



Metric Assessments of Books As Families of Works

Zuccala, Alesia Ann; Breum, Mads; Bruun, Kasper; Wunsch, Bernd Thomas

Published in:
Journal of the Association for Information Science and Technology

DOI:
[10.1002/asi.23921](https://doi.org/10.1002/asi.23921)

Publication date:
2017

Citation for published version (APA):
Zuccala, A. A., Breum, M., Bruun, K., & Wunsch, B. T. (2017). Metric Assessments of Books As Families of Works. *Journal of the Association for Information Science and Technology*, 69(1).
<https://doi.org/10.1002/asi.23921>

Metric Assessments of Books as Families of Works

Alesia Zuccala¹, Mads Breum, Kasper Bruun, and Bernd T. Wunsch

¹a.zuccala@hum.ku.dk

Royal School of Library and Information Science
University of Copenhagen
Njalsgade 76, 2300 København S, Denmark

Abstract

We describe the intellectual and physical properties of books as manifestations, expressions and works and assess the current indexing and metadata structure of monographs in the Book Citation Index (BKCI). Our focus is on the interrelationship of these properties in light of the Functional Requirements for Bibliographic Records (FRBR). Data pertaining to monographs were collected from the Danish PURE repository system as well as the BKCI (2005-2015) via their International Standard Book Numbers (ISBNs). Each ISBN was then matched to the same ISBN and family-related ISBNs cataloged in two additional databases: OCLC-WorldCat and Goodreads. With the retrieval of all family-related ISBNs, we were able to determine the number of monograph expressions present in the BKCI and their collective relationship to one work. Our results show that the majority of missing expressions from the BKCI are emblematic (i.e., first editions of monographs) and that both the indexing and metadata structure of this commercial database could significantly improve with the introduction of distinct expression IDs (i.e., for every distinct editions) and unifying work-related IDs. This improved metadata structure would support the collection of more accurate publication and citation counts for monographs and has implications for developing new indicators based on bibliographic levels.

1. Introduction

In the past, bibliographic data and citation data pertaining to books were inaccessible, if not difficult to retrieve. Now, as digital resources have improved, so has the priority to advance book-related metrics. This is partly due to the introduction of Thomson Reuter's Book Citation Index (BKCI) (Adams & Testa, 2011)¹ and the addition of books to Elsevier's Scopus. These commercial databases, however, are not the 'be-all and end-all' for the discerning bibliometrician. Recent assessments of the BKCI point to numerous indexing problems, which can lead to flawed evaluations (Gorraiz et al., 2013; Leydesdorff & Felt, 2013; Torres-Salinas et al., 2014). Still, researchers continue to use the BKCI and/or Scopus by finding ways to extract book citations from journal articles (Hammarfelt, 2011; Zuccala et al., 2014). Some have chosen instead to work with alternative resources, like Google Books (Kousha & Thelwall, 2009), Google Scholar (Kousha & Thelwall, 2011) and OCLC-WorldCat (Torres-Salinas & Moed, 2009; White et al., 2009). Concerted efforts are even being made to compare data that has been retrieved from multiple databases (Kousha et al., 2016; Zuccala & Cornacchia, 2016; Zuccala et al., 2015a; Zuccala & White, 2015b).

¹ At the time this research was carried out the Book Citation Index was owned by Thomson Reuters. It is now part of the parent company Clarivate Analytics.

The metric community is making rapid progress, but this is related primarily to the exploration of new data sources. The BKCI indexing problem therefore persists. One solution is to avoid studies based on citations and work with library holding counts instead (Torres-Salinas & Moed, 2009). With this option books cataloged in various international libraries (e.g., the OCLC-WorldCat Union catalog) may be evaluated according to “their perceived impacts on culture and the life of the mind” (White et al., 2009, p. 1086). Thus far, the libcitation has generally been accepted, though researchers are reluctant to separate libcitations from citation counts, suggesting that both indicators might be used in a complementary manner (Linmans, 2010; Zuccala & White, 2015b). To a large extent, the citation is inexorable: it is the principal indicator upon which the BKCI was founded, and remains pertinent to the use of other databases as well (e.g., Scopus and Google Scholar). Another solution for improving book-related metrics is to take the problem of book indexing more seriously and put more emphasis on index-related improvements. This approach does not rest entirely with the bibliometrician’s expertise, yet most studies that rely on indexes/book catalogs still point to the same issue: *regardless of where and how bibliographic and citation data are collected, it is essential to recognize that books often belong to bibliographic families.*

Since bibliographic families may be examined both theoretically and empirically, the aim of this study is to do both. First, we will examine and explain several interrelated concepts linked to a family-oriented entity-relationship model, known as the Functional Requirements for Bibliographic Records (FRBR). We have chosen to use this model, because it can effectively illustrate the extent to which books, as complex entities are not always indexed accurately in the BKCI, using appropriate metadata. In the second empirical part of this study, we will present some data collected specifically from the Book Citation Index (BKCI), OCLC-WorldCat, and Goodreads, and use this data to demonstrate why a robust model is necessary, in order to improve upon the accuracy of book-oriented metrics (i.e., citation counting). The empirical aspect of our research is based on the following question: *Do books currently indexed in the Book Citation Index (BKCI) have adequate metadata and data designed to reflect inherent familial components and relationships?*

2. Background to the Problem

2.1 Bibliographic entities and their properties

Counts of books as publications and/or counts of their received citations may be compounded or not, depending on how we recognize their *intellectual* and *physical* properties. According to Lubetzky (1953), all bibliographic entities possess at least two: an *intellectual property*, which we refer to as the *work* and a *physical property*, which is the container for the *work*. It is worth noting that when Lubetzky (1953) first established these definitions, digital media had not yet been introduced. Attempts have also been made since then to elaborate upon the term *work*; hence the general consensus today is that what we observe from a *work* is the synthesis of its ideational and semantic content (Smiraglia, 2001).

If we examine a journal article, we are likely to observe familial components based on a one-to-one, or a one-to-many relationship. An article’s *intellectual property* begins as a piece of *work* and its *physical property* can manifest as an official print publication and/or a digital publication with a Digital Object Identifier (DOI). The purpose of the DOI is to provide a persistent link to both the print and digital object. Circa Lubetzky’s time period (1950s) there would have been little confusion about what is counted when a journal article was accepted for publication, printed and indexed. Today, with print and

digital publishing, it can be interesting to examine when an article is officially published – i.e., if it is available online with a DOI, or printed and indexed at a later date (Haustein et al., 2015).

A monograph is similar to a journal article in that it typically appears first as one *intellectual contribution*, or *work*. Like a journal article, it may be published in print or digital form. Unlike the journal article, the monograph can be re-itemized as a new *edition*. Chi et al., (2015) initially reflect on this problem when they note that the BKCI sometimes includes different *editions* of the same *work*:

The BKCI distinguishes different editions of a book for some of its source items and indexes one or more editions of a work. For example, “CRIME SCENE TO COURT: THE ESSENTIALS OF FORENSIC SCIENCE, SECOND EDITION (2004)” and “CRIME SCENE TO COURT THE ESSENTIALS OF FORENSIC SCIENCE, 3RD EDITION (2010)” coexist in the database. Therefore, the citation links provided by the BKCI to the different editions of a book are edition sensitive and may need further judgment or weight for an additional evaluation process.

While it is clear that these items have been published as distinct editions, Chi et al.’s (2015) use of the term *work* needs further attention. A basic Google search for the second edition and third edition of “CRIME SCENE TO COURT” confirms that both have been published under the same title, with the same editor (WHITE, PC), but in different publication years. Moreover, a closer examination indicates that not only do they possess unique International Standard Book Numbers [i.e., ISBN: 978-1-84755-065-1 for the second edition and ISBN: 978-1-84755-882-4 for the third edition] they also do not share the same content. This is evidenced by the fact that each volume is made up of different chapter titles and different authors corresponding to each chapter. Chi et al., (2015) have the impression that both editions of “CRIME SCENE TO COURT” are the same *work*, but we show that this may not be the case.

