collections, LCSH-organized collections can serve as data sources for producing a novel map of science that covers all knowledge domains expressed in non-fiction books. Thus, LCSH serves as rich grounds for data analysis and visualization that can express in novel ways the predominant knowledge domains within science and their relations.

3. Research questions

An LCSH-based map of science would provide a novel representation of scientific literature published in non-fiction books since LCSH co-assignments express the likelihood that two topics are covered in the same book. Although it is a promising approach (Klavans & Boyack, 2009) there has never been an LCSH-based map of science, and this study therefore seeks to answer the following research questions:

RQ1. Can a map of science be produced from LCSH assignments?

RQ2. How is an LCSH-based map of science like and different from existing citation-based maps?

4. Methodology

This section describes the datasets used, their manipulation, and calculation of the topic relationship strengths. Finally, a resulting singular dataset is filtered to reduce visual elements to a manageable number to produce a usable map of science.

4.1. Data

A collection organized using LCSH necessarily entails two datasets: 1) bibliographic records that are assigned at least one LCSH topic that may have been modified to suit the needs of the local collection (also called *LCSH string*), and 2) the LCSH authority records that contain the *established LCSHs* and their structure of topic relationships using broader/narrower topics, related topics, and synonyms (i.e., *used for* terms). Related topics represent horizontal relations, which are not considered by the current study. Broader topics are directed vertical relationships that define a hierarchy of broad (e.g., *Science*,[7] *Medicine*) to narrow terms (e.g., *Mechanics*, *Surgery*). For example, let us assume *Aircraft accidents–Human factors– Quebec (Canada)* is a subject heading assigned to a bibliographic record. *Aircraft accidents–Human factors* is the established subject heading listed in the LC topical authority records. *Aircraft accidents* is the main heading whose topical authority record states that it has a broader term: *Transportation accidents*. Note that, as described in Section 2, when assigned to a specific book, established LCSHs are often modified to increase their specificity. These modifications are called *subdivisions:* in practice, they are indicated in the data by a double-dash (e.g., Quebec (Canada) is a subdivision of Aircraft accidents–Human factors), which must generally be parsed and removed before the assigned LCSH can be reverted to its authority record version and its broad to narrow topic structure.

The specific datasets are comprised of 204,430 topical LCSH authority records and 122,393 bibliographic records representing the physical material housed in a major North American University Science and Engineering library. The subset of LCSH representing the domain of science is chosen since there is an existing citation-based maps of science with which to compare (Leydesdorff and Rafols (2009)), and this knowledge domain is an extreme test for this study since it offers the deepest and most connected portion of the LCSH hierarchy (Julien, Tirilly, Leide, & Guastavino, 2012; Yi & Chan, 2010).

Fig. 2 presents the distribution of publication dates found in the bibliographic collection: it contains works published between 1901 and 2008, with the clear majority from the 1960's onward. Since the bibliographic records represent material housed in a science and engineering library they can be assumed to represent typical literature about *Science*.

[6] https://www.loc.gov/aba/publications/FreeLCSH/LCSH39%20Main%20intro.pdf.

[7] The reader should not confuse the italicized LCSH term *Science* with the equivalent journal. The journal is never referenced in this text.

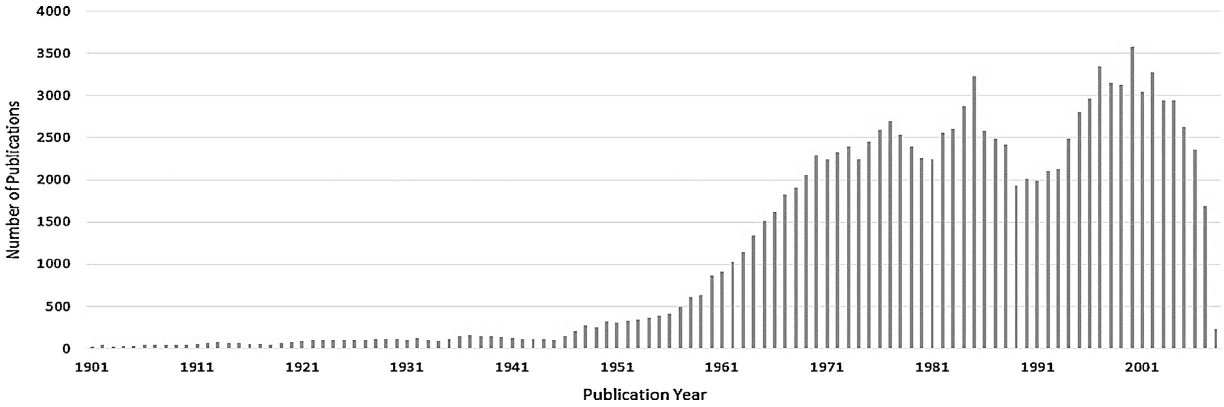


Fig. 2. Distribution of publication years within the bibliographic collection.

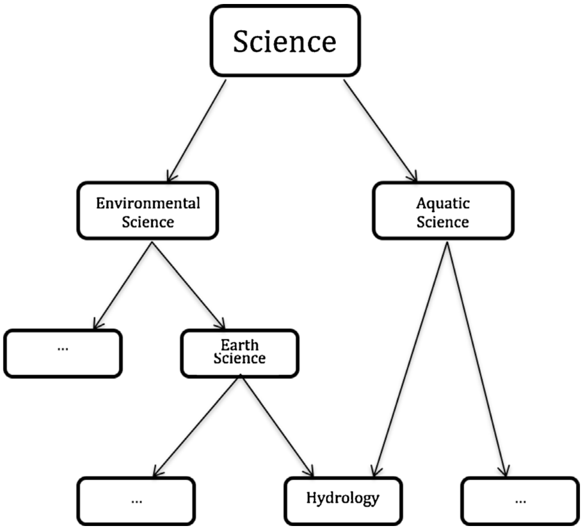


Fig. 3. Example of LCSH hierarchy multiplie inheritance.

