QSS

an open access journal

Citation: Nguyen, B. X., Luczak-Roesch, M., Dinneen, J. D., & Larivière, V. (2022). Assessing the quality of bibliographic data sources for measuring international research collaboration. *Quantitative Science Studies*, *3*(3), 529–559. https://doi.org/10.1162/qss_a_00211

DOI:
https://doi.org/10.1162/qss_a_00211

Peer Review:
https://publons.com/publon/10.1162/qss_a_00211

Supporting Information:
https://doi.org/10.1162/qss_a_00211

Received: 8 September 2021
Accepted: 5 July 2022

Corresponding Author:
Ba Xuan Nguyen
ba.nguyen@vuw.ac.nz

Handling Editor:
Ludo Waltman

The MIT Press

RESEARCH ARTICLE

Assessing the quality of bibliographic data sources for measuring international research collaboration

Ba Xuan Nguyen[1,2], Markus Luczak-Roesch[1,3], Jesse David Dinneen[4], and Vincent Larivière[5]

[1]School of Information Management, Victoria University of Wellington, Wellington, New Zealand
[2]Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam
[3]Te Pūnaha Matatini, Aotearoa New Zealand's Centre of Research Excellence for Complex Systems, Auckland, New Zealand
[4]School of Library and Information Science, Humboldt-Universität zu Berlin, Berlin, Germany
[5]École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec, Canada

Keywords: bibliographic data sources, data quality assessment, data quality dimensions, data quality metrics, international research collaboration measurement

ABSTRACT

Measuring international research collaboration (IRC) is essential to various research assessment tasks but the effect of various measurement decisions, including which data sources to use, has not been thoroughly studied. To better understand the effect of data source choice on IRC measurement, we design and implement a data quality assessment framework specifically for bibliographic data by reviewing and selecting available dimensions and designing appropriate computable metrics, and then validate the framework by applying it to four popular sources of bibliographic data: Microsoft Academic Graph, Web of Science (WoS), Dimensions, and the ACM Digital Library. Successful validation of the framework suggests it is consistent with the popular conceptual framework of information quality proposed by Wang and Strong (1996) and adequately identifies the differences in quality in the sources examined. Application of the framework reveals that WoS has the highest overall quality among the sets considered; and that the differences in quality can be explained primarily by how the data sources are organized. Our study comprises a methodological contribution that enables researchers to apply this IRC measurement tool in their studies and makes an empirical contribution by further characterizing four popular sources of bibliographic data and their impact on IRC measurement.

1. INTRODUCTION

As collaboration across national borders promises advantages of shared resources and knowledge between nations (Wagner, 2005), many governments have an interest in encouraging international research collaboration (IRC) through their science policy (Peters, 2006). Because of that, it is essential to examine the productivity and impact of IRC between countries (Zhou, Zhong, & Yu, 2013). However, developing measurements of IRC activities is a topic that has not been given much attention in bibliometrics scholarship (Chen, Zhang, & Fu, 2019).

The most common indicator for IRC mentioned in bibliometric studies is coauthorship (Aksnes, Piro, & Rørstad, 2019), which is often obtained from bibliographic data sources (Nguyen, Luczak-Roesch et al., 2022). As credible data, together with appropriate models, are the two main contributors to the precise findings of an empirical study (Heckman, 2005), the quality of bibliographic data sources used in measuring IRC is essential.

The quality of bibliographic data sources cannot be evaluated until the definition of data quality (DQ) has been well described for the particular task executed on bibliographic data. In the literature, DQ has commonly been defined as "fitness for use" (Wang & Strong, 1996). The definition "fitness for use" implies that an aspect of DQ considered essential for one task may not be appropriate for another task. For instance, consistency is argued to play an important role in judging patents' validity (Burke & Reitzig, 2007) but accuracy is considered a core dimension of data quality in citation analysis (Olensky, 2015). (It should be noted that dimensions refer to the aspects of DQ or sets of DQ attributes in DQ studies). Therefore, we should make clear what the key dimensions are in bibliographic studies that measure IRC. Furthermore, one dimension of DQ may have different definitions and corresponding metrics to measure it. The case of currency and timeliness dimensions is an example. Some studies consider them separate dimensions, while others treat currency as timeliness (Zaveri, Rula et al., 2016). These varieties lead to the need to establish a dedicated DQ framework for the specific case of IRC measurement from bibliographic data.

In addition to these theoretical considerations DQ quality related to IRC measurement, there are practical research challenges. These challenges arise because there are different bibliographic data sources available to researchers on which IRC may be measured. This list includes multidisciplinary bibliographic data sources (such as Scopus, Web of Science [WoS], Dimensions, Crossref, and Microsoft Academic Graph) and domain-specific data sources (such as PubMed, IEEE Xplore, and ACM DL). These data sources vary in the licensing costs for their use, the range of data, and the intuitive "fitness for use." For example, Microsoft Academic Graph is a multidisciplinary bibliographic data source that can be freely downloaded from the Internet. At the same time, PubMed, a database of references and abstracts on life sciences and biomedical topics, is behind a paywall. It is of interest to researchers to choose the most suitable data source from a wide range of available options for IRC measurement.

This study is an attempt to establish a dedicated DQ framework for IRC measurement. In detail, we address the following main research question and the three subquestions:

- How well are different bibliographic data sources suited to measure International Research Collaboration?

   1. Which dimensions are relevant to a data quality assessment (DQA) framework for IRC measurement?
   2. Which dimensions from the DQA framework reflect differences in the primary data sources for IRC measurement?
   3. How can the DQA framework developed be applied to choose the most suitable data source for IRC measurement?

To answer these above research questions, we developed an instrument for DQA in IRC. We then validated this instrument by using it to assess and compare the DQ of four widely used bibliographic data sets.

Our study contributes to understanding DQ in the IRC measurement domain. We identify a list of possible DQs relevant to reflect data quality for IRC measurement. We also implement a "metadata crosswalk" to see how attributes of bibliographic data sources connect to the Functional Requirements for Bibliographic Records (FRBR) model's constructs. We apply this "metadata crosswalk" to select the relevant DQs for our DQA framework. This approach implies a methodological contribution to the DQ domain. In addition, our study has practical implications. We propose a complete set of computable metrics for each specific DQ in the DQA framework built to evaluate bibliographic data sources. Our DQA framework and its sets of computable metrics provide a baseline for researchers to apply in their own IRC measurement studies. We also prove how to apply our DQA framework to evaluate DQ for IRC measurement and suggest the most suitable data sources from a list of common bibliographic data sources surveyed in the present study (Nguyen, Dinneen, & Luczak-Roesch, 2022).

This paper is structured as follows: First, Section 2 introduces a brief description of the fundamentals and related work of DQA for IRC measurement. We then break down our analyses and results into three distinct parts: design, implementation, and application of a new DQA framework for IRC measurement. Section 3 reports our design of a DQA framework for IRC measurement. Section 4 describes the implementation of the DQA framework being designed. Section 5 explores the application and validation of this DQA framework. Next, Sections 6 and 7 present discussions and limitations of this paper, respectively. The paper ends with conclusions and suggestions for future work in Section 8.

Figure 1 represents the main phases of the process implemented in Sections 3–5.

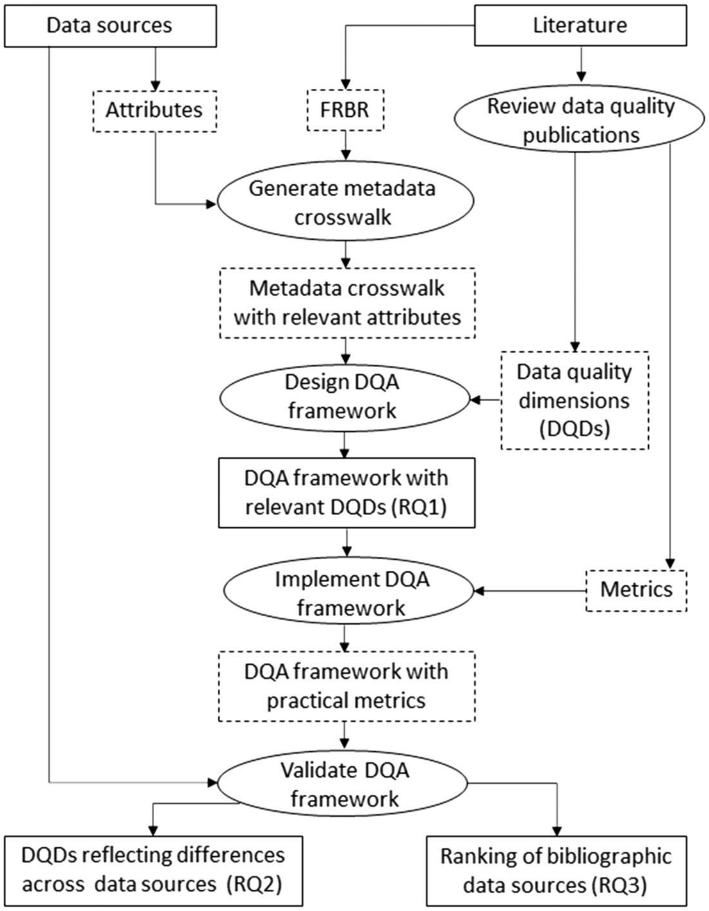
Data sources
Literature
Attributes
FRBR
Review data quality publications
Generate metadata crosswalk
Metadata crosswalk with relevant attributes
Design DQA framework
Data quality dimensions (DQDs)
DQA framework with relevant DQDs (RQ1)
Implement DQA framework
Metrics
DQA framework with practical metrics
Validate DQA framework
DQDs reflecting differences across data sources (RQ2)
Ranking of bibliographic data sources (RQ3)

Figure 1. Overview of the process assessing the quality of bibliographic data sources for IRC measurement.