2.2 The structure of bibliographic ‘families’

In the Functional Requirements for Bibliographic Records (FRBR) the term *work* is an abstract entity, which serves as the focal point for a full conceptual model of the bibliographic universe (Tillett, 2005). The FRBR was first developed by a study group, affiliated with the International Federation of Library Associations and Institutions (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998), but Tillett (2005), one of the original members of this group, explains that it was written to serve as “a generalized view... independent of any cataloguing code or implementation” (p. 24). Now it is often recommended for the restructuring of catalogs:

the number of records we make is a decision made up front by the cataloger based on local policies reflecting local user needs. We may choose to catalog at various levels: the collection of works (FRBR calls this an aggregation), an individual work, or a component of a work. At the description level, we may include a description of all the parts and should provide access to each component. At the component level, we should provide a link to relate to the larger ‘whole.’
(Tillett, 2005, p. 27)

Long before the FRBR was introduced, O’Neill and Vizine-Goetz (1989) were the first to examine the term *work* as part of an entity-relationship model of the bibliographic family. In this early model, the top concept of *work* refers abstractly to a common origin and content. Subsequent concepts – i.e., *text*,

edition, printing, and book – are used to gradually represent a more narrow understanding of a *work* down to the individual printed book on the shelf of a library. *Book* is the only term for a physical object and thus the only one that is not abstract. O’Neill and Vizine-Goetz (1989) explain also how a *work* and its physical object are linked on the basis of a one-to-many relationship: each *book* is affiliated with one *work*, but one *work* can have multiple books with which it is affiliated.

Tillett (2005) agrees with an abstract notion of *work*, but refers to a *text* and its specific arrangement of sentences, paragraphs chapters, etc. as an *expression*. The *expression* is then manifested by a specific version, leading to one example, which she calls an *item* (p. 25). These four concepts – i.e., *work, expression, manifestation, and item* – belong to a family tree with inherent relationships. It is a bibliographic family because “all texts of a work are derived from a single progenitor” (Smiraglia, 2001, p. 75). At the level of the original *work* there may be expressed *equivalent* works, such as copies (e.g., hardcopy or paperback) or reprints. There might also be expressed *derivatives*, which can include multiple editions, revisions, translations, etc. At the *descriptive* level, the family tree could also include reviews, commentaries, annotated editions or critical evaluations of the original work (Tillett, 2001).

Figure 1 illustrates what the FRBR entity-relationship model might look like as a guide to evaluating the current structure of the BKCI. This is an adapted version of Tillett’s (2001) figure, which was printed first in *Relationships in the Organization of Knowledge* and reprinted later in *What is FRBR? A conceptual model for the bibliographic universe* (Tillet, 2005). Note that our figure is designed to focus solely on scholarly *work* and indicates the cut-off point when an *expression* may be recognized as a new *work*. Below Figure 1, we present a list of concepts, which have also been adapted from Tillet (2001, 2005). Our definitions do not deviate too much from the classical definitions, but we include references to other texts in some cases for further clarification.

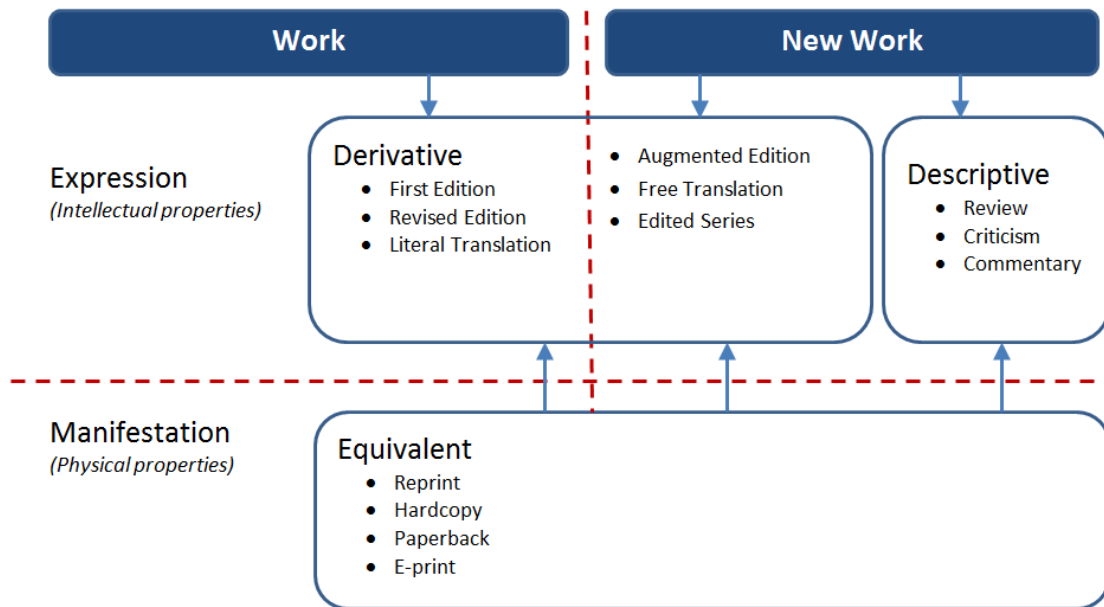


Figure 1. Modified model of bibliographic families for a scholarly work (Tillet, 2001; 2005).

1. **First Edition:** the emblematic or original version of a *work* as an intellectual contribution
2. **Revised Edition:** an edition that includes small corrections made to the original *work*
3. **Literal Translation:** a direct translation of the original language text into another specified target language text (e.g., Danish to English) whereby the intellectual domain and the historical-temporal context of the original work is recognized and maintained (Pellizzi, 2015).
4. **Augmented Edition:** a new edition of a *work* that is based on an earlier *work* with augmented or new intellectual content
5. **Free Translation:** an approach to translating a text, which intentionally recognizes the cultural gap between the “intellectual world of the author and that of the translator” (Pellizzi, 2015, p. 10); it modifies parts of the original language text, so that it appeals differently to the audience of the target language text.
6. **Edited Series:** by default every new *expression* of an edited series with new intellectual content will become a new *work*, even if the title of the edited series remains the same.
7. **Review:** a focused piece of *work* written by a new author to describe and review the intellectual content of the original, emblematic *work* or one of its *expressions* (e.g., a book review)
8. **Criticism:** an extensive piece of *work* written by a new author which critically evaluates the intellectual content of the original, emblematic *work* or one of its *expressions* in connection with other similar *works* (e.g., literary criticism)
9. **Commentary:** a *work* that explains and annotates an original *work* (e.g., a commentary on one or more expressions of the Bible).

2.3. The monograph as a complex ‘work’

At present little is known about why certain books are included in the BKCI. Most books that have been indexed have been published in 2005 or after, and there is currently a book-by-book editorial selection process in place at Thomson Reuters (Testa, 2012). One of the goals of Thomson Reuter’s development team is to include books with a relatively high citation impact, yet it will always be unclear as to which particular *item*, was originally used by the citing person(s). The distinct *item* that was used, however, is not important, as long as it has been accurately referenced. This means that all *manifestation* details for an indexed item need to be accurate (i.e., author name, title, ISBN, publication date) so that a decision can be made as to which expressions are equivalent and which shall be characterized as new work. This is one of the key recommendations of FRBR, and thus far, it has had some impact already on other bibliographic structures like OCLC-WorldCat (see Bennett et al., 2003). According to Bennett et al. (2003), “the majority of benefits associated with applying the FRBR” may be “obtained by concentrating on a relatively small number of complex works” (p. 45).