4.2. Multiple inheritance

The LCSH hierarchy is not a true tree since topics often have multiple parents, also called *multiple inheritance*, which hampers the calculation of the LCSH relationship strength. Fig. 3 shows the example of *Hydrology*, which is a narrower term of *Aquatic Science* and *Earth Science:* this illustrates the issue that one parent topic, *Science* in this example, can reach a descendant topic, *Hydrology* in this example, via multiple branches, and a reason why Klavans and Boyack (2009) argue that the LCSH structure does not lend itself to visual mapping. Multiple inheritance entails that there can be multiple paths between a specific topic and its broader ancestor(s); thus, a co-assigned pair of LCSHs in one single book can create multiple relationships between a pair of their broader ancestors. This multiplicative effect must be controlled such that, for any abstraction level of the LCSH hierarchy, relationship strengths between topics express the number of unique books that relate the topics.

4.3. LCSH hierarchy modifications

For the purposes of this study the LCSH *Science* is considered the root or topic that has no parent and is the ultimate ancestor to all topics. Fig. 4 presents a sample of the LCSH hierarchy related to *Science* and 4/59 of its predominant disciplinary branches in terms of their respective numbers of accessible assignments. Since *Science* has no parent it is labelled level 0 of the hierarchy. Visual inspection reveals that Leydesdorff and Rafols (2009)'s citation-based map of science does not contain the broad disciplinary subdivisions offered by the level 1 LCSHs shown in Fig. 4 (i.e., *Physical sciences*, *Technology*, *Life sciences*, *Natural history*). For example, Fig. 4 shows that *Engineering* is a level 2 topic belonging to *Technology* (see bottom left branch in Fig. 4), and *Chemistry* is a level 2 topic belonging to *Physical Sciences* (see top left branch in Fig. 4); this is not consistent

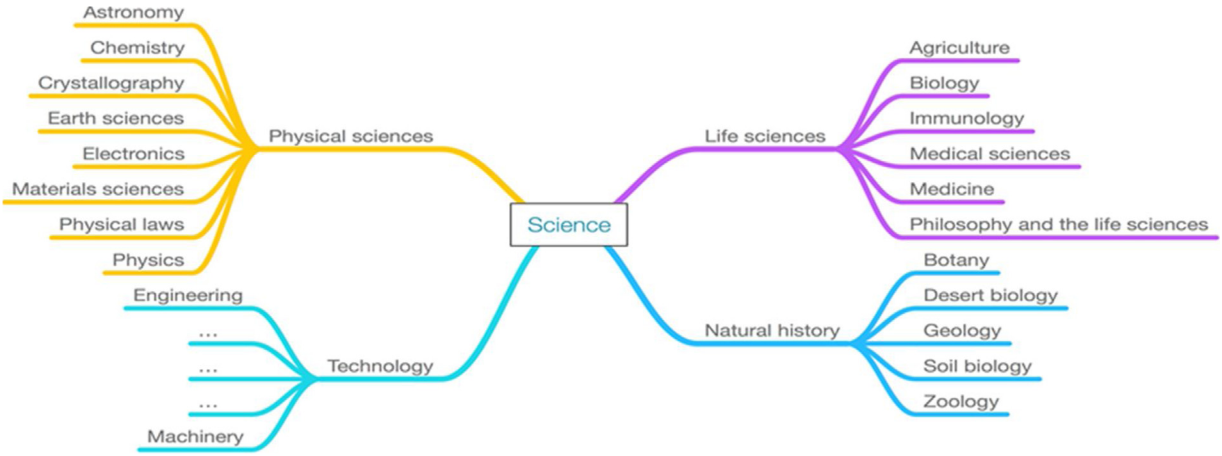


Fig. 4. Sample of LCSH hierarchy showing Science and some of its major disciplinary branches.

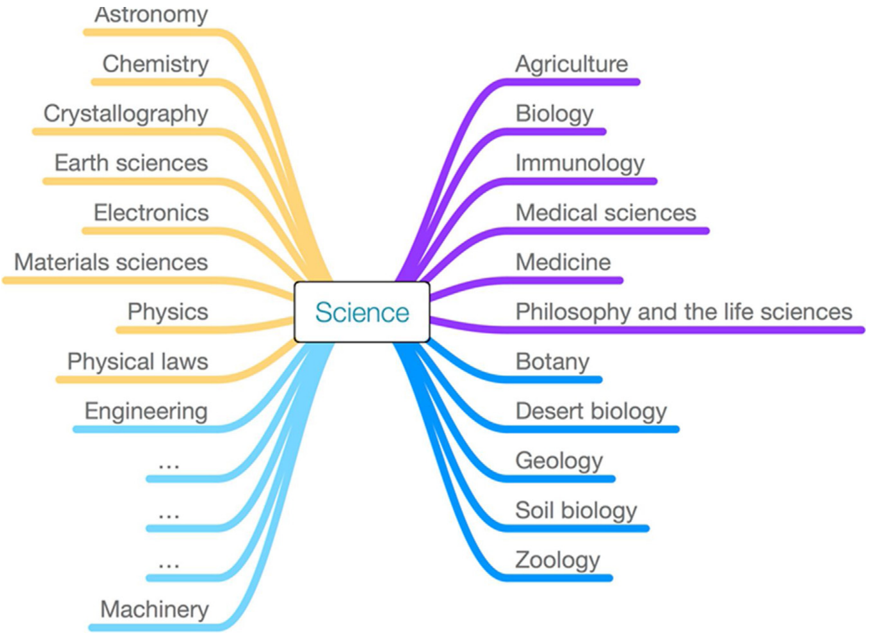


Fig. 5. Sample of LCSH hierarchy from Fig. 4 where broad disciplinary subdivisions are removed to permit comparison with previous map of science from Leydesdorff and Rafols (2009).

with Leydesdorff and Rafols (2009) where *Engineering* and *Chemistry* are level 1 topics and immediate narrower terms of *Science*.

Visual inspection reveals that replacing the level 1 LCSHs shown in Fig. 4 with their respective narrower terms produces an organization of *Science* that is comparable to Leydesdorff and Rafols (2009). Fig. 5 shows the modified portion of the LCSH hierarchy shown in Fig. 4 where, for example, LCSHs *Engineering* and *Chemistry* are now immediate narrower terms of *Science*. Fig. 4Fig. 4's replacement of level 1 topics with their immediate narrower terms produces Fig. 5Fig. 5's map that represents primary disciplines of science whose scopes are comparable to Leydesdorff and Rafols (2009). Note that the replaced LCSHs' bibliographic assignments are re-assigned to their immediate broader parent (i.e., *Science* in this case); for example, bibliographic records assigned to *Life sciences* are reassigned to *Science*. This changes topic assignments' specificity but the impact is marginal since non-fiction books are generally not about such broad topics as *Life Sciences* or *Natural Science*. Indeed, these structural changes to the LCSH semantic hierarchy changed less than 0.05% (i.e., 466/1,016,538) of the collection's LCSH assignments.