2. FUNDAMENTALS AND RELATED WORK

2.1. Conceptualization of International Research Collaboration

Within an academic context, the term *collaboration* refers to various activities, including work on a research project undertaken by a team, cooperation between stakeholders from industry and academics, or the collaboration between students and teachers. In general, such collaboration is carried out to share resources, ideas, equipment, and data (Wagner, 2005) between nations, organizations, and individuals.

In the present paper, *international research collaboration* is a construct that refers specifically to scientific research activities between individuals from different countries. Although cross-border cooperation in science started as early as the 19th century (Beaver & Rosen, 1978), international collaboration multiplied after the Second World War and has since become an ever-growing trend following globalization (Beaver, 2001) and facilitated by advanced technology, tools, and workflows (Anuradha & Urs, 2007; Zhou et al., 2013) and government support (Hatakenaka, 2008). Consequently, policy makers need to benchmark and measure IRC over time to assess the impact of pro-IRC policy, initiatives, and support (i.e., to examine how much they have improved their "level" of IRC). Therefore, IRC measurement has become a central focus of IRC research (Chen et al., 2019).

2.2. Bibliometric Approaches to the Measurement of International Research Collaboration

In general, there are various approaches to measure research collaboration. Traditional bibliometrics and other approaches to measuring scholarly and scientific publishing are available, such as altmetrics or webometrics. Among them, traditional bibliometrics is frequently used in studies of research collaboration (Downing, Temane et al., 2021).

IRC measurement in bibliometric studies can be implemented variously. One difference stems from the different ways to operationalize "international" according to different definitions of or ideas about IRC. Studies have, for example, used either authors' listed affiliations or authors' PhD locations and countries of birth. However, using affiliations listed in publications has become the standard and convenient practice for operationalizing the "international" attribute (Chen et al., 2019). The development of international coauthored papers has perhaps reinforced this practice. As publications with international research collaborations, on average, receive a higher number of citations (Glänzel & Schubert, 2001; Schmoch & Schubert, 2008), it seems beneficial for researchers to engage in collaborative research. Over time, both the number and the ratio of multinational publications have been on the rise (Fortunato, Bergstrom et al., 2018).

In IRC measurement, the variety of data sources used to analyze coauthorship may be problematic. Various data sets can be used for coauthorship analysis, available from different sources (e.g., WoS, Google Scholar and nationally funded research projects). However, different data sources may lead to different results (De Stefano, Fuccella et al., 2013). This fact raises the questions of how to evaluate the quality of these data sources for IRC measurement and what criteria should be used to rank them so that IRC can be measured accurately.

2.3. DQ Assessment

There is a considerable body of literature about DQ spread across the fields of management, business, computer science, and information systems (Xiao, Lu et al., 2014), in which DQ is most commonly defined as "fitness for use" (Strong, Lee, & Wang, 1997). Data quality is often operationalized via a framework of data dimensions for measurement, such as the conceptual framework of information quality proposed by Wang and Strong (1996). This framework includes dimensions of DQ considered essential by data consumers, organized into four categories:

1. *Intrinsic DQ*, which includes *Believability*, *Accuracy*, *Objectivity*, and *Reputation*.
2. *Contextual DQ*, which includes *Value-added*, *Relevancy*, *Timeliness*, *Completeness*, and *Appropriate amount of data*.
3. *Representational DQ*, which includes *Interpretability*, *Ease of understanding*, *Representational consistency*, and *Concise representation*.
4. *Accessibility DQ*, which includes *Accessibility,* and *Access security*.

Each dimension of DQ can be measured by a list of specific metrics. For example, the *Completeness* dimension may be measured by relevant subdimensions: *schema completeness*, *property completeness*, and *population completeness*, each with its own implemented metric (Zaveri et al., 2016).

Therefore, the quality of data is evaluated by the process of *DQA*, to examine whether some data meet the consumers' needs in a specific use case (Bizer & Cyganiak, 2009). In this process, each dimension of data is evaluated subjectively or objectively. Subjective DQAs reflect stakeholders' needs and experiences, while in so-called objective assessments, organizations follow a set of principles to develop metrics specific to their needs (Pipino, Lee, & Wang, 2002).

The framework by Wang and Strong (1996) is adopted by the present study as a starting point for designing a DQ assessment framework for bibliographic data because of its empirical generation of DQ categories and its canonical role in DQ assessment (Cichy & Rass, 2019); it is widely cited in DQ literature and has, for example, led to the development of a subgroup of DQ assessment studies (Xiao et al., 2014) that developed further metrics for DQDs such as completeness and relevance (Zhu & Wu, 2011), currency (Heinrich & Klier, 2010), and accuracy (Närman, Holm et al., 2011).

Scientometrics has recently been concerned about the effects of the quality of bibliographic data and altmetrics on their studies (Bornmann & Haunschild, 2018; Strotmann & Zhao, 2015). For example, the accuracy of name disambiguation can change the results of coauthorship network models (Kim, Kim, & Diesner, 2014) and statistical analysis methods of author co-citation analysis (Strotmann & Zhao, 2012). There are also many publicly available data sets for building citation networks that shape the scientific influence (Van Holt, Johnson et al., 2016) so the quality of data is important to scientometrics.

Although there have been studies researching different aspects of bibliographic data sources' DQ, these studies have not examined DQA with possible dimensions thoroughly. These studies have often examined the bibliographic data's quality in two approaches. The first approach is to evaluate a specific dimension of bibliographic sources' DQ quantitatively. Regarding the *completeness* dimension, two major multidisciplinary databases—Scopus and Thomson-Reuters databases (Martín-Martín, Orduna-Malea et al., 2018)—have been explored to assess the extent to which data elements are absent (Jacsó, 2009). The results show that the rate of missing country data is high (e.g., there is a 34% omission rate of country metadata in Scopus and 14% in Thomson-Reuters' WoS). Another example of examining a specific dimension is the study by Sinha, Shen et al. (2015), in which the *accuracy* of the MAG data source is proved to maintain 95% accuracy.

The second approach is to compare various bibliographic data sources for IRC measurement by analyzing specific criteria, such as suitability (Hennemann, Wang, & Liefner, 2011) or coverage (Singh, Singh et al., 2021). Regarding journal coverage, for example, Dimensions had more unique journals than Scopus, and WoS had the least number (Singh et al., 2021).

However, these studies show the differences between bibliographic data sources rather than evaluating them with a relevant DQ framework.

Our literature review confirms that DQA in IRC measurement is an understudied area.

3. DESIGN OF A DQA FRAMEWORK FOR IRC MEASUREMENT

3.1. Objectives

To assess the quality of bibliographic data sources for IRC measurement, we wanted to identify relevant dimensions. For this purpose, two objectives needed to be achieved. First, we wanted to create an inventory of possible data quality dimensions (DQDs). Second, the DQDs identified needed to be assessed for their relevance to IRC measurement. The result was a selection of DQDs that apply to IRC measurement.