Figure 2, below, illustrates what is meant by the term “complex work”. The example that we use is a monograph that was first written and published in Dutch, titled *De Vergeten Wetenschappen: Een Geschiedenis van de Humaniora* (Bod, 2010). *De Vergeten Wetenschappen* has been reprinted in its first edition language, and has also been ‘translated’ to Polish and Ukrainian (i.e., as two new expressions of the same ‘work’). Note that the 2013 Polish and the 2016 Ukrainian expressions are linked back to the original ‘work’; thus were not (according to international catalogers linked to OCLC-WorldCat) recognized as new works. In OCLC-WorldCat they have been recorded as direct or ‘literal translations’ of the Dutch progenitor, but the latest English expression has not (see Figure 2).

The term ‘literal translation’ generally means that a source language text is rendered to a target language text, while retaining similar meaning and structure of content (Bassnett, 2002). In this sense,

the translation process seems relatively straightforward; however, several factors can influence the exercise. For instance, it may become more complex if there is a deeper focus on the cultural or historical background of the source language text and its author, as well as the target language text and translator (e.g., a free translation). With some freely translated texts, changes are often rooted in the historical period in which the translation was carried out, including the conditions surrounding the translation and the intellectual world of the translator herself. With other translated texts, more emphasis is placed on the reception and influence of the translation on the target language and culture. In simpler terms, an author may have a work re-written by a translator, or she may translate her own work, but the translated *work* can only be recognized later as a new *work* if it includes significant changes.

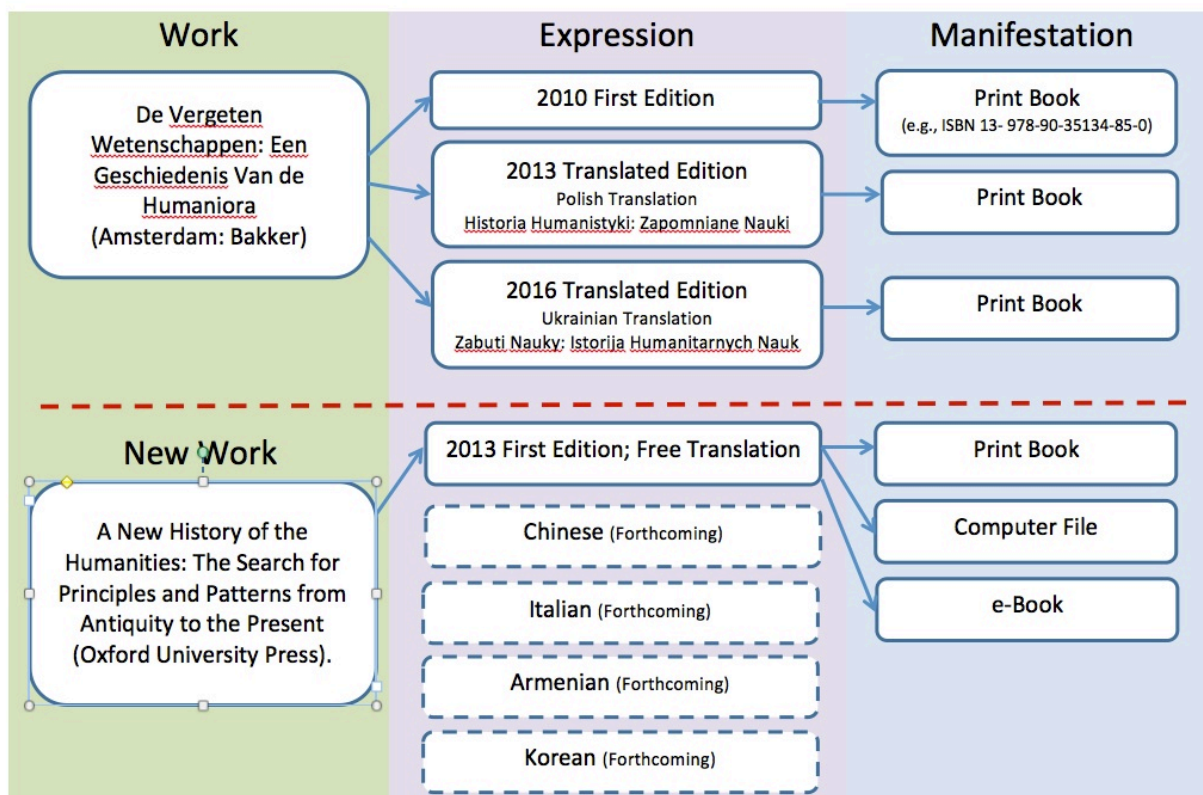


Figure 2. Model of a complex work with expressions and manifestations of a new work

Note from Figure 2 that Rens Bod has translated and published an English derivative of *De Vergeten Wetenschappen*, titled *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present* (Bod, 2013). Again, this English expression, unlike the Dutch-to-Polish expression and Dutch-to-Ukrainian expression has been identified (in WorldCat) as a new work. In an e-mail exchange with the author, we received the following information:

I would consider the English [expression] as a kind of improved edition of the Dutch book. When the Dutch work was translated into English, I sent it to OUP [Oxford University Press] and incorporated the comments by the 5 OUP reviewers into the English version; I also had the book read by an arabist,

indologist and a sinologist, and incorporated their comments as well. And, I added a few additional humanists to the book (e.g. Mabillon)² as well as some additional concepts, such as the Chinese theory of parallel perspective (R. Bod, personal communication, June 16, 2016).

In the future, Rens Bod notes that there will be new expressions of his work in “Chinese (just finished), Italian, Armenian and Korean versions...translated directly from the English version” (R. Bod, personal communication, June 16, 2016).

Clearly one *work* has potential to possess complex family relationships, and in the case of *De Vergeten Wetenschappen*, we see that the bibliographic family is still growing. With many more *works* like this, there are multiple implications for the structural design of the BKCI. Currently “it is possible to distinguish bibliometrically between monographs and edited volumes among the books [in the BKCI]” but according to Leydesdorff and Felt (2012) “monographs may be underrated in terms of citation impact or overrated because individual chapters are counted separately” (p. 1). This problem of over-counting or under-counting pertains solely to the metadata used for indexing monographs in the BKCI and their components as familial entities.

3. Research

3.1 Databases

Our research focuses primarily on the BKCI, but in order to assess its reliability as a data source for bibliometric analyses, we have chosen to compare it to three other catalogs: 1) the Danish PURE repository system for scholarly research outputs, 2) the OCLC-WorldCat, and 3) Goodreads. Each database/catalog was selected for a specific reason.

The Danish PURE repository system is a collection of repositories corresponding to eight universities across Denmark. Each university has created its own PURE database in order to register and maintain records of all scholarly research outputs. Due to this system’s nation-wide adoption, it is often used in conjunction with the performance-based evaluation system in Denmark. As of 2009, all Danish scholars across the country received a mandate to register their scholarly publications in PURE. Each year, performance points are then calculated on the basis of these PURE records and used to determine the amount of leftover government funding to be distributed across departments or research centers (i.e., 25% of the new basic funds, which are 5% of the total basic funding). Monographs are included, and each registration earns a department or research center 5.00 points (level 1 authority publisher) or 8.00 points (level 2 authority publisher) (Giménez-Toledo et al., 2016). The data retrieved for our study was a set of monographs that had been registered in eight University PURE repositories between the years of 2005-2015. Our main reason for working with these PURE repositories was to examine their current indexing quality, and to determine the extent to which books published by Danish scholars have been indexed also in the new BKCI.

² Jean Mabillon was a French Benedictine monk and scholar and Bod (2013) has referred to his *De re diplomatice* (‘On the Science of Diplomatics’) in *A New History of the Humanities*.

The OCLC-WorldCat and Goodreads were also chosen for this study because both catalogs comply to some degree with the FRBR standard. The BKCI does not; hence by matching ISBNs and extracting all related data from these two extra databases, it is possible to assess the extent to which the BKCI is an accurate index of monographs as family-based entities.