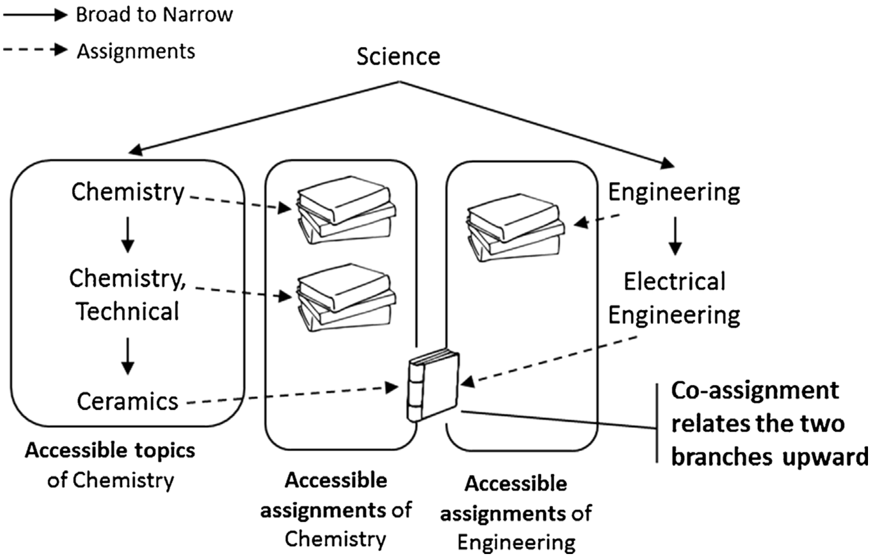

Broad to Narrow
Assignments
Science
Chemistry
Chemistry, Technical
Ceramics
Engineering
Electrical Engineering
Co-assignment relates the two branches upward
Accessible topics of Chemistry
Accessible assignments of Chemistry
Accessible assignments of Engineering

Fig. 6. Illustrative example of relation of LCSH branches by a co-assignment between their descendants.

4.4. *Strength of topical relationships*

We define the strength of the relationship between any two LCSHs as the count of unique bibliographic co-assignments between them or any of their respective sets of descendant LCSHs. In other words, given two LCSH branches, the strength of the relationship between the two root LCSHs is the number of times that one LCSH from both branches is assigned as a pair to a bibliographic record (i.e., a book). This section formalizes these definitions.

We define:

Accessible Topics (AT): AT(x) is the set of topics that includes an arbitrary LCSH x and all its narrower term descendants, if any. AT(x) is generated by travelling the hierarchy starting from x and recursively following all its narrower term options while collecting the unique LCSHs visited until the bottom of the hierarchy has been reached and no additional narrower terms can be followed (i.e., until leaf nodes are reached).

Accessible Assignments (AA): AA(x) the set of bibliographic records assigned to AT(x). Direct LCSH assignments are part of each book's metadata, and Julien et al. (2012) describe a method to associate the assigned LCSHs to their closest established LCSH within the topic hierarchy. AA(x) is generated by listing book identification numbers (e.g., BookId) that have been assigned any LCSH from AT(x).

Two LCSHs are related if they share at least one AA element or book: the count of shared books, *S*, is the measure of topic relatedness used in this study or the likelihood that two topics are covered by the same book.

For example, Fig. 6 shows a sample of the LCSH hierarchy where *Electrical Engineering* is a narrower term to *Engineering*, whose broader term is *Science*. *Ceramics* is a narrower term to *Chemistry, Technical*, whose broader term is *Chemistry*, then *Science*. The accessible topics of *Chemistry* include it and all its descendants (i.e., *Chemistry, Technical*; *Ceramics*), while its accessible assignments include any bibliographic record that is assigned any of its accessible topics. Fig. 6 shows a book that is assigned both *Electrical Engineering* and *Ceramics*, which creates a co-assigned LCSH relationship between the two branches; in other words, *Chemistry* and *Engineering* are related because they share an accessible assignment, for the same reason *Chemistry, Technical* and *Electrical Engineering* are related. This process effectively propagates co-assigned relationships up the chain of broader term ancestors, which circumvents the structural issue shown in Fig. 3 and allows mapping of co-assigned LCSH relationships at any higher level of the LCSH hierarchy to address the concerns of Klavans and Boyack (2009).

Therefore, *S* for any two LCSHs x and y, is calculated as the cardinality of the intersection between the AA(x) and AA(y); in other words, the count of books assigned both x and y, and any of their respective descendants. This is denoted as:

$$S(x, y) := |AA(x) \cap AA(y)|$$

where | AA(x) ∩ AA(y) |: the cardinality of the intersection between AA(x) and AA(y) (as defined above) returns the number of unique books that relate x (or any of its descendants) and y (or any of its descendants).

Fig. 7 presents a fictitious topic hierarchy whose broadest terms A and B respectively have three and one narrower term(s) as represented by solid single arrow lines, and dashed double-arrow lines represent a co-assignment of the two topics in the same book. To determine the strength of the relationship between A and B one would sum all co-assignments between them and any between their descendants, which would produce a strength of 8.

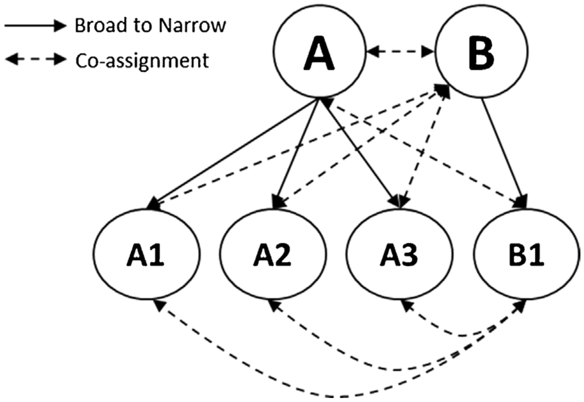


Fig. 7. Sample demonstration of LCSH structure and co-assignments. In this case the strength of topical relationship A-B, or the sum of co-assignments between all pairs of topics to which they provide access, is 8.