3.2. Methods

A systematic review of the literature was conducted to create an inventory of possible DQDs. Specifically, our focus was on the DQDs of bibliographic data sources, and we used Google Scholar as the first tool to retrieve literature. We selected Google Scholar because this tool has been observed to always find more citations for each journal than any others among ResearchGate, WoS, and Scopus (Thelwall & Kousha, 2017). We searched for publications having the terms "bibliographic data" or "bibliographic records" in the title. Furthermore, we then filtered the retrieved articles further for those with additional keywords in their content (the content keywords used were "quality dimension," "data quality," and "quality assessment"). The papers found were initially skimmed to determine whether they discussed DQDs. We then applied the citation pearl-growing method (Harter, 1997) to find relevant sources on this topic. The collection of DQDs discussed in these papers was the inventory of possible DQDs for the present study.

Because, to our knowledge, there are no explicit studies of IRC DQDs, we had to make such a list by assessing the relevance of DQDs for IRC tasks. In detail, we assessed which attributes from each data source are necessary and sufficient for IRC measurement and whether the definition of each dimension (of DQ in general, not just of bibliographic DQ) could be relevant to this task. This approach includes two phases. To begin, we mapped the attributes of the most popular bibliographic data sources to entities in the FRBR model, an entity-relationship model of bibliographic records (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). By mapping attributes of data sources to the corresponding FRBR entities, we could easily compare them and find which attributes were needed for IRC measurement. We then assessed the relevance of each DQD by considering how it could be meaningfully applied to measure IRC using the attributes found.

3.3. Results

3.3.1. Inventory of possible DQDs

Possible data dimensions were gathered from the list of papers found in reviewing the literature. Table S1 (in the Supplementary material) shows these DQDs with their definitions. For each DQD, one definition relevant to bibliographic data, or at least relevant to a broader concept than bibliographic data, was extracted. The earliest definition was chosen when there were many definitions for a data dimension. In the case that many data dimensions had similar definitions across multiple papers, only the data dimension described at first was chosen.

The chosen dimensions from the above table were assessed for their relevance to IRC in the next section.

3.3.2. Relevant DQDs for IRC

In the first phase of listing relevant DQDs for IRC, an existing model of bibliographic resources—Functional Requirements for Bibliographic Resources or FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998)—was applied to the four data sources' attributes (i.e., fields were put into the model's categories) to enable comparison across them. The four data sources examined were Microsoft Academic Graph (MAG)[1], Dimensions (publications)[2], WoS Core Collection, and ACM Digital Library (ACM DL)[3]. The first was a domain-specific resource covering the computing sciences, while the others were considered among the most important bibliographic data sources covering all fields of study (Waltman & Larivière, 2020). The summary of the four data sources is given in Table 1.

The result of the categorization is shown in Table S2. In this table, the two entities *person* and *corporate body* were presented together because the bibliographic data sources discussed here do not always store them separately. For instance, the attribute "Author Address" of WoS might contain information about either the personal authors' home addresses or their affiliations' addresses. Additionally, only general or article attributes are shown for legibility, while proceedings' attributes (e.g., found only in the ACM DL) are omitted.

To measure IRC, the following information was needed: the countries of authors collaborating on a work (e.g., derived from affiliation data), and the date that work was published. Therefore, information about the time of *manifestation* of that work (from now on called *time published*) and the country of (the *corporate body* of) each *person* creating that work (from now on called *countries involved*) had to be presented in bibliographic records for the particular task of IRC measurement. The corresponding attributes (implementing *manifestation*, and implementing *person* and/or *corporate body* of a *work*) in the four data sources being studied were presented together in a "metadata crosswalk" in Table S3. In this table, the necessary attributes related to *time published* or *countries involved* were presented in bold.

As explained above, the second phase of listing relevant DQDs for IRC is assessing the relevance of the chosen DQDs for IRC measurement. In this phase, we assessed how these DQDs could be evaluated with the attributes found. Consequently, there is one functional requirement applied in this phase: The DQs should be evaluated with only the bibliographic data source. Table S4 indicates whether or not the definition of each DQD (being chosen in phase 1) can be meaningfully applied to measure IRC using the attributes found and provides the rationale for each.

After completing the second approach described above, seven DQDs were found to be relevant to IRC studies: *Accuracy*, *Appropriate amount of data*, *Completeness*, *Concise representation*, *Ease of Understanding*, *Relevancy*, and *Representational consistency*. In our objective assessment approach, the relevance criterion was that the DQDs selected could be evaluated by the attributes found in the bibliographic data sources. These seven DQDs were among the 15 most important dimensions to data customers, presented in the conceptual framework of DQ in the study by Wang and Strong (1996). In their study, Wang and Strong (1996) came up with these 15 most important dimensions by asking data consumers to rate the importance of

[1] MAG data were downloaded as a part of OAG v1, which was publicly available from mid-2017 (https://www.microsoft.com/en-us/research/project/open-academic-graph/).

[2] Dimensions data were downloaded via Dimensions API in April 2020 (https://app.dimensions.ai/api/auth).

[3] ACM DL data were retrieved by FTP download in March 2019 (ftp://pubftp.acm.org).

Table 1. Summary of the four data sources under the survey

Features	ACM DL	Dimensions	MAG	WoS
Total works	182,791	116,971,505	166,192,182	54,549,343
Date range	1951–2017	1665–2019	1965–2017	1980–2019

possibly relevant dimensions. Therefore, this fitness suggested a benefit that we could apply these DQDs' importance weights when we used our DQA framework to evaluate the bibliographic data sources (in Section 5 of the present study). Without applying these DQDs' importance weights from Wang and Strong's study, we would have had to repeat the survey ourselves to get the customers' rates, which would have been time-consuming.

Figure 2 shows how these DQDs fit into the conceptual framework of data quality proposed by Wang and Strong (1996).

The other eight DQDs were not selected because they could not be evaluated with the bibliographic data sources' attributes relevant to IRC measurement (as presented in Table S3). In detail, three of them could only be assessed objectively (Accessibility, Security, and Timeliness) but there was insufficient information. The remaining five DQDs (namely Believability, Interpretability, Objectivity, Reputation, and Value Added) could be assessed objectively with external data sources, or be assessed subjectively (e.g., with users' opinions) (Zaveri et al., 2016). For example, Reputation could be evaluated by asking the data users to rate the data sources, or by using available ranking sources. Although adding these remaining five DQDs could add more information for the data sources' evaluation, it would be time-consuming (e.g., doing surveys) or out of scope here (using external ranking data sources would need additional assessment of these data sources' quality as well). Therefore, it was impractical to include the DQDs that could not be evaluated with the bibliographic data sources' attributes and they might be considered in future work.

The seven DQDs that were considered relevant to IRC measurement in this section were operationalized in the next section to implement the DQ assessment.

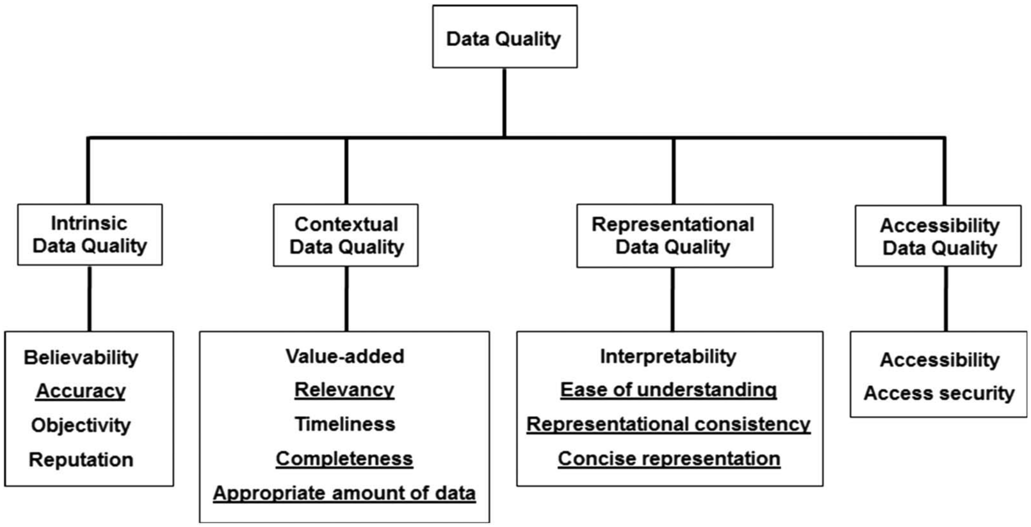

Figure 2. The conceptual framework of information quality, proposed by Wang and Strong (1996), with the seven relevant DQDs (underlined) examined in the present study.

4. IMPLEMENTATION OF THE DQA FRAMEWORK

4.1. Objectives

There were two objectives to operationalize DQDs identified in the prior section. First, we wanted to finalize a list of relevant metrics for each DQD. These metrics had to be practical to implement and appropriate for IRC measurement. Second, each metric's implementation form had to be specified to measure the data sources surveyed.