3.2. Data retrieval and data curation method

The procedural list below explains how all monograph data for this study were collected (over six months in 2015-2016), integrated and 'curated' into a new database for all research queries:

1. ISBNs from the BKCI (2005-2015) were retrieved and added to a new SQL database. This included monographs only, which had been indexed with the following metadata tags: a) Pubtype=book/books, b) Doctype=book, c) Norm_doctype=book, and d) Role=author (n=16,392).
2. ISBNs from the eight Danish PURE repositories were also retrieved, based on the following indexing tags: a) doc_type=db, b) doc_level=sci, and c) person role=NOT editor, and added to the SQL database (n=8,604)
3. All duplicate ISBNs from both the BKCI and PURE were removed and the two datasets were merged to produce a total of n=24,961 ISBNs (note: only 35 records between the two original lists were duplicates).
4. With OCLC-WorldCat and Goodreads, we used an Application Programming Interface (API) to retrieve additional metadata (e.g., book title, author name, publisher, publication year) matched to our initial list of ISBNs (n=24,961), including all extra related ISBNs (i.e., additional manifestations of the same work).
5. Our final research dataset in the SQL database included a total of n=56,445 unique ISBNs
6. All ISBN-13 numbers were trimmed to create a new numerical ISBN for retrieval purposes (e.g., 978-92-95055-02-5 was reduced to 929505502) and to minimize errors in SQL queries
7. An OCLC-Work-ID was created as a distinct metadata field for a *work* and all of its related ISBNs for the OCLC-WorldCat data.
8. We also created a Goodreads-Work-ID for a *work* and all of its related ISBNs from Goodreads
9. In all cases where there was a relational overlap of the same *work* in Goodreads and/or OCLC-WorldCat, we created a Final-Work-ID. This enabled us to identify the most comprehensive relational overview of one *work*. If a particular ISBN was not identified at all in Goodreads or OCLC-WorldCat, the individual item was given its own Final-Work-ID.
10. A final Expression-ID was created for each *work* based on the following rules. First, if a *manifestation* (of a book) was published in the same year and in the same language then it was identified as being the same *expression*. If a *manifestation* (of a book) was published in the same year but in a different language, then it was categorized as being a different *expression*. The last part of the Expression-ID was designed to show the number of *manifestations* related to one *expression*.

Table 1: Sample list of related ISBNs extracted from the BKCI, OCLC-WorldCat, and Goodreads.

	Manifestations of a work		Found in the Book Citation Index	Monograph Title	PubYear	Language	Expression-ID based on all ISBNs with PubYear and Language metadata		Unambiguously determined editions of the same work
	ISBN#	Numeric-ISBN#					BKCI	Expression-ID	
1	978-03-33257-16-6	033325716		Manias, panics, and crashes : a history of financial crises	1978	eng	64914_36719_3	1	1st Edition
2	978-04-65043-80-4	046504380		Manias, panics, and crashes : a history of financial crises	1978	eng	64914_36719_3	1	1st Edition
3	978-04-65044-02-3	046504402		Manias, panics, and crashes : a history of financial crises	1978	eng	64914_36719_3	1	1st Edition
4	978-03-33521-89-2	033352189		Manias, panics and crashes : a history of financial crises	1989	eng	64914_36720_2	2	2nd Edition
5	978-04-65044-03-0	046504403		Manias, panics, and crashes : a history of financial crises	1989	eng	64914_36720_2	2	2nd Edition
6	978-04-65044-04-7	046504404		Manias, panics, and crashes : a history of financial crises	1995	und	64914_36721_1	3	PubYear Error?
7	978-03-33670-40-8	033367040		Manias, panics and crashes : a history of financial crises	1996	eng	64914_35694_1	4	3rd Edition
8	978-04-71161-71-4	047116171		Manias, panics and crashes : A history of financial crisis	1997	eng	64914_36887_1	5	PubYear Error?
9	978-04-71389-45-3	047138945		Manias, panics, and crashes : a history of financial crises	2000	eng	64914_36891_1	6	4th Edition
10	6613839752	661383975		Manias, Panics and Crashes	2001				
11	978-03-33970-29-4	033397029		Manias, panics and crashes : a history of financial crises	2002	eng	64914_35724_1	7	PubYear Error?
12	978-14-03936-51-6	140393651	1	Manias, panics and crashes : a history of financial crises	2005	eng	64914_36892_2	8	5th Edition
13	978-04-71467-14-4	047146714		Manias, panics, and crashes : a history of financial crises	2005	eng	64914_36892_2	8	5th Edition
14	6610818096	661081809		Manias, Panics and Crashes	2005				
15	978-02-30575-96-7	023057596		Manias, panics and crashes : a history of financial crises.	2009	eng	64914_34677_2	9	PubYear Error?
16	978-02-30575-97-4	023057597		Manias, panics and crashes : a history of financial crises.	2009	eng	64914_34677_2	9	PubYear Error?
17	1280818093	128081809		Manias, Panics and Crashes: A History of Financial Crises (Revised)	2010				
18	978-02-30365-35-3	023036535		Manias, panics and crashes : a History of Financial Crises	2011	eng	64914_59531_2	10	6th Edition
19	1449886922	144988692		Manias, Panics, and Crashes: A History of Financial Crises	2011	eng	64914_59531_2	10	6th Edition
20	978-02-30365-38-4	023036538	1	MANIAS, PANICS AND CRASHES: A HISTORY OF FINANCIAL CRISES, 6TH EDITION	2011				
21	8502210807	850221080		Manias, Panicos e Crises: A Historia das Catastrofes Economicas Mundiais	2013	por	64914_64914_1	11	Translated Edition
22	1283527308	128352730		Manias, Panics and Crashes	2014				
23	1-137525754	113752575		Manias, Panics, and Crashes: A History of Financial Crises	2015	eng	64914_48167_1	12	7th Edition
24	1440744807	144074480		Manias, Panics, and Crashes: A History of Financial Crises					

Table 1, above, presents a sample list of related ISBNs from our final dataset. Note that we have retrieved 24 unique ISBNs (i.e., physical manifestations) of the same work titled *Manias, Panics, and Crashes*. In the BKCI, we found two ISBNs (i.e., 978-14-03936-51-6 and 978-02-30365-38-4), and both were indexed separately as distinct entities rather than two *manifestations* linked together to represent the same *work*. The ISBN at line 20 was not found in any other database except the BKCI; thus according to our methodology (point 9) we have counted it as a separate work. With the OCLC-WorldCat we found 14 more unique ISBNs related to this title, and with Goodreads, we found an additional 8 ISBNs.


Note also, that out of the 24 unique ISBNs for *Manias, Panics, and Crashes: A History of Financial Crises* 18 could be categorized with a full Expression-ID, based on the rules that we developed for identifying expressions and the metadata available to confirm these rules. Only 12 were labeled formally as unique expressions. In the end, only seven editions could be unambiguously identified when we carried out a search using Google for all expressions related to each publication year. This means that some of the publication years may have been recorded in error.

While OCLC-WorldCat tends to be a more reliable database for retrieving complete metadata, Goodreads tended to support the retrieval of more unique ISBNs. OCLC-WorldCat was particularly useful for identifying ‘expressions’ of a particular work due to its regular indexing of publication year and language. Goodreads, on the other hand, supported a much better understanding of the relationship between the ISBNs because, unlike OCLC-WorldCat, all ISBNs were united under one work-related metadata tag.

3.3. Results

Table 2 indicates the results of our data crawling and matching procedure beginning with two original datasets – 1) an ISBN list from the BKCI, and 2) an ISBN list from the Danish Pure Repository. In relation to the 16,392 ISBNs retrieved from the BKCI, an extra 30,903 ISBNs (65% more) were found following the API procedures with OCLC-WorldCat and Goodreads. With the Danish PURE repository only a few extra related ISBNs (19%) were found using the APIs.

Table 2. ISBN matching and retrieval results for total manifestations, expressions, and works.