Note that beyond the potential effects of multiple inheritances, a single book with multiple co-assigned LCSH pairs can also relate two broader topics multiple times. To illustrate let us assume that co-assignments B ↔ A1 and B ↔ A2 are found in the same book; in other words, a single book relates A ↔ B twice. For any hierarchy level, we posit that any one book should be able to relate a pair of broader topics only once. Duplicated relations between topics, whether resulting from multiple inheritance or duplicated co-assignments, are excluded from the AA sets (i.e., by set theory definition). In this example, the resulting strength of the A↔B relationship would be reduced from 8 down to 7 since a single book relates these topics twice but counts towards the relationship only once.

4.5. Data analysis

Two datasets are required to produce an LCSH topic map: 1) a list of LCSHs with their respective number of accessible assignments, and 2) co-assigned LCSH pairs and their respective number of co-assignments. Our datasets contained 1370 level 2 LCSHs, of which 843, or one of their descendants, were assigned at least once. The data analysis of the LCSH assignments per bibliographic record resulted in 31,532 unique co-assigned LCSH pairs for a total of 40,675,353 co-assignments.

Fig. 8 shows the distribution of LCSH assignments for the 843 level 2 LCSH that represent the collection. The top curve shows a highly skewed distribution where a small group of topics are assigned often while most topics are seldom assigned. This is confirmed by the bottom curve that indicates the relatively small number of topics that represent the clear majority of the collection. Specifically, the top 12.7% (107/843) most assigned LCSHs account for 95% of all assignments. This number of topics is an estimate of the beginning of the long tail of LCSHs beyond which topics can be filtered while maintaining the accuracy of the collection's topical representation. Note that highly skewed distributions are common in organized information found in other knowledge domains (Egghe, 2005; Julien, Tirilly, Dinneen, & Guastavino, 2013; Newman, 2005); thus, there are strong indications that organized collections in general follow the same kind of distribution within their respective organizing structure. This highly skewed distribution shows that, as far as non-fiction books are concerned, a few topics tend to receive most authors' attention, while most topics receive little or no attention.

Fig. 9 shows the distribution of co-assigned pairs of level 2 LCSHs (N = 31,532). The top-curve shows the number of times each pair was found co-assigned in a unique bibliographic record, plotted on a log scale to visually reveal the highly skewed distribution that is otherwise hidden by the chart axes. This figure shows that very few LCSH pairs are co-assigned over 10,000 times while most pairs are found 1000 times or fewer. The bottom curve of Fig. 9 shows that the top 5.9% (1882/31,532) most co-assigned LCSH pairs account for 95% of all co-assignments. This is an estimate of the start of the long tail beyond which the strength of the topical relationship could be considered less representative of the broad collection topics.

4.6. Data visualization

There are no strict rules regarding the selection of representative data for visualization (McCain, 1990) but thresholds have been frequently used in science mapping (Leydesdorff & Rafols, 2009; Yan, Ding, & Zhu, 2010; Zhao & Strotmann, 2008). Two threshold filters were applied to reduce our dataset to a manageable number of visual elements: 1) mapped LCSHs should provide access to least 500 bibliographic records (i.e., count of accessible assignments >500) and 2) LCSH relationships should have at least 2000 co-assignments. These threshold values were chosen to include the smallest subset of topics and pairs that account for at least 95% of assignments and co-assignments. Fig. 10 shows that threshold 1 leaves 128 LCSHs, representing 96.6% of the total assignments (982,212/1,016,538), and threshold 2 leaves 100 LCSHs, which relate 1787 pairs representing 95% of co-assignments. The resulting map of science represents the overall topical pattern of the collection and how these topics are related in terms of how many times they are read together in the same book.

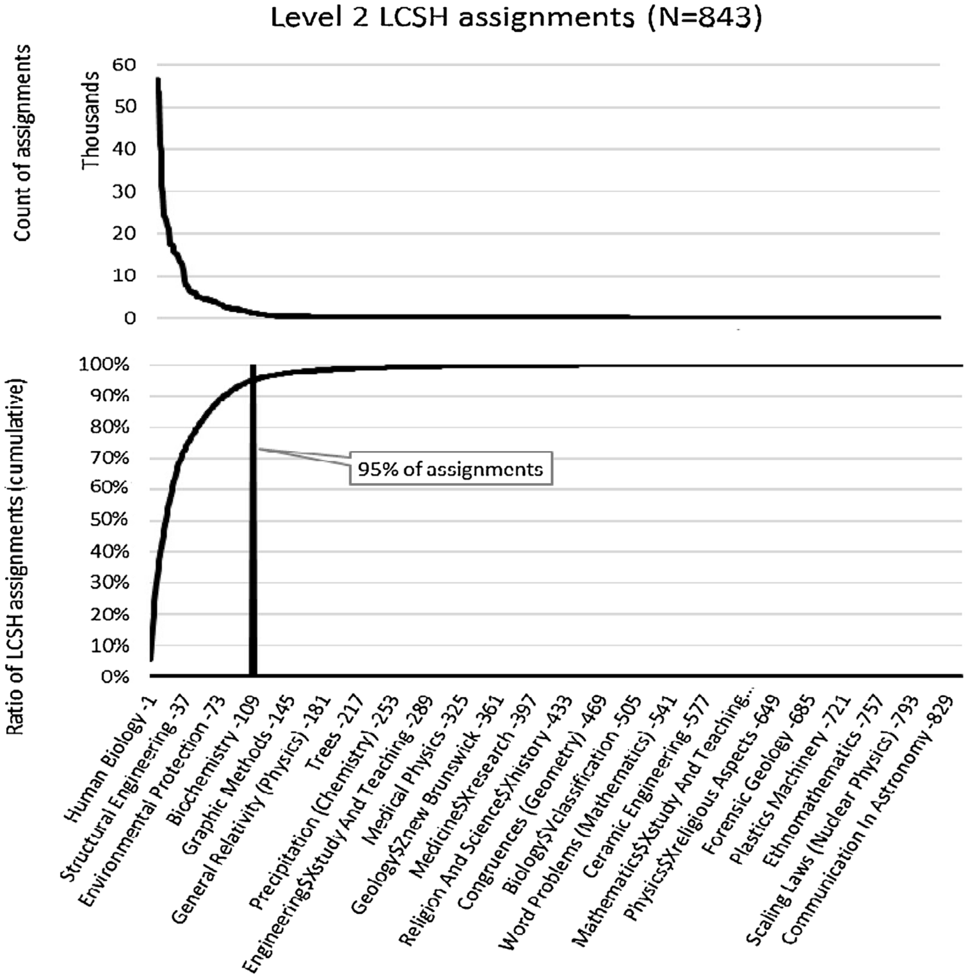


Fig. 8. Highly skewed distribution of level 2 LCSH assignments. Top curve shows the number of times the LCSH is assigned, the bottom curve shows the relatively small number of topics that account for 95% of assignments.