4.2. Methods

The present study followed two steps for implementing the DQDs framework. First, we listed possible operationalizations for metrics. To do that, we considered the metrics mentioned in papers studying the quality of bibliographic data (identified in Section 2). However, most of these papers did not fully describe the metrics' definitions. Therefore, the descriptions of possible metrics for the above DQDs, reviewed by Zaveri et al. (2016), were used as the initial collection of possible metrics (as displayed in Table S5). In this table, these metrics were also assessed as to whether or not they could be used for IRC measurement. There is one functional requirement that our study applied to choosing metrics for evaluating data sources' quality: The metrics should be computable. Some new metrics relevant to IRC measurement and practical for implementation were also built, from the definitions of relevant dimensions by Wang and Strong (1996). These consist of the *explicitly appropriate amount of data* and *implicitly appropriate amount* (for measuring Appropriate amount of data)*, consistent standard* and *consistent syntax* (for measuring Representational consistency), *separate form of time and country information* (for measuring *Concise representation*), and *presence of relevant vocabularies* and *correct spelling* (for measuring Ease of Understanding). In this step, six of the seven dimensions selected were successfully operationalized by practical metrics.

Second, we specified specific types of operationalization for the metrics. The metrics chosen from the above table were then implemented to measure data quality for the task of IRC measurement. There are three functional (arithmetic) forms among the implementations of the metrics: Simple Ratio, Min or Max Operation, and Weighted Average (Pipino et al., 2002). While Simple Ratio is the measure that shows the ratio of desired outcomes to total outcomes of every single metric, Min or Max Operation and Weighted Average are used to measure the combination of many metrics. Therefore, these above metrics were first implemented in the form of a simple ratio, as presented in Table S6. The two remaining functional forms (i.e., Min or Max Operation and Weighted Average) were considered to be used in comparing different ways to aggregate many metrics of a DQD in Section 5.3.3 of the present study.

4.3. Results

Table S6 shows the metrics' operationalization with the explanation.

To clearly show the dependence between the assessments for metrics and aid in assessing the independence of DQDs across different data sources, we provide mathematical formalizations (Table S7). We define a metric assessment (MA$_x$) as the implementation of a metric on a data source. Therefore, we define a *set* of metric assessments MA{} as

$$\mathrm{MA} = \{[\mathrm{MA}_1, \mathrm{MA}_2, \ldots, \mathrm{MA}_{10}] | \mathrm{MA}_i \text{ is the implementation of a metric } \mathrm{M}_i \text{ listed in Table S6}, \quad i = 1, 2, \ldots, 10\}$$

MA_i is implemented on a data source's sample with m observations. Therefore, a metric assessment MA_i is a set of measurements as

$$\mathrm{MA}_i = \left\{[\mathrm{MA}_{i1}, \mathrm{MA}_{i2}, \ldots, \mathrm{MA}_{im}] | \mathrm{MA}_{ij} \text{ is a measurement on the observation } j, j = 1, 2, \ldots, m\right\}$$

In our study, a measurement MA_{ij} examines whether an observation j satisfies the defined requirement of metric i, as described in Table S5. Consequently, each set of measurements (MA_i) has an unsatisfying measurement $\mathrm{MA}_{i\mathrm{F}}$ subset, which is the set of measurements that return failed results when they are checked on a data source's sample D:

$$\mathrm{MA}_{i\mathrm{F}}(\mathrm{MA}_i, \mathrm{D}) = \left\{[\mathrm{MA}_{i1}, \mathrm{MA}_{i2}, \ldots, \mathrm{MA}_{im}] | \forall \ \mathrm{MA}_{ij} : \mathrm{R}\left(\mathrm{MA}_{ij}, \mathrm{D}\right) = \varnothing, \mathrm{MA}_{ij} \in \mathrm{MA}_i, j = 1, 2, \ldots, m\right\}$$

Sample D has the following specific subsets:

$\mathrm{D_M}$: the set of data points that have missing values, $\mathrm{D_M} \subseteq \mathrm{D}$

$\mathrm{D_E}$: the set of data points that have explicit information of affiliations' nationalities, $\mathrm{D_E} \subseteq (\mathrm{D} - \mathrm{D_M})$

$\mathrm{D_I}$: the set of data points that do not have explicit information of affiliations' nationalities, but their information can implicitly refer to affiliations' nationalities, $\mathrm{D_I} \subseteq (\mathrm{D} - \mathrm{D_M} - \mathrm{D_E})$

We also have two relevant populations used in the measurement assessments:

$\mathrm{P_C}$: the set of possible countries may be included in a set of observations of D. In our study, this population includes all countries in the list ISO 3166 published by the International Organization for Standardization (ISO).

$\mathrm{P_Y}$: the set of possible years may be included in a particular set of observations of D. For instance, we checked the availability of each year in the time coverage from 1980–2017.

Table S7 shows how metric assessments depend on others. For example, the value of $\mathrm{EoU_{Voc}}$ depends on the value of $\mathrm{Com_{Pro}}$. In other words, the more observations satisfy the $\mathrm{EoU_{Voc}}$ measurement, the more observations will be tested with the $\mathrm{Com_{Pro}}$ measurement.

The metrics identified and built in this section were used to assess the independence of DQDs across different data sources and to rank these data sources in the next section.

5. APPLICATION AND VALIDATION OF THE DQA FRAMEWORK

5.1. Objectives

From the prior section, a DQA framework was built with 10 specific metrics for six DQ dimensions. In this section, three consecutive objectives needed to be achieved to illustrate how this DQA framework works for IRC measurement. First, we wanted to obtain the results of operationalizing DQDs to data sources. Such results reflected the data sources' data quality. Second, the independence of DQDs across data sources should be assessed. In other words, we wanted to know whether the results measured by our DQA framework changed according to a particular bibliographic data source used for IRC studies. Third, the data sources' ranking should be gained by applying the developed DQA framework to determine the most suitable data source for IRC measurement.

5.2. Methods

5.2.1. *Method to apply operationalized DQDs to data sources*

To demonstrate the metric framework developed in Section 4, we calculated the 10 selected metrics (of the six selected DQDs) on each of the data sources. In other words, we quantified the data quality of each data source for IRC measurement.

Specifically, 10 metrics (Table S7, "Metric" column) were used to measure the data quality of data sources.

The 10 metrics were calculated as the ratios of data points that satisfied these metrics' definitions to the total data points examined for each metric (Table S6, "Formula" column). It was impractical to do a calculation on the whole data set because the numbers of publications in Dimensions, MAG, and WoS were quite large (Table 1) so it would take several months to calculate the metrics' values. To avoid the length of processing time incurred due to data size, we sampled these data sources for calculation instead. For each of the largest data sources (Dimensions, MAG, and WoS) a sample size of 40 blocks of data, 10,000 data points each, was randomly selected to be used for this purpose. Specifically, the sampling process was done in two steps for each data source. First, the whole set of data points was split into a list of blocks of 10,000 data points each. Second, a random number generator was initiated by the function setseed(0), and then a random sample of size 40 was generated by using the function sample() with replacement. Estimates were made to give the likely ranges for metrics' values of these data sources in the period 1980–2017, while measures on ACM DL (which had 416,439 data points correspondingly) gave the exact metrics' values. Figure 3 shows the distributions of data points in the Dimensions, MAG, and WoS data sources' samples, and the whole ACM DL data source per year. In this figure, the four data sources' distributions had similar temporal trends. The numbers of publications' data points increase over time in general, with a decrease in the last year 2017 of ACM DL, Dimensions, and MAG (possibly due to the incomplete data of this year in these data sources). The only exception in the trends is the case of ACM DL, which shows a sharp decrease in a short period after the year 2000. This exception can be explained as the dot-com bubble crisis' impact on computing research expenditures in the late 1990s.