	The Book Citation Index		The Danish Pure repository
1 Number of ISBNs crawled	16,392 (35%)		8,604 (81%)
2 Number of overlapping ISBNs		35 (0.41%)	
3 Extra related ISBNs found in OCLC-WorldCat and Goodreads	30,903 (65%)		2,042 (19%)
4 Total unique ISBNs in the dataset under study	47,295 (100%)		10,646 (100%)
5 ISBNs with distinct language and publication year	34,236		8,362
6 Total Expressions	20,284		7,844
7 Total Works	16,311		8,195

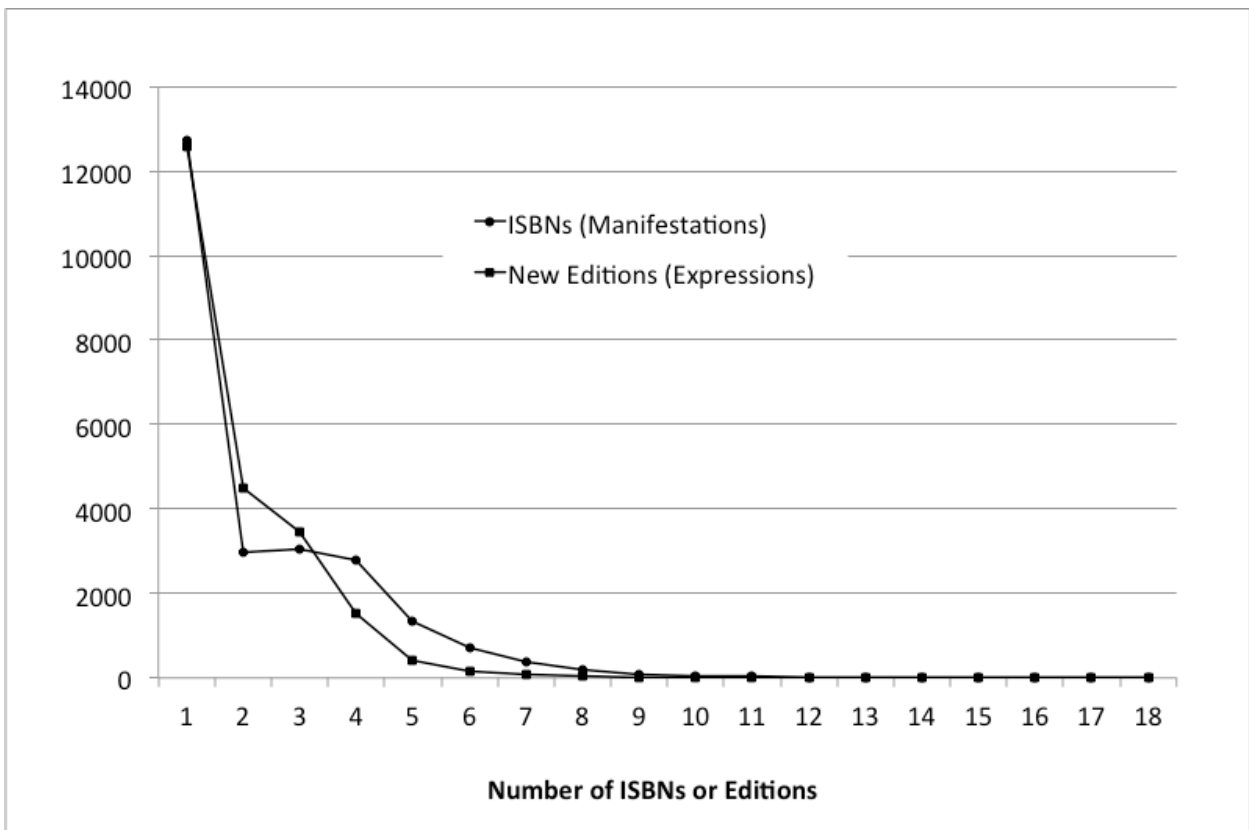


Figure 3. Frequency distribution of works with one or more ISBNs and published as one or more edition.

Figure 3 indicates the number of works from the full dataset with one or more ISBNs (i.e., physical manifestations), including those that had been published as one or more edition (i.e., distinct expressions). Although a little more than half (52%; n=12,723) were published with only one ISBN, almost half (48%, or n=10,249) could also be identified as having two or more ISBNs. The highest count of ISBNs was a total of n=28 for one work, and the lowest was 1, but on average, a scholarly work is likely to be published as two editions, each with approximately 3 different ISBNs.

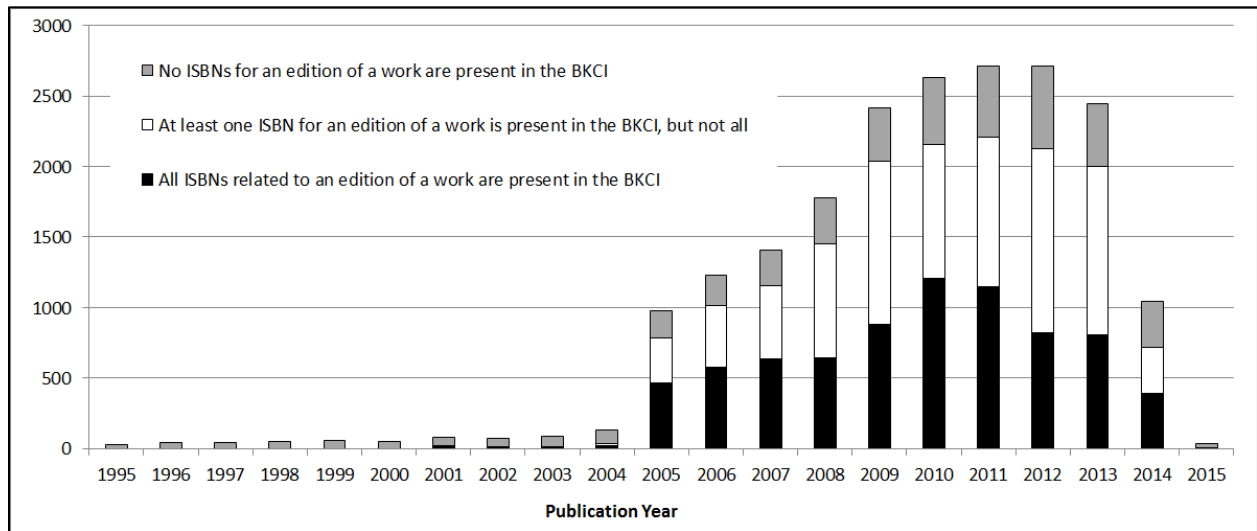


Figure 4. Indexing quality of the BKCI based on ISBNs per edition for publication years 1995-2015.

Figure 4 presents the indexing quality of the BKCI pertaining to editions or expressions of a work published in 1995, up to and including 2015. We selected this time frame because 98% of the ISBNs originally retrieved from the BKCI were for at least one edition of a work that had been published between these years. The black portion of each column per year indicates that all ISBNs related to an edition of a work are present in the BKCI. With the ISBN as the counted variable, this means that several works in their entirety have been accurately indexed. The white part of the column indicates that there is at least one ISBN indexed for a particular edition of a work, but that ISBNs for additional family-related editions are missing. The grey portion at the top of each column then represents all of these missing editions, which were confirmed to exist based on data matching with Goodreads and OCLC-WorldCat, but were not recorded in the BKCI.

Note that for the publication year of 2005, most editions (i.e., *expressions*) had been fully indexed in the BKCI, as shown by the proportionally longer black column. For the publication year of 2009, more editions in general were added to the BKCI, but a full indexing of each edition (i.e., *expression*) and related ISBN seems to decrease, as shown by the proportionally longer white column. Again, the gray column indicates the proportion of editions that have no representation in the BKCI. For the publication year of 2010 and onward there is no real observable pattern other than the fact that the indexing quality for all editions (*expressions*) has remained inconsistent.