LCSHs and their co-assignments are imported into graph-drawing software (Gephi, 2015) to generate a visual map of science where LCSHs are nodes drawn as coloured circles and co-assignments are drawn as edges (i.e., lines) between them. Node color signifies its parent LCSH, and edge color matches the node colors if they share the same parent; if they do not, the edge is drawn as the color between two node colors. Although there are different versions of the established force-directed layout (e.g., Fruchterman Reingold; Yifan Hu Multilevel), only the Force Atlas offers weighted edges, which was used to represent the number of co-assignments between two LCSHs. The resulting map is visually inspected and compared with that of Leydesdorff and Rafols (2009) in terms of the topics and their relationships.

5. Results

Fig. 11 shows the LCSH map of science containing 100 LCSH nodes and 1787 edges remaining after threshold filtering. Nodes are level 2 LCSHs (i.e., the grandchildren of *Science*) coloured to match their level 1 parent LCSH (i.e., the direct children of *Science*). Edge width is proportional to the number of co-assignments between the two LCSHs, and the node and label sizes are proportional to the number of accessible assignments. Different modularity settings (resolution level at 0.5, 1, and 2) were tested to reveal node clusters but no obvious clusters were produced. The label adjust option was used to remove label overlaps, and no other layout options were applied. Node placement serves only to highlight predominant pairs of related topics.

Fig. 11 is best interpreted using its color-coded legend of broader disciplines (see bottom right of Fig. 11); for example, *Medical Sciences* (light purple in Fig. 11) includes *Clinical Medicine*, and *Medicine, Physical*, which are in the top 3 most prevalent LCSHs assigned to the collection (see Table 2) as indicated by their large node size in Fig. 11. The relationships between *Human biology, Clinical Medicine*, and *Medicine, Physical* are some of the strongest as indicated by the thick lines between these topics; in other words, books about *Clinical, Medicine* are often also indexed with *Human Biology* and/or

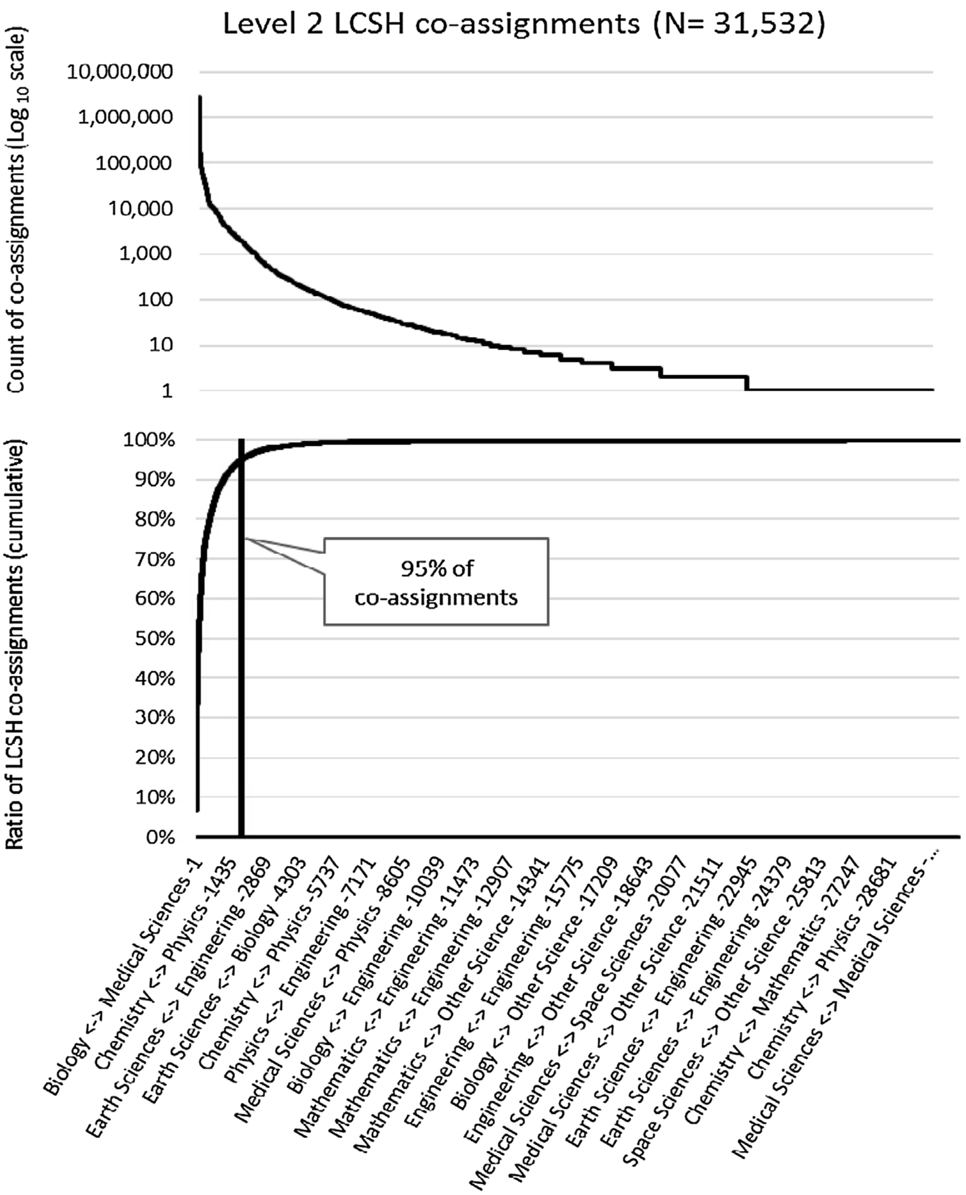


Fig. 9. Highly skewed distribution of level 2 LCSH co-assignments. Top curve shows the number of times the LCSH pair is co-assigned (log-scale), the bottom curve shows the relatively small number of LCSH pairs that account for 95% of co-assignments.