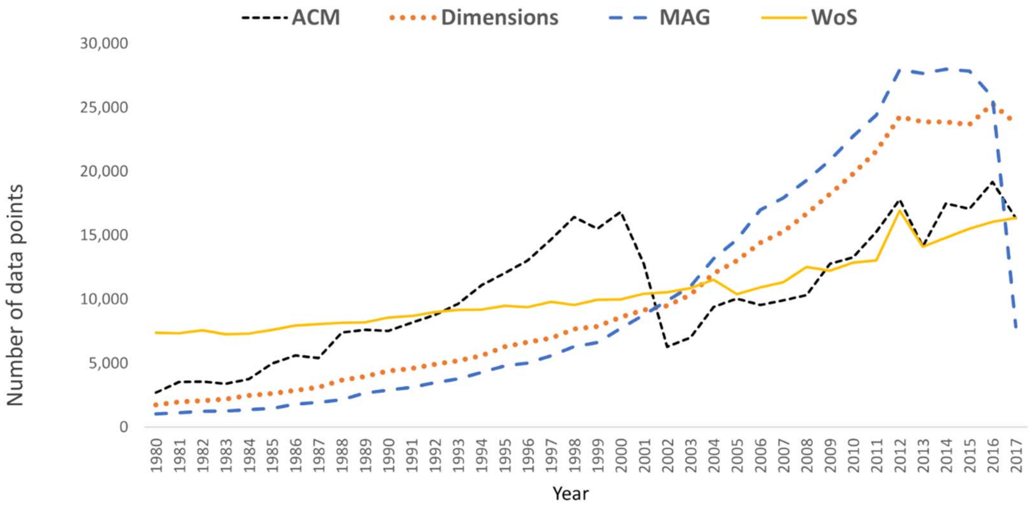

Figure 3. The number of data points (i.e., collaborations; *y*-axis) of four data sets (the plotted lines): all 40 random data samples for each of the three bibliographic data sets (Dimensions, MAG, WoS) and of the whole ACM DL data set, across the years (*x*-axis).

Because data samples were randomly taken from each of WoS, Dimensions, and MAG, we also checked whether these samples are biased samples. For this purpose, we compared the distributions of the three data sources' samples by year. Figure 4 shows that the distributions of data points' years across data blocks in each data set's sample are not notably

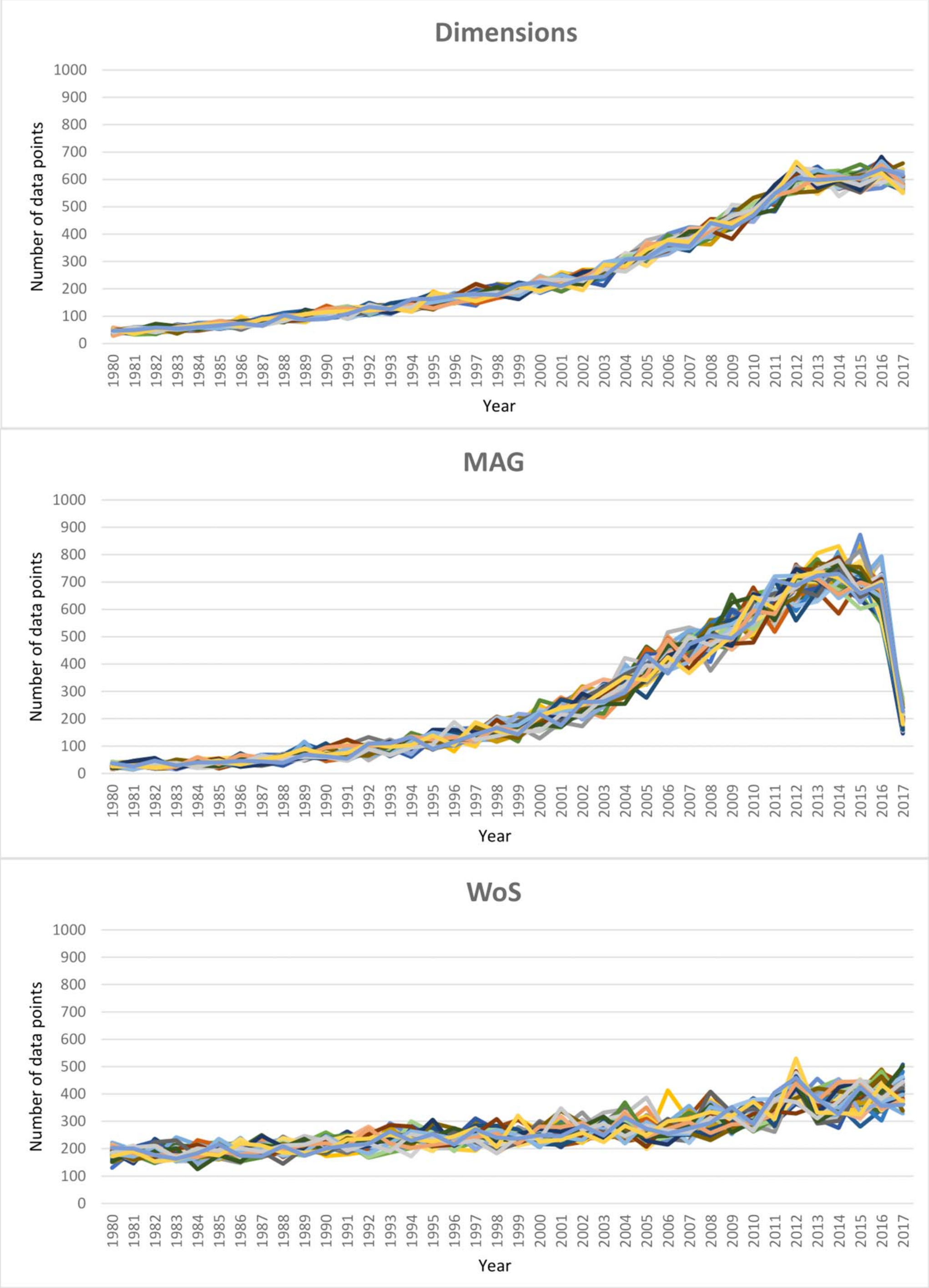

Figure 4. The distribution of the 40 random data samples (lines) across the years (*x*) for each of the three bibliographic data sets; each sample consists of 10,000 data points (i.e., collaborations; *y*).

different (and they are also similar to the corresponding data source's total trend in Figure 3).

As we estimated the values from sampled data blocks, the two types of values for each metric surveyed were evaluated as below.

We first calculated the average values of the 10 selected metrics on the above data sources. The variances of these values among 40 data blocks were also presented to evaluate the average values' spread.

To be more exact, we wanted to find the likely range for the metrics' actual values. We then estimated the confidence interval for each metric measured on Dimensions, MAG, and WoS. According to the central limit theorem, the distribution of either the sum or the mean of a random sample of large size (a sample size of 30 is a rule of thumb for large size) is approximately normal (Anderson, 2010). Because we had a large sample with 40 data blocks (≥ 30), the central limit theorem could be applied in this case. In other words, the average values of each metric measured on randomly selected data blocks are approximately normally distributed (even though the data from which they are sampled is not necessarily normal). As a result, a 95% confidence interval for each metric's average value on the whole data source could be estimated.

5.2.2. Methods to assess the independence of DQDs

We wanted to check how much the results measured by the chosen metrics across different data sources varied. We carried out this activity by doing analyses at the data source level: (a) comparing the metrics' values across data sources, and (b) comparing the correlations between each pair of metrics among data sources. First, we used ANOVA tests to check whether the differences between the averages of metrics across data sources were statistically significant. Second, we compared the correlations between each pair of metrics across data sources. Pearson correlation coefficient with a confidence level of 95% was produced for the metric values of 40 data blocks from each data source.

Considering that ACM DL is a domain-specific bibliographic source in the computing sciences, we also wondered whether the nature of a specific domain could affect the data quality. In other words, we wanted to check at the data source's subset level: (c) whether the metrics' values vary by discipline or not. We assumed that the above metrics worked consistently across different disciplines on data sources, and therefore, the validity of including a domain-specific source in this study was ensured. To access this consistency, we measured their values on subsets of the data sources: Dimensions, MAG, and WoS (ACM DL, meanwhile, contained records of the Computer Science discipline only and was not examined).

The whole process of the above three tests is summarized in Figure 5. In this process, the task of assessing the independence of DQDs included three small steps.

First, we prepared data for calculating the metrics. As it was impractical to calculate the values of metrics for the whole data source, we wanted to calculate on every sample of 40 randomly selected data blocks, each having the size of 10,000 data points for each data source. Similarly, we wanted to calculate the metrics' values on samples of data sources' subsets reflecting different disciplines for Dimensions, MAG, and WoS. Therefore, we needed to separate each of these data sources into many discipline subsets. This task of preparing discipline subsets for each data source was implemented as follows.

We separated Dimensions into subsets by disciplines. The Dimensions data source could be easily split into 22 subsets, using its single-valued "field of research" attribute (*category_for*).