To illustrate this indexing problem more clearly, we refer back to the sample title list shown in Table 1. From this table, note that both the fifth and sixth editions of *Manias, Panics, and Crashes: A History of Financial Crises* had been indexed in the BKCI, but all earlier editions published (or printed) in the years 1978, 1989, 1996 and 2000, each with their own related ISBNs, were not added. Overall, what we found is that for all of the monographs originally identified with ISBNs in the BKCI, approximately 21% of their related editions (or expressions) were not represented.

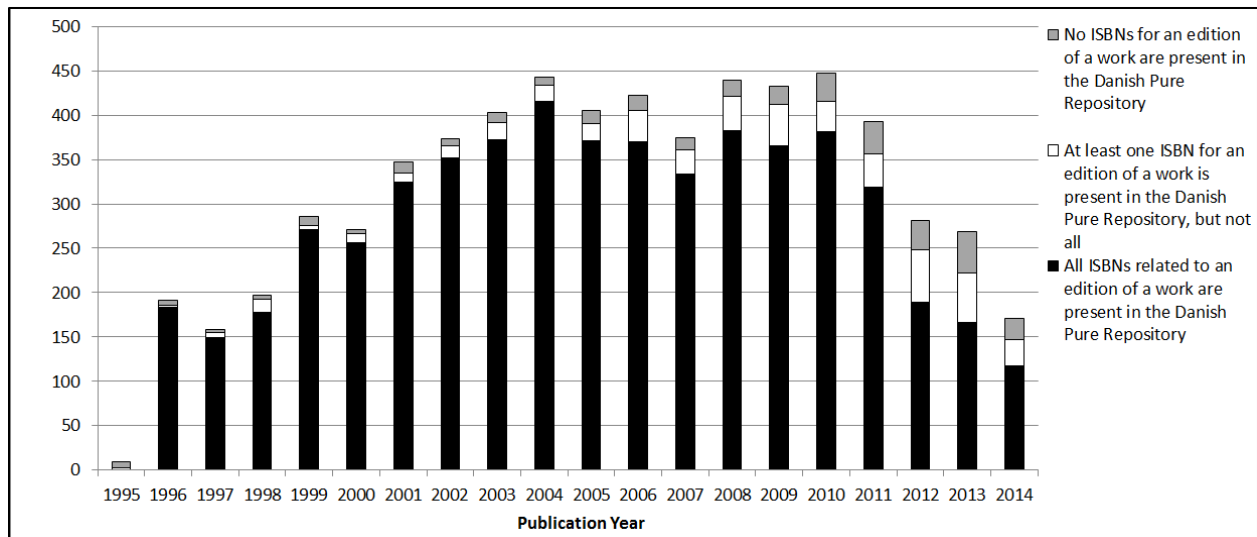


Figure 5. Indexing quality of the Danish PURE repository system based on ISBNs per edition for publication years 1995-2015.

Figure 5 shows the same information shown in Figure 4, but this time for the Danish PURE repository system. Here the indexing quality for editions per work tends to be much better. Note also that most of the works that had been registered in PURE do not have more than one associated ISBN (as shown by the black and white portion of the columns). There could be two reasons for this: 1) many works were never published or reprinted again as second or third editions with new ISBNs, or 2) the Danish author decided to only register his/her work under a single ISBN. Also, if an author had been responsible for producing and publishing both a Danish and English edition of a work, both would have had to be indexed. For some works identified as having a non-indexed edition (i.e., the proportionally smaller grey bars), we found that only a Danish edition of a work was registered, but not the original language one. If the Danish author-as-translator did not produce the original language edition; he or she would not have been required to register this in PURE.

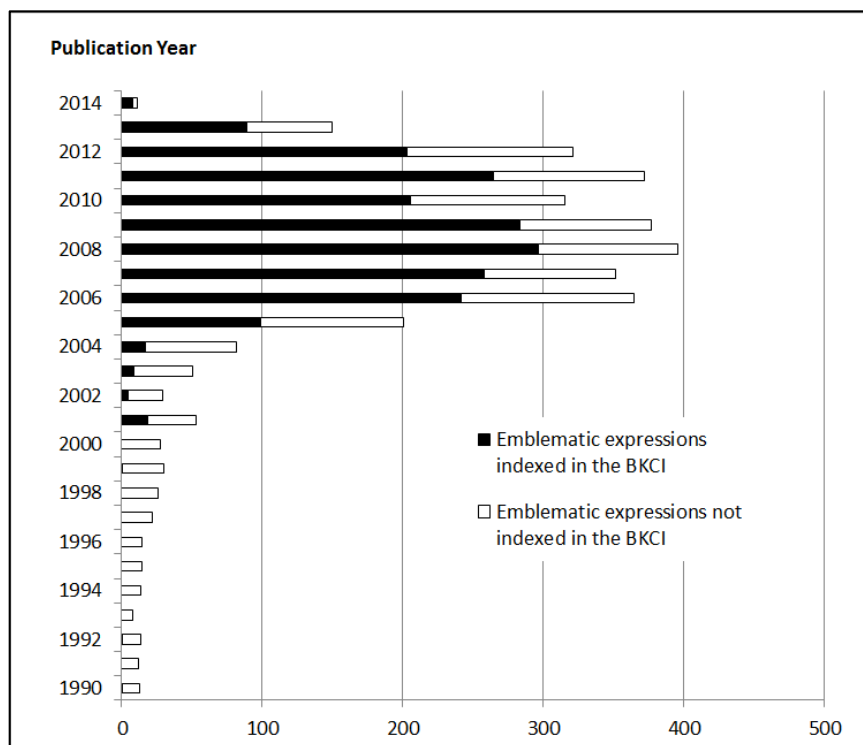


Figure 6. Indexing of emblematic expressions of a work in the BKCI based on ISBNs per edition for publication years 1995-2015.

Figure 6 illustrates the extent to which emblematic expressions were indexed in the BKCI for the publication years of 1990 up to and including 2014. For all works with more than one edition (i.e. expression) in our dataset ($n=10,731$) we were able to identify a total of $n=3,370$ that were emblematic. Again, the emblematic edition or expression is the first publication and printing of a work as an original intellectual contribution. According to our data, approximately 40% of these emblematic expressions had not been indexed, even though they are represented in the BKCI in the form of later editions.

4. Discussion: Metrics for Monograph ‘Families’

With the Book Citation Index currently as it is, counting citations to monographs is problematic; hence a discussion is needed both in light of FRBR standards and our study results. While many similar problems apply to edited books, here we will focus strictly on monographs.

One of the data accuracy problems related to the BKCI stems directly from the referencing practices of researchers. With the BKCI structured as it is now “monographs may be underrated in terms of citation impact or overrated because individual chapters are counted separately” (Leydesdorff and Felt, 2012, p. 1). For instance, if a scholar who writes a research paper refers repeatedly to a specific chapter, (s)he may choose to cite only that chapter. If the scholar refers to several chapters from the same monograph, (s)he may choose to cite the full monograph. There is no rule regarding this practice, but different research associations often set guidelines. According to the Publication Manual of the

American Psychological Association (2016), referencing a chapter from a monograph is in fact not recommended (note: only a chapter from an edited book), yet there are instances in the BKCI where this occurs. For example:

- Full Monograph: Moed, H. (2005). *Citation analysis in research evaluation*. Dordrecht, NL: Springer
- Chapter in Monograph: Moed, H. (2005). Assessing social sciences and humanities. In *Citation analysis in research evaluation* (pp. 145-166). Dordrecht, NL: Springer.

If this practice continues, and the BKCI is re-developed to follow FRBR, the problem of citation undercounting would cease to exist. In other words, separate citation counts might still be attributed to the Moed (2005) chapter-based reference as well as the monograph-based to reference, but the implementation of a *work*-related identifier would confirm that the two records are related.