Table 2
Top 10 level 2 LCSHs in terms of their number of accessible assignments.

LCSH	Number of Assignment	Percentage of Assignment	Parent (Level 1)
Human Biology	56,667	5.93%	Biology
Clinical Medicine	53,731	5.63%	Medical Sciences
Medicine, Physical	53,654	5.62%	Medical Sciences
Mathematical Physics	41,353	4.33%	Physics
Mechanics	39,306	4.12%	Physics
Electric Engineering	31,550	3.30%	Engineering
Earth Sciences	29,586	3.10%	Environmental Sciences
Dynamics	24,230	2.54%	Mathematics
Statics	24,076	2.52%	Physics
Algebra	23,041	2.49%	Mathematics

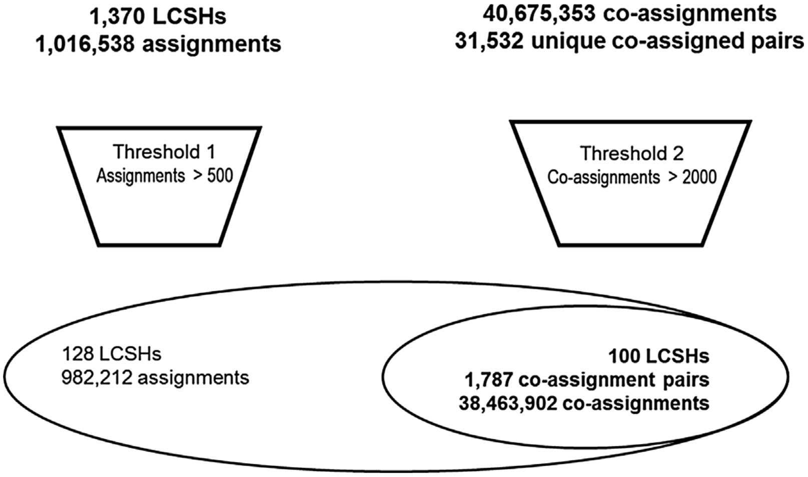


Fig. 10. Dataset threshold filtering for visualization.

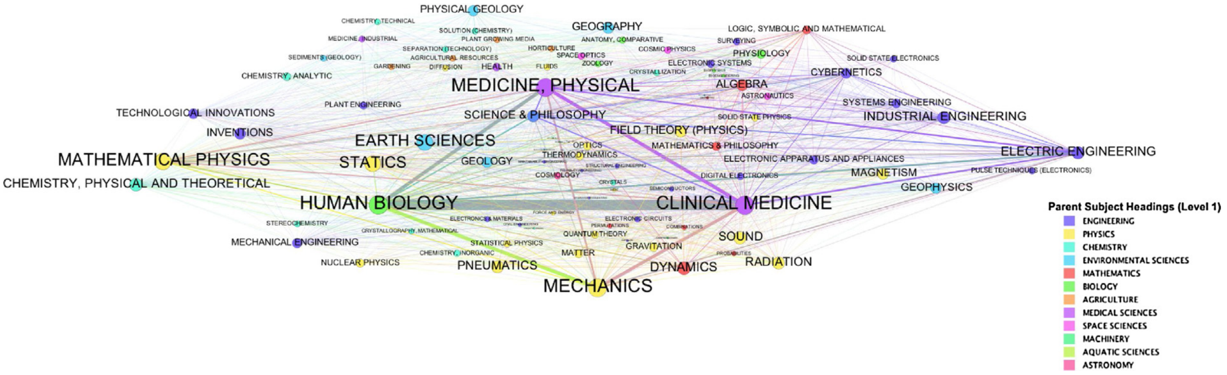


Fig. 11. High-level map of science derived from LCSH topic relatedness.

Table 3
Top 10 subject heading co-assignment (level 2).

Co-assigned 1	Co-assigned 2	Co-assignments	Percentage of co-assignments
Human Biology	Clinical Medicine	2,812,204	7.31%
Human Biology	Mechanics	954,666	2.48%
Human Biology	Medicine, Physical	937,293	2.44%
Clinical Medicine	Medicine, Physical	937,293	2.44%
Clinical Medicine	Mechanics	852,269	2.22%
Medicine, Physical	Mechanics	521,813	1.36%
Human Biology	Electric Engineering	418,860	1.09%
Clinical Medicine	Electric Engineering	418,858	1.09%
Medicine, Physical	Electric Engineering	418,730	1.09%
Human Biology	Mathematical Physics	328,722	0.89%

Medicine, Physical. Fig. 11 reveals that *Medical Sciences* (light purple nodes in Fig. 11) draw from the broad disciplines of *Physics* (yellow nodes in Fig. 11) and *Engineering* (dark purple or indigo nodes in Fig. 11); for example, as listed in Table 3, *Mechanics* (see yellow node at bottom of Fig. 11) is often co-assigned with *Clinical Medicine* and/or *Human Biology*, and *Electrical Engineering* (see dark purple node at right side of Fig. 11) is often co-assigned with *Medicine, Physical* and/or *Clinical Medicine*.

Despite our efforts to filter the dataset to a manageable number of visual elements, Fig. 11 is visually complex due to high connectivity between the nodes; overlapping edges make it difficult to interpret, which is sometimes called a *spaghetti* or *hairball* effect. To address these issues, Fig. 12 shows a simpler map containing level 1 LCSHs representing parent topics of the level 2 LCSHs shown in Fig. 11. Fig. 12 shows twelve level 1 LCSHs (see Table 4) that are related with 157 edges or pairs of co-assigned LCSHs. To further reduce visual clutter, the visualization excludes edges representing less than 2.5% of the total co-assignments or about 1,000,000 co-assignments; note that if this excludes all edges leading to a node then the largest of its edges is retained to keep the node visible and linked with the graph.