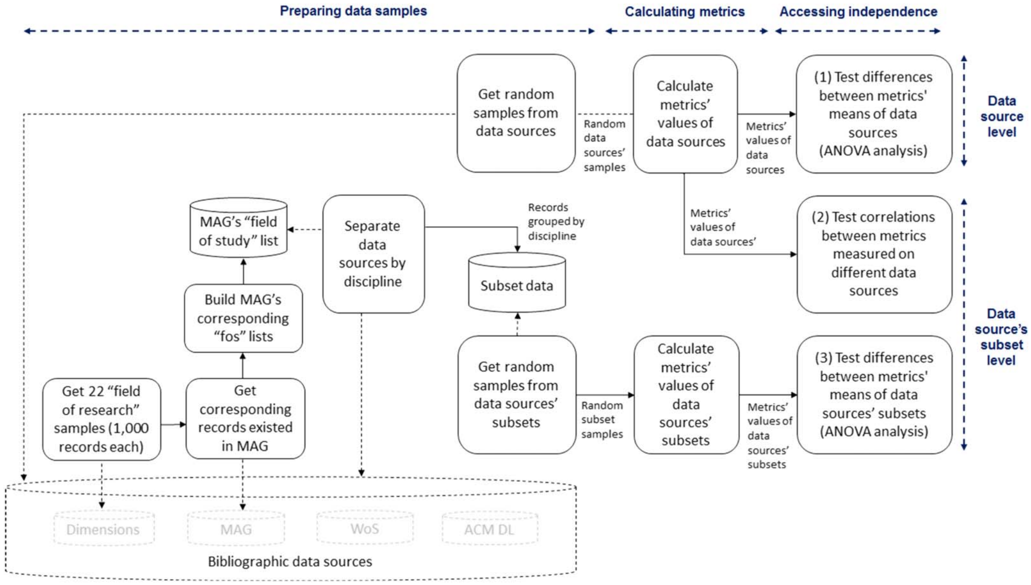
Preparing data samples
Calculating metrics
Accessing independence
Get random samples from data sources
Random data sources' samples
Calculate metrics' values of data sources
Metrics' values of data sources
(1) Test differences between metrics' means of data sources (ANOVA analysis)
Data source level
MAG's "field of study" list
Separate data sources by discipline
Records grouped by discipline
Metrics' values of data sources'
(2) Test correlations between metrics measured on different data sources
Subset data
Build MAG's corresponding "fos" lists
Data source's subset level
Get random samples from data sources' subsets
Random subset samples
Calculate metrics' values of data sources' subsets
Metrics' values of data sources' subsets
(3) Test differences between metrics' means of data sources' subsets (ANOVA analysis)
Get 22 "field of research" samples (1,000 records each)
Get corresponding records existed in MAG
Dimensions
MAG
WoS
ACM DL
Bibliographic data sources

Figure 5. A summary of the process of assessing the independence of DQDs.

We then separated MAG into subsets by disciplines. However, the MAG data source did not appear to have clearly defined categories for disciplines. We realized that MAG did not include such a single-valued attribute for determining a publication's solely discipline as Dimensions did. The "field of study" attribute (*fos*) of MAG, of which values were generated by applying a natural language processing technique (Shen, Ma, & Wang, 2018), could have multiple values. Therefore, we split the MAG data source into 22 corresponding subsets in three steps. First, we took 22 samples, each of 1,000 random papers from each subset of Dimensions. Second, each sample's papers were checked to see whether they existed in the MAG data source or not. If yes, we obtained a list of relevant *fos* values found from matched papers in the MAG data source for each sample. These lists were then filtered with the most relevant values. (Some overlapped *fos* were detected and removed to keep these lists exclusive. For example, the *fos* "Mathematics" appeared in both the lists for "Mathematical Sciences" papers and "Information and Computing Sciences" papers. However, it was kept in the list for the former only.) Third, we organized MAG papers into 22 subsets by comparing their *fos* values with the above 22 lists of separate "fields of studies."

The last data source to separate into disciplines was WoS. This data source could be easily split into five subsets using its "research area" attribute. Because there were multiple research areas assigned to each paper, only the first area for each paper was used. For example, if a paper has the assigned research areas "Business & Economics" and "Women's Studies" we categorized it into the discipline "Business & Economics."

Second, we calculated the metrics' values. These metrics were applied both to different data sources and to different disciplines in each data source.

Third, we assessed the independence of the DQDs in our DQA framework. We applied an ANOVA (single factor) technique to inspect the differences in metrics' values across the four data sources under study. We then tested correlations between the metrics measured on these data sources to check whether there were any notable relationships between every pair of these metrics. We also applied ANOVA to inspect the differences in metrics' values across the subsets of each data source.

5.2.3. Methods to rank data sources using combined measures

A possible method to combine the metrics' values in a DQA framework is to assess the consistency of people's responses across the metrics' uses by an internal consistency reliability test. This test would be useful to examine the extent to which the metrics measure the same DQDs, but it was impractical for this study because surveys would be needed to collect users' evaluations. Therefore, we tried to apply the following alternatives.

In the present study, the values measured by different DQDs were aggregated at two levels: the metric level and the dimension level. First, the values measured by different metrics in a DQD, in general, could be aggregated by applying Min/Max operation or by assigning a Weighted Average. For instance, the *Completeness* dimension could be calculated in the present study by the two metrics: property completeness and population completeness. A conservative approach would be choosing the lowest value among those received from the three metrics mentioned above. However, this approach did not work for every metric in the present study. For example, the *Appropriate amount of data* dimension had two metrics: *Appropriate Data - Explicitly*, and *Appropriate Data - Implicitly*. These two metrics were exclusive because they were measured on two different subsets of each data source. Therefore, this approach was excluded from the present study. Another approach was calculating a weighted average for the two corresponding metrics' values in the Completeness dimension. Because we had no evidence about each metric's contribution to the DQDs applied for IRC measurement, the equally weighted average was chosen. For the example of the *Completeness* dimension, two metrics measuring a DQD were assigned an equal value of 0.5 each.

Second, the aggregated values of DQDs were also combined, either with a set of Equally Weighted Average or a set of weights derived from Wang and Strong (1996). The first set contained equal weights for each of the dimensions. Because we had six dimensions implemented, each was assigned an equal weight of 0.17. The second set applied the importance rating values of DQDs. The corresponding values were derived from the study by Wang and Strong (1996) for this purpose. This derivation was because the six dimensions implemented fit Wang and Strong's conceptual framework of data quality (as mentioned in Section 3), and our metrics were also selected and built following the definitions of dimensions in this framework (as mentioned in Section 4).

5.3. Results

5.3.1. Values of DQDs operationalized to data sources

The 10 metrics were calculated as the ratios of data points that satisfied these metrics' definitions to the total data points. For example, the metric *Concise representation - Separate Form of Time and Country Information* was assessed by examining data sources' structures to see whether a separate attribute existed for each of time and country information. For IRC measurement, just information of the year has been commonly used in the calculation. Table S2 shows that all four data sources surveyed (ACM DL, Dimensions, MAG, and WoS) have a particular attribute storing information about the years of IRC papers published. However, only Dimensions has a separate attribute (which is *research_org_countries*) indicating the country of affiliations, while MAG and ACM DL store this information and other information (e.g., affiliations' names, state codes, state names) in a combined attribute (authors.org and content.article_rec.authors.au.affiliation, respectively). We calculate the ratio of data points having the smallest set of complete data to the total number of data points. A data point was considered compact when all attributes storing information about *year* and *country* were complete and optimized in length. For example, a data point containing "1995" and "Humboldt-Universität zu Berlin, Germany" was not considered

Table 2. Average values of metrics calculated by data sources chosen

Data sources	**ACM DL** (measured on the whole data source)	**Dimensions** (measured on 40 data blocks)	**MAG** (measured on 40 data blocks)	**Web of Science** (measured on 40 data blocks)
Metrics	Value	Avg.	Avg.	Avg.
Completeness - Property (M1)	0.859	0.761	0.399	1.000
Completeness - Population (M2)	0.359	0.420	0.308	0.389
Appropriate Data - Explicitly (M3)	0.499	0.893	0.678	0.916
Appropriate Data - Implicitly (M4)	0.418	0.115	0.170	0.800
Accuracy - Free of Malformed Datatype (M5)	0.965	0.980	0.958	0.993
EoU - Presence Relevant Vocabularies (M6)	0.917	0.882	0.910	0.879
EoU - Correct Spelling (M7)	0.579	0.491	0.393	0.182
Concise representation - Compact Form of Time and Country Information (M8)	0.005	0.798	0.005	0.560
Consistency - Standard (M9)	0.750	0.605	0.456	0.478
Consistency - Syntax (M10)	0.991	0.993	0.891	1.000

compact. The reason was that although “1995” was the compact form for the *year* attribute, the phrase “Humboldt-Universität zu Berlin, Germany: was not the smallest set of complete data for the *country* attribute as only “Germany” was needed to identify the nationality.