Applying the FRBR standard to the BKCI would, in general, ensure that all *expressions* of a *work* are indexed distinctly with an identification code. This is our first recommendation, and to some degree it has already been accomplished. For instance, currently there are two separate indexed editions of *Manias, Panics, and Crashes* in the BKCI (see Table 4), but not all editions have been indexed (as the data illustrate in Figure 4) and with the two that are present, there is no linking ID that shows they are part of the same *work* or progenitor. For all *expressions* and not just these two, a primary *work* identifier is critical, and will show the extent to which different editions within the BKCI belong to the same bibliographic family. The follow-up effect of this practice is that bibliometricians would also have new options for collecting citation counts at specific family levels. A suggested indexing structure for the BKCI, including levels for citation counting, is outlined in Figure 7.

Note from Figure 7, that a *work* is the highest proposed target entity for a citation count; while all individual *expressions* (editions) constitute the lowest proposed target entity. Each *expression* of *Manias, Panics, and Crashes* has been labeled from #1 to #7 (note: see the same list in Table 2). The first four *expressions* link back to the same *work*, and the last three *expressions* may potentially be indexed as new *work(s)*, as illustrated by the line leading to the box labeled “New Work ID”.

Earlier, we indicated that Bod’s (2013) English translation of *De Vergeten Wetenschap*, newly titled as *A New History of the Humanities*, was said to possess augmented properties that make it identifiable as a new work. With Figure 7, we also show that when the fifth, sixth and seventh editions of *Manias, Panics, and Crashes* were published, C. P. Kindelberger was no longer writing alone, but with R. Z. Aliber as his co-author. For these later editions, particularly the sixth one, a note on Amazon.com indicates that there have been changes to the content: “This highly anticipated sixth edition has been revised to include an in-depth analysis of the first global crisis of the twenty-first century” (Amazon.com, 2016). Sometimes small revisions appearing in a new edition still fit the abstract and intellectual concept of the work as a whole, but because the revisions in this case are substantial, one might apply both a new author and augmented text rationale towards indexing the last three editions of *Manias, Panics, and Crashes* under a new work ID.

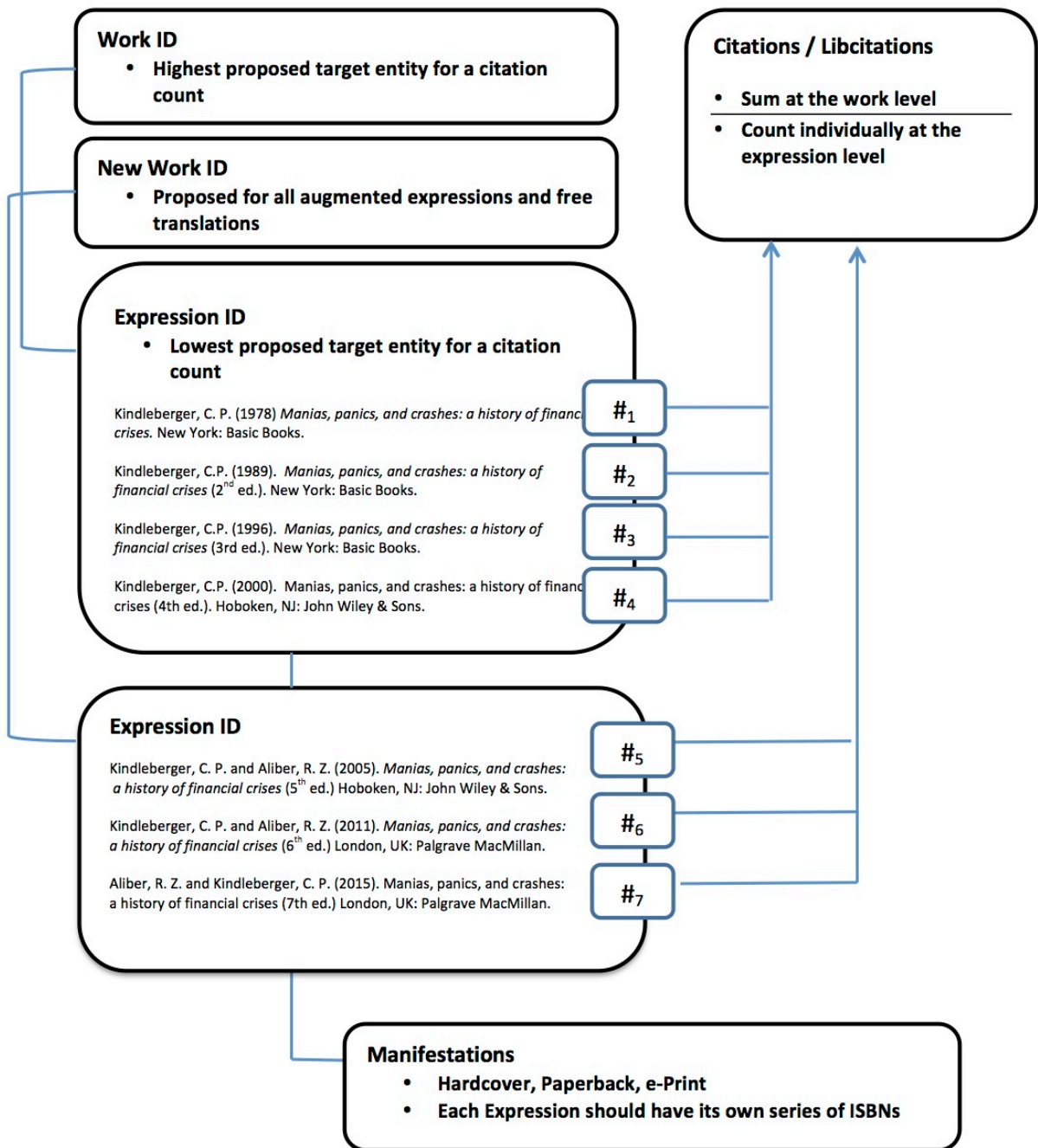


Figure 7. Recommended indexing structure for the BKCI.

At Figure 7, the arrows leading to the box labeled “*citations/libcitations*” illustrate how our proposed indexing structure would support metric assessments books at different bibliographic levels, and for two different types of metric indicators. For example, one could analyze the sum of *citations* given to the first four *expressions* of *Manias, Panics, and Crashes* at #1 to #4 (i.e., the *work* as a whole), or evaluate the individual counts of *citations/libcitations* given to each *expression* at #1 to #4. The same process may be repeated again, for every *expression* indexed as a new work (i.e., #5 to #6). Again, the indexer has little control over the appearance of references in the academic literature, but if most scholars adhere to proper guidelines, a reference should always be given to the correct edition of a monograph used at the time of writing. Figure 7 also illustrates that the two different counting options may be applied to *libcitations* or library holdings for each cataloged edition (e.g., using OCLC-WorldCat).

The value in calculating indicators at different bibliographic levels is that it can help to identify whether or not a specific *expression* or edition of a monograph is receiving more attention than the *work* as a whole. For instance, one specific *expression* of a *work* may be cataloged in libraries, used, referred to, or reviewed more frequently than another. This could be the literal translation of a non-English edition of a work to English, with the new English-language edition potentially having a wider appeal. For some types of translated works, in fact, an author might even have more than one metric profile. At Figure 2, we see how distinct metrics could be calculated for *De Vergeten Wetenschappen* (Bod, 2010) as well as for *A New History of the Humanities* (Bod, 2013). The delineation between new monograph expressions (editions) would also support the identification of associated descriptive works (e.g., book reviews; commentaries). Last but not least, bibliographic levels present better opportunities for bibliometricians to discuss the merits of certain weighting options.

5. Conclusion

The purpose of this study was to investigate the extent to which books currently indexed in the Book Citation Index (BKCI) have adequate metadata and data designed to reflect inherent familial components and relationships. Our research focuses primarily on monographs, and results confirm that some familial components are present in the BKCI, but not all. In terms of ISBNs, many are missing for extra editions of the same *work* and many in particular that need to be indexed are the ISBNs of emblematic (original/first) editions. The purpose of including all ISBNs is to ensure that every physical *manifestation* of a monograph is recognized (e.g., print, paperback, hardcopy, e-print) and that each ISBN is indexed as part of the correct edition or *expression*. This, in turn, ensures that all monograph editions can clearly be identified as being part of the same intellectual contribution, or *work*. Thus, publication counts and citation counts would be more accurate in the BKCI, and new metric indicators could be calculated more effectively.