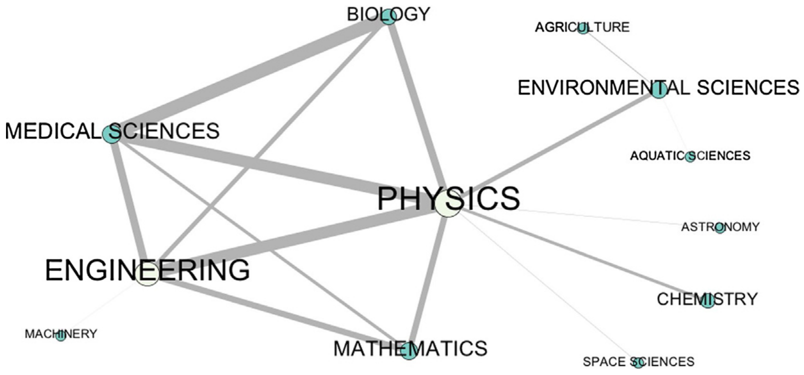


Fig. 12. A map of science derived from the relatedness of LCSH topics (Level 1).

Table 4
List of 12 level 1 LCSHs.

LCSH	Share of assignments	Number of Children (Level 2)
Physics	25.2%	21
Engineering	21.4%	25
Medical Sciences	12.2%	4
Mathematics	11.1%	10
Environmental Sciences	10.0%	10
Biology	8.5%	8
Chemistry	6.3%	11
Agriculture	1.7%	4
Space Sciences	0.6%	3
Astronomy	0.3%	1
Aquatic Sciences	0.2%	1
Machinery	0.2%	1

Fig. 12 and Table 4 confirm the strong relationships between *Medical sciences, Engineering, Biology* and *Physics* that were shown in Fig. 11. Fig. 12 also shows a strong relationship with *Mathematics* that is difficult to see in Fig. 11 since it is diffused across its sub-disciplines (see red nodes in Fig. 11). This highlights the importance of judiciously choosing an abstraction level to map; for example, Fig. 11 suggests that *Medical Sciences* is a predominant topic of this collection (e.g., *Medicine, Physical*; *Clinical Medicine*) but this is partly due to the fact that this topic has relatively few descendants (see number of children column in Table 4); therefore, its assignments stay concentrated in Fig. 11, as compared to *Physics* or *Engineering* whose predominance in the collection is diffused across their multiple important sub-disciplines in Fig. 11. In fact, Fig. 12 and Table 4 show that once the sub-disciplines of *Physics* are combined, it accounts for more than a quarter of accessible assignments, and it is the gateway hub that connects the highly-connected disciplines on its left (e.g., *Engineering, Medical Sciences, Mathematics*) with relatively independent topical satellites on its right (e.g., *Astronomy, Chemistry*). For example, books that are assigned *Engineering* (or one of its descendants) are often assigned *Physics* (or one of its descendants), while books on *Chemistry* have a good chance of also being about *Physics*.

6. Discussion

Our LCSH-based map differs from the map provided by Leydesdorff and Rafols (2009), which shows *Chemistry* as the hub connecting a large cluster of *Medical sciences* (e.g., *Biomedical Science, Clinical Medicine, Neuroscience*) and hard sciences such as *Physics, Engineering,* and *Material Science*. In contrast, Fig. 12 shows that *Physics* is the most prominent topical hub in our LCSH-based map. Books about *Medical sciences* are more likely to also be related to *Physics* (or one of its descendants) than *Chemistry*, which is more often cited by scholarly journal articles in the *Medical sciences.*

The maps also differ in that the LCSH map does not require normalization; this contrasts with the map provided by Leydesdorff and Rafols (2009), which visualizes a broad range of citation counts using a logarithmic scale. As a result, Fig. 11Fig. 11's node sizes directly represent relative numbers of accessible documents without transformation. LCSH assignments can therefore be said to be a more *direct* representation of the structure of science since it is not affected by every discipline's specific citing practices; indeed, the map by Leydesdorff and Rafols (2009) does not show *Mathematics*, whose citation rate is too low to visualize even after normalization. In contrast, the results shown in Fig. 12 reveal that *Mathematics* is co-assigned with most major sub-disciplines of science (e.g., *Engineering*, *Physics*, and *Medical Sciences* in Fig. 12).

In hindsight, it may seem *a priori* obvious that science and engineering make heavy use of mathematics. This may be intuitive for those with experience in science education, but does not necessitate that co-assigned major topics will include mathematics. Maps of science derived from LCSH have never been reported before now, and our maps now confirm the intuitive importance of mathematics across science's sub-disciplines, which is supported by Balaban and Klein (2006)'s findings from their map derived from university course pre-requisites. This partly validates our novel LCSH mapping method since it reveals credible topic relationships that differ from those expressed by existing maps of knowledge.

7. Limitations

Some of the differences between our LCSH-based map and prior works are due to differences between the respective topical structures used; for example, Leydesdorff & Rafols's (2009) map is based on the ISI topic structure, which includes *Biomedical Sciences,* whereas LCSH includes *Medical Sciences;* the equivalency of these topics would have to be verified. Differences might also be partly due to differences in visualization methodologies, in turn stemming from choice of visualization software, normalization method, or filtering threshold; for example, Leydesdorff & Rafols (2009) use Pajek for visualization and normalize the data using cosine normalization, whereas we used Gephi for visualization and did so without data normalization. Finally, the bibliographic collection used by this study was retrieved from physical material housed in a major North American University Science and Engineering library, and results should thus be validated with other academic library collections; however, due to the wide practice of copy cataloging, LCSH assignments are likely to be relatively consistent across collections. In addition, since social sciences and humanities are not covered in our LCSH map, it is inappropriate to compare our map to citation-based maps where social science and humanities are represented.

8. Conclusion

This study mapped science using LCSHs assigned to a science and engineering academic library collection where LCSH co-assignments represent relationships between topics. Two LCSH-based maps of science were generated to represent topical relationships expressed in books that are not covered by traditional citation-based maps, which is a novel methodological contribution of this study.