Table 2 shows the average values of these metrics, which were measured using ANOVA.

The values in Table 2 show measurements made of the whole ACM DL and a sample of only 40 randomly selected data blocks of each of the other data sources (WoS, MAG, and Dimensions). Notably, the variances among the averages calculated from the data blocks of each data source (i.e., WoS/MAG/Dimensions) were calculated and it was seen that they were relatively small (< 0.01). In other words, there was minimal skew in the data blocks. Therefore, these average values of metrics measured on samples could be considered closely approximate to the true values of metrics that would be observed if they were measured on the whole data sources.

Table S8 shows a 95% confidence interval for the average value of each metric on the whole data source. The table shows that there are only two out of 10 metrics that had notable variations (> 0.5%) in their confidence interval for MAG data source and WoS data source. They are *Completeness - Population* (M2) and *Appropriate Data - Implicitly* (M4). However, these variation values are small in comparison to the differences between the mentioned two metrics across data sources. Therefore, we could use the average values of metrics when assessing the independence of DQDs across data sources in the following step.

5.3.2. The independence of DQDs

The independence of DQDs across data sources All 10 metrics were calculated on data points from these data sources. Table 3 shows the values of these 10 DQDs’ metrics measured on different data sources.

Table 3. Comparing the values of metrics across different data sources

Metrics	**ACM DL's values** (measured on the whole data source)	**Dimensions' averages** (measured on 40 data blocks)	**MAG's averages** (measured on 40 data blocks)	**WoS's averages** (measured on 40 data blocks)	**Significance rating**
Completeness - Property (M1)	0.859	0.761	0.399	1.000	***
Completeness - Population (M2)	0.359	0.420	0.308	0.389	***
Appropriate Data - Explicitly (M3)	0.499	0.893	0.678	0.916	***
Appropriate Data - Implicitly (M4)	0.418	0.115	0.170	0.800	***
Accuracy - Free of Malformed Datatype (M5)	0.965	0.980	0.958	0.993	***
EoU - Presence Relevant Vocabularies (M6)	0.917	0.882	0.910	0.879	***
EoU - Correct Spelling (M7)	0.579	0.491	0.393	0.182	***
Concise representation - Compact Form of Time and Country Information (M8)	0.005	0.798	0.005	0.560	***
Consistency - Standard (M9)	0.750	0.605	0.456	0.478	***
Consistency - Syntax (M10)	0.991	0.993	0.891	1.000	***

The table shows that the notable differences between the 10 metrics' values across data sources are statistically significant, proved by the small *p*-values in the tests (*** means *p*-value ≤ 0.001). The metrics that reflected apparent differences (at least 15%) between a particular data source and the others are highlighted and represented as follows:

- *Completeness of Property* (M1): The average ratio of this metric on MAG is lower than those on ACM DL, Dimensions, and WoS (39.9% compared to 85.9%, 76.1%, and 100%, respectively). These differences mean that MAG has more missed or empty values (e.g., Null/NA) for the expected affiliations of corresponding authors than other data sources do.
- *Appropriate Data - Explicitly* (M3): The average ratios of this metric on Dimensions and WoS are notably higher, and that on ACM DL is lower than the value measured on MAG (89.3%, 91.6%, and 49.9% compared to 67.8%, respectively). These differences mean that Dimensions and WoS have more explicit "country" information in the nonempty data points while ACM DL has the least ratio of explicit "country" information.
- *Appropriate Data - Implicitly* (M4): The average ratios of this metric on Dimensions and MAG are notably lower, and that on WoS is the highest in comparison to the value measured on ACM DL (11.5%, 17% and 80% compared to 41.8%, respectively). This difference means that Dimensions and MAG have less implicit "country" information from the data points that do not include explicit information than ACM DL, while WoS has the highest implicit "country" information ratio.
- *EoU - Correct Spelling* (M7): The average score of this metric on WoS is lower than those on ACM DL, Dimensions, and MAG (18.2% compared to 57.9%, 49.1%, and 39.3%, respectively).

- *Concise representation - Compact Form of Time and Country Information* (M8): The average scores of this metric on ACM DL and MAG are especially lower than those on Dimensions and WoS (0.5% and 0.5% compared to nearly 79.8% and 56%, respectively).
- *Consistency - Standard* (M9): The average ratios of this metric on ACM DL and Dimensions are higher than those on MAG and WoS (75% and 60.5% compared to 45.6% and 47.8%, respectively). This difference means that ACM DL and Dimensions have more affiliations following a consistent standard in the nonempty data points (than MAG and WoS do).

As we noted from the previous step, the estimated values of M2 and M4 had a slightly notable variation (> 0.5%) in their confidence intervals for the MAG data source. We then were cautiously afraid that such notable variation might affect the accuracy of results in accessing the independence of DQDs across data sources (in this step). However, as we can notice in Table 3, the *p*-values of testing M1 and M2 are < 0.001. In other words, they provided strong evidence that the differences between ACM DL, Dimensions, MAG, and WoS are statistically significant.

There are three results from the above findings. First, the values of different DQDs' metrics varied across data sets. Second, five dimensions reflect the differences in data from primary sources for IRC measurement: *Appropriate amount of data*, *Completeness*, *Concise representation*, *Ease of Understanding*, and *Representational consistency*. Third, there are more "better" results when the metrics were measured on WoS and Dimensions than on MAG.

The independence of DQDs across disciplines Table S9 shows the *p*-values of the ANOVA test analyzing metrics calculated across these disciplines in Dimensions. These values measured on Dimensions subsets by every metric were different clearly. All the *p*-values were small (*** means *p*-value ≤ 0.001), showing that these differences were statistically significant.

Then we analyzed metrics calculated across these disciplines in MAG. Because the MAG data source did not include an attribute mentioning the papers' disciplines, we needed to classify MAG papers into relevant subsets. Using the *fos* values of sampled papers appearing in both the MAG data source and each of Dimensions' 22 subsets divided by discipline, we could identify and separate 85% of MAG papers into 22 corresponding subsets. Table S10 shows the *p*-values of the ANOVA test analyzing metrics calculated across these disciplines in MAG. These values measured on MAG subsets by every metric were clearly different. All the *p*-values were small (*** means *p*-value ≤ 0.001), showing that these differences were statistically significant.

Table S11 shows the variance values of 10 metrics calculated across disciplines, measured on WoS. Except for the two metrics M1 and M10, which had all values at 1, the other metrics show significant differences (*p*-value ≤ 0.001) when they were measured on WoS subsets.

Tables S11 show that the values of our developed DQDs' metrics were different across disciplines, and these differences were statistically significant.

The correlation values of each pair of metrics are given in Tables S12–S15 for ACM DL, Dimensions, MAG, and WoS, respectively (all *p*-values were nearly 0).

In Tables S12–S15, MAG shows two strong linear relationships (correlation coefficient value > 0.7) between metrics and WoS shows a strong linear relationship. For MAG, the relationships are between *Completeness - Property* (M1) and one of the two metrics: *Completeness - Population* (M2), *Accuracy - Free of Malformed Datatype* (M5). For WoS, the relationship is between *Concise representation - Compact Form of Time and Country Information* (M8) and *Consistency - Standard* (M9). These relationships are not common across the data sources. In other words, the metrics applied reflected different aspects of the bibliographic sources' quality dimensions. Therefore, it is not necessary to remove or restructure any metric above.

Table 4. The weights built from the importance ratings by Wang and Strong (1996)

	Average of importance ratings (Wang & Strong, 1996)	Inverse values of (1)	Weights in proportions of the sum of (2)
DQD	*(1)*	*(2)*	*(3)*
Accuracy	3.05	5.95	0.20
Completeness	3.88	5.12	0.17
Appropriate amount of data	5.01	3.99	0.13
Concise representation	4.75	4.25	0.14
Representational consistency	4.22	4.78	0.16
Ease of Understanding	3.22	5.78	0.19

5.3.3. Ranks of data sources

As mentioned in Section 5.2.3, we considered using the importance ratings of DQDs proposed by Wang and Strong (1996). This study computed the average of the importance ratings for dimensions from data consumers, but these were in reversed order (i.e., lower values indicate higher importance of the respective dimension). Consequently, we converted these values by subtracting the maximum value of the survey's Likert-type scale (9, on a scale from 1 to 9) from each of these average values. The weights were then calculated by taking the proportions of the inverse values. The results are shown in Table 4.