Part of this research was also designed to compare the indexing of monographs in the BKCI with the Danish PURE repository system. Only a small percentage of books (0.41%) that had been indexed in eight Danish university PURE databases were also present in the BKCI. The BKCI is therefore not a reliable or accurate tool for citation-based evaluations of Danish scholars who mainly publish books.

At present, the Danish evaluation system does not focus on citations, or citation-based approaches to evaluation. However, indexing problems still point to some drawbacks related to the PURE system when taking a performance-based approach. If monographs continue to be indexed without recognizing that they are family-based entities, a few problems might arise. For example, if co-authoring scholars from two different Danish universities register two manifestations of the same work differently in PURE,

this could result in a single BFI point given towards each university department. Normally, if two scholars are responsible for the same *work*, each department should actually receive a *fractionalized* BFI point for the shared contribution. Until it is clear whether or not FRBR might be applied to the PURE system, the Ministry of Higher Education and Science in Denmark is at least making an effort to improve upon the accuracy of book registrations, by producing and publishing a set of document registration guidelines (Uddannelses-og Forskingsministeriet, 26 January 2017).

6. References

- Adams, J., & Testa, J. (2011). Thomson Reuters Book Citation Index. In E. Noyons, P. Ngulube & J. Leta (Eds.), *The 13th Conference of the International Society for Scientometrics and Informetrics* (Vol. I, pp. 13-18). Durban, South Africa: ISSI, Leiden University and the University of Zululand.
- Amazon.com. (2016). *Manias, panics and crashes: a history of financial crises*, sixth edition, paperback – September 27, 2011. Retrieved October 10, 2016 from <https://www.amazon.com/Manias-Panics-Crashes-History-Financial/dp/0230365353>.
- American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC.
- Bassnett, S. (2002). *Translation studies*. London: Routledge.
- Bennett, R., Lavoie, B. F., & O'neill, E. T. (2003). The concept of a work in WorldCat: an application of FRBR. *Library Collections, Acquisitions, and Technical Services*, 27(1), 45-59.
- Bod, R. (2010). *De vergeten wetenschappen. Een geschiedenis van de humaniora*. Amsterdam: Bert Bakker.
- Bod, R. (2013). *A new History of the humanities. The search for principles and patterns from antiquity to the present*. Oxford, UK: Oxford University Press.
- Chi, P., Thijs, B., & Glänzel, W. (2015). The challenges to embody a new data source: The Book Citation Index. *ISSI Newsletter*, 11(1), 24-29.
- Giménez-Toledo, E., Manana-Rodríguez, J., Engels, T. C. E., Ingwersen, P., Pölönen, J., Sivertsen, G., Verleysen, F.T. and Zuccala, A. A. (2016). Taking Scholarly Books into Account. Current Developments in Five European Countries. *Scientometrics*, 107(2), 685-699.
- Gorraiz, J., Purnell, P., & Glänzel, W. (2013). Opportunities and limitations of the book citation index. *Journal of the Association for Information Science and Technology*, 64(7), 1388–1398.
- Hammarfelt, B., (2011). Interdisciplinarity and the intellectual base of literature studies: citation analysis of highly cited monographs. *Scientometrics*, 86(3), 705-725.
- Haustein, S., Bowman, T.D., & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. In Salah, A.A., Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Eds.), *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*, Istanbul, Turkey, 29 June to 3 July, 2015, (pp. 1170 - 1179). Bogaziçi University Printhouse.

- IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional requirements for bibliographic records, final report. UBCIM Publications New Series, Vol. 19.* Munchen: K.G. Saur. Retrieved February 15, 2017 from <http://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf>.
- Kindleberger, C. P. (1978) *Manias, panics, and crashes: a history of financial crises.* New York: Basic Books.
- Kousha, K. & Thelwall, M. (2009). Google book citation for assessing invisible impact? *Journal of the American Society for Information Science and Technology*, 60(8), 1537-1549.
- Kousha, K. & Thelwall, M. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62(11), 2147-2164.
- Kousha, K., Thelwall, M. & Abdoli, M. (2016, in press). Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology*. Retrieved November 1, 2016 from <https://wlv.openrepository.com/wlv/handle/2436/619162>.
- Leydesdorff, L. & Felt, U. (2012). "Books" and "book chapters" in the book citation index (BKCI) and science citation index (SCI, SoSCI, A&HCI). *Proceedings of the American Society for Information Science and Technology*, 49(1), 1-7. [DOI: 10.1002/meet.14504901027]
- Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library holdings, and productivity measures, *Scientometrics*, 83(2), 337-354.
- Lubetzky, S. (1953). Development of cataloging rules. *Library Trends*, 2(2), 179-186.
- Moed, H. (2005). *Citation analysis in research evaluation.* Dordrecht, NL: Springer
- O'Neill, E. T., & Vizine-Goetz, D. (1989). Bibliographic Relationships: Implications for the Function of the Catalog. In E. Svenonius (Ed.), *The Conceptual Foundations of Descriptive Cataloging* (pp. 167-179). San Diego: Academic Press.
- Pellizzi, F. (2015). Art historical and anthropological translation: Some notes and recollections. *Art in Translation*, 4(1), 9-16.
- Smiraglia, R. P. (2001). *The nature of a work: implications for the organization of knowledge.* London: The Scarecrow Press, Inc.
- Testa, J. (2012). *The book selection process for the Book Citation Index in Web of Science.* Thomson Reuters. Retrieved June 24, 2016 from http://wokinfo.com/media/pdf/BKCI-SelectionEssay_web.pdf
- Tillet, B. (2001). Bibliographic relationships. In C. A. Bean and R. Green (eds.) *Relationships in the Organization of Knowledge* (pp. 19-35). Dordrecht: Kluwer Academic Publishers.
- Tillet, B. (2005). What is FRBR? A conceptual model for the bibliographic universe. *The Australian Library Journal*, 54(1), 24-30. DOI: 10.1080/00049670.2005.10721710.

- Torres-Salinas, D. & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics*, 3(1), 9–26.
- Torres-Salinas, D., Robinson-Garcia, N., Cabezas-Clavijo, A. & Jimenez-Contreras, E. (2014). Analyzing the citation characteristics of books: edited books, book series and publisher types in the Book Citation Index. *Scientometrics*, 98(3), 2113–2127.
- Uddannelses-og Forskingsministeriet. (26 January 2017). Retrieved March 10 2017 from <http://ufm.dk/forskning-og-innovation/statistik-og-analyser/den-bibliometriske-forskningsindikator/filer/retningslinjer-for-forskningsregistrering-til-bfi-v-1-0.pdf>
- White, H., Boell, S.K, Yu, H., Davis, M., Wilson, C.S. and Cole, F.T.H. (2009). Libcitations: a measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, 60(6), 1083-1096.
- Zuccala, A. & Cornacchia, R. (2016). Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics*, 108(1), 465-484.
- Zuccala, A., Guns, R., Cornacchia, R., & Bod, R. (2014). Can we rank scholarly book publishers? A bibliometric experiment with the field of history. *Journal of the Association for Information Science and Technology*, 66(7), 1333-1347.
- Zuccala, A. A., Verleysen, F., Cornacchia, R., & Engels, T. (2015a). Altmetrics for the Humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Proceedings*, 67(3). <http://dx.doi.org/10.1108/AJIM-11-2014-0152>
- Zuccala, A. A., & White, H. D. (2015b). Correlating libcitations and citations in the humanities with WorldCat.org and Scopus data. In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of the 15th International Society for Scientometrics and Informetrics Conference*, Istanbul, Turkey, 29 June to 3 July, 2015, (pp. 305-316). Bogazici University Printhouse.