The comparison of traditional citation-based maps of science to the LCSH-based maps broadens our understanding of the relationships between the major sub-disciplines of science. The differences with the prior maps of science are partly due to different meanings of the relationship measures: our maps show knowledge areas that must be learned together, as seen in non-fiction books, while citation analysis expresses how one discipline draws knowledge from or builds upon another. This is the main contribution of this study to existing knowledge, and proves that mathematical concepts are still an integral and critical part of knowledge required to understand non-fiction books about science and engineering.

Author contributions

Conceived and designed the analysis: Fei Shu, Jesse David Dinneen, Charles-Antoine Julien.
Collected the data: Banafsheh Asadi, Charles-Antoine Julien.
Contributed data or analysis tools: Fei Shu, Jesse David Dinneen.
Performed the analysis: Fei Shu, Jesse David Dinneen.
Wrote the paper: Fei Shu, Jesse David Dinneen, Banafsheh Asadi, Charles-Antoine Julien.

References

Börner, K., Theriault, T. N., & Boyack, K. W. (2015). Mapping science introduction: Past, present and future. *Bulletin of the American Society for Information Science and Technology*, *41*(2), 12–16.

Balaban, A. T., & Klein, D. J. (2006). Is chemistry The Central Science? How are different sciences related? Co-citations, reductionism, emergence, and posets. *Scientometrics*, *69*(3), 615–637.

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics*, *44*(3), 323–345.

Bernal, J. D. (1939). *The social function of science*. London, UK: George Routledge & Sons Ltd.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Boyack, K. W., & Klavans, R. (2014a). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685.

Boyack, K. W., & Klavans, R. (2014b). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics*, *8*(3), 569–580.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.

Boyack, K. W. (2008). Using detailed maps of science to identify potential collaborations. *Scientometrics*, *79*(1), 27–44.

Braam, R. R. (1991a). Mapping of science by combined Co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233–251.

Braam, R. R. (1991b). Mapping of science by combined Co-citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, *42*(4), 252–266.

Chan, L. M., & Salaba, A. (2016). *Cataloging and classification: an introduction*. Lanham: Rowman & Littlefield.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817–842.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. San Diego, CA: Elsevier Academic Press.

Ellingham, H. (1948). Divisions of natural science and technology. *Paper presented at the report and papers submitted to the royal society scientific information conference* [Burlington House, London]

Gephi. (2015) (Version 0.8.2): Gephi Consortium.

Gomez-Nunez, A. J., Vargas-Quesada, B., Moya-Anegon, F., Chinchilla-Rodriguez, Z., & Batagelj, V. (2016). Visualization and analysis of SCImago Journal & Country Rank structure via journal clustering. *Aslib Journal of Information Management*, *68*(5), 607–627.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics*, *75*(3), 607–631.

Julien, C.-A., Tirilly, P., Leide, J. E., & Guastavino, C. (2012). Constructing a true LCSH tree of a science and engineering collection. *Journal of the American Society for Information Science and Technology*, *63*(12), 2405–2418.

Julien, C.-A., Tirilly, P., Dinneen, J. D., & Guastavino, C. (2013). Reducing subject tree browsing complexity. *Journal of the American Society for Information Science and Technology*, *64*(11), 2201–2223.

Klavans, R., & Boyack, K. W. (2007). Is there a convergent structure of science? A comparison of maps using the ISI and Scopus databases. In *Paper presented at the 11th international conference of the international society for scientometrics and informetrics*.

Klavans, R., & Boyack, K. W. (2008). Thought leadership: A new indicator for national and institutional comparison. *Scientometrics*, *75*(2), 239–250.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.

Klavans, R., & Boyack, K. W. (2015). Exploring the relationships between a map of altruism and a map of science. *Bulletin of the American Society for Information Science and Technology*, *41*(2), 30–33.

Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagnè, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, *57*(8), 997–1004.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from web-of-science data. *Journal of Informetrics*, *6*(2), 318–332.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, *94*(2), 589–593.

Leydesdorff, L., Moya-Anegón, F., & Nooy, W. (2015). Aggregated journal–journal citation relations in scopus and web of science matched and compared in terms of networks, maps, and interactive overlays. *Journal of the Association for Information Science and Technology*, *67*(9), 2194–2211.

Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, *11*(5-6), 295–324.

Leydesdorff, L. (2015). The dynamics of journal–journal citation relations: can hot spots in the sciences be mapped? *Paper presented at the The 78th ASIS&T Annual Meeting*.

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, *18*(4), 209–223.

Library of Congress. (2016). *Subject heading manual: H 0180 assigning and constructing subject headings*. [Retrieved 25 May, 2017, from]. http://www.loc.gov/aba/publications/FreeSHM/freeshm.html

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41*(6), 433–443.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, *61*(1), 129–145.

Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, *58*(14), 2167–2179.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Pan, J., Zhang, X., & Wang, X. (2013). Mapping the structure and evolution of science. *Journal of Information Processing and Management*, *52*(8), 255–266.

Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, *4*(3), 415–422.

Peters, H. P. F., & van Raan, A. F. J. (1993a). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy*, *22*(1), 23–45.

Peters, H. P. F., & van Raan, A. F. J. (1993b). Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy*, *22*(1), 47–71.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, *6*(6), 381–400.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. *Journal of Information Science*, *11*(4), 147–159.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies*, 17–40.

Small, H. (1993). Macro-level changes in the structure of co-citation clusters: 1983–1989. *Scientometrics*, *26*(1), 5–20.

Small, H. (1999). Visualizing science by citation mapping? *Journal of the American Society for Information Science*, *50*(9), 799–813.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476.

Tijssen, R. J. W., Raan, A. F. J. v., Heiser, W. J., & Wachmann, L. (1990). Integrating multiple sources of information in literature-based maps of science. *Journal of Information Science*, *16*(4), 217–227.

Waltman, L., & Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics*, *83*(1), 115–131.

Yi, K., & Chan, L. M. (2010). Revisiting the syntactical and structural analysis of Library of Congress Subject Headings for the digital environment. *Journal of the American Society for Information Science and Technology*, *61*(4), 677–687. http://dx.doi.org/10.1002/asi.21295

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, *4*(2), 185–193.

Zhao, D., & Strotmann, A. (2008). Information science during the first decade of the web: An enriched author cocitation analysis. *Journal of the American Society for Information Science and Technology*, *59*(6), 916–937.