Therefore, the values received in column 3 in Table 4 were weights derived from Wang and Strong (1996). They were then used as a way to weigh the DQDs. Table 5 presents weights for metrics and two different sets of weights for DQDs. The evaluated values of data sources were calculated by applying these different options of weights and are presented in Table 6.

Table 5. Weights of different options for metrics and DQDs

	Weights		
	For metrics	***For DQDs***	
DQDs—Metrics	***Equal weights***	***Equal weights***	***Weight derived from Wang and Strong***
Completeness - Property (M1)	0.50	0.17	0.17
Completeness - Population (M2)	0.50		
Appropriate Data - Explicitly (M3)	0.50	0.17	0.13
Appropriate Data - Implicitly (M4)	0.50		
Accuracy - Free of Malformed Datatype (M5)	1	0.17	0.20
EoU - Presence Relevant Vocabularies (M6)	0.50	0.17	0.19
EoU - Correct Spelling (M7)	0.50		
Concise representation - Compact form of time and country information (M8)	1	0.17	0.14
Consistency - Standard (M9)	0.50	0.17	0.16
Consistency - Syntax (M10)	0.50		

Table 6. Evaluated values of data sources with weights added for metrics

	Weights	
Data sources	***Equal weights***	***Weights derived from Wang and Strong (1996)***
ACM DL	0.609	0.406
Dimensions	0.726	0.521
MAG	0.511	0.348
WoS	0.729	0.548

The results in Table 6 show that WoS was ranked as the highest quality data source, by using either equal weights or weights derived from Wang and Strong (1996).

In summary, this section presents the 10 metrics' values of six relevant DQDs for evaluating bibliographic data sources. Five out of six dimensions (except for *Accuracy*) reflect the significant differences ($p < 0.001$) across the data sources under the survey. These differences show that there are more "better" results when the metrics were measured on WoS and Dimensions than on MAG. In addition, the dimensions' values are significantly different ($p < 0.001$) across disciplines. The metrics have no strong relationships with each other so they can be used in evaluating the bibliographic data sources. The evaluation shows that WoS received the highest scores for its fitness to use in IRC measurement. These results are discussed in the next section.

6. DISCUSSION

The goal of our investigation was to assess the quality of bibliographic data sources for measuring IRC. The main findings of this study were discussed around the research questions as follows:

6.1. Relevancy of Dimensions for IRC Measurement

RQ1: Which dimensions are relevant to a DQA framework for IRC measurement?

With an inventory of possible dimensions identified from the literature review, we selected seven dimensions that we considered relevant to IRC measurement (*Accuracy*, *Appropriate amount of data*, *Completeness*, *Concise representation*, *Ease of Understanding*, *Relevancy*, and *Representational consistency*). Except for *Relevancy*, the other six (Table S5, 'DQD' column) among the seven dimensions selected were successfully operationalized by practical metrics. As they were selected specifically for the task of IRC measurement, this list of six dimensions was not identical to task-independent dimensions suggested by other studies, which required specific attributes from the data sources. For example, the framework of computable dimensions by Rajan, Gouripeddi et al. (2019) included the dimension *Currency* (also named Timeliness in some studies). This dimension required information about the average "out of date" values of data, which were not provided by the data sources under the survey. Another example is the list of dimensions selected specifically for Linked Open Data (Zaveri et al., 2016). This list was selected to reflect the nature of linked data (e.g., the *Availability* dimension was measured with metrics involving the accessibility of the SPARQL endpoint and the server, and the accessibility of the RDF dumps). These metrics were not applicable for the task of IRC measurement because the SPARQL endpoint and RDF did not exist in bibliographic data. The above examples suggested that a specific set of metrics should be built for each task at hand.

To the best of our knowledge, our study was the first attempt to operationalize DQDs for IRC measurement. Previous studies about IRC measurement either ignored the reason why their data sources were chosen or chose particular data sources because these data were available during their studies. These practices implied that the findings in previous studies might vary differently according to which data sources were used in the studies (Nguyen, Luczak-Roesch, & Dinneen, 2019). Another implication is that we have not known whether IRC-data-quality would be different from general-data-quality. The DQDs and their built-in metrics that we selected will help researchers in this specific domain to evaluate and determine the most suitable bibliographic data sources needed for future studies.

6.2. Meaningful Differences Among the DQDs

> **RQ2**: Which dimensions from the DQA framework reflect differences in the data from primary sources for IRC measurement?

We found five dimensions (except for *Accuracy*) reflecting notable differences across data sources (*Completeness, Appropriate amount of data, Concise representation, EoU - Correct Spelling,* and *Representational consistency*). Each of the remaining dimensions was measured by metrics reflecting different aspects of that dimension on the data sources surveyed. For example, the *Completeness* dimension was evaluated with two metrics: *Completeness - Property* and *Completeness - Population*. Both metrics performed differently with a statistical significance of 0.001. These differences show that the data sources under study performed differently for the task at hand.

The exceptional dimension that did not reflect notable differences was *Accuracy*, which scored high (95%–99%) for the data sources in the survey, similar to the results of the study by Sinha et al. (2015). This dimension was measured in the present study by its only metric—"the detection of malformed datatype"—because other possible metrics for *Accuracy* were either inapplicable or impractical in the context of IRC measurement (as presented in Table S5). This metric reflected the "free of error" status of the data sources and showed that all the data sources surveyed performed well at this aspect of *Accuracy*. The study might have shown different values of *Accuracy* across these data sources if other metrics had been applicable to measure the other aspects of *Accuracy*. In other words, the findings received in the current study might have been different if more metrics had been included successfully in the evaluation of the *Accuracy* dimension. As *Accuracy* was considered the key dimension of data quality (Olensky, 2015), the inclusion of only one metric reflecting one aspect of this dimension may not fully express how accurate the data sources are. Although the approximate scores of *Accuracy* across the four bibliographic data sources implied that we could exclude this dimension from our DQA framework, we kept the DQA framework unchanged for general use because other data sources might show notable differences.

Among the four data sources, MAG had notably lower quality scores, while Dimensions and WoS had notably higher quality scores. The poor performance of MAG agrees with other studies about the quality of bibliographic data sources for tasks beyond IRC. For example, Huang, Neylon et al. (2020) showed that MAG, while having higher coverage for journals and conferences in comparison with WoS and Scopus, has "less complete affiliation metadata." The lower scores of the dimensions *Appropriate data (explicitly)* and *Consistency* in MAG can also be explained by the fact MAG data set was built from web pages indexed by Bing (Sinha et al., 2015). Consequently, many affiliations from these web pages may lack information about nationality or may not be correctly spelled. Because MAG is an openly available bibliographic data source for scientometrics, the use of this data source in IRC measurement studies has become widespread and this circumstance may lead to IRC measurement results of low quality. Therefore, researchers should be aware of and consider MAG's weaknesses in choosing bibliographic data sources for their studies.

In contrast, the high scores of Dimensions and WoS can be explained by how these data sources organize the affiliation records. Dimensions and WoS both show notably high scores for *Appropriate data (explicitly)* and *Concise representation - Compact Form of Time and Country Information*. The differences are because Dimensions and WoS stored affiliations' nationality and year information in a separate attribute, so their scores for *Concise representation - Compact Form of Time and Country Information* are higher in comparison with the scores of ACM DL and MAG. For *Appropriate data (explicitly),* Dimensions data are enriched with GRID—a global research identifier database (Orduña-Malea & Delgado-López-Cózar, 2018). This data infrastructure allows assigning each institution to a persistent GRID identifier, so the number of name variants of each institution will be minimized. As a result, the ratio of explicit information about nationality in Dimensions affiliations can be further improved. WoS also scored the highest (100%) for *Completeness - Property* (M1). This result was unexpected because it was inconsistent with the result of Jacsó (2009), which showed that 14% of WoS data was missing country information. However, WoS had low scores for *EoU - Correct Spelling* (M7) and *Consistency - Standard* (M9). These scores are low because many records of WoS were in uppercase and/or acronyms (e.g., "UNIV CALIF BERKELEY, DEPT GEOL & GEOPHYS"). Overall, the combined DQDs' scores led to the highest rank of WoS (as presented in Table 6), which reflected the time and country disambiguation ability of WoS in comparison to other data sources. Another notable point is that WoS has a higher proportion of data points at the beginning of the period surveyed (1980–2017) than other data sources (shown in Figure 3). The difference may be because WoS was the commercial data source that came into operation earlier than other data sources. An implication here is that WoS may be more useful for research surveying IRC before the 1990s than other data sources.

It is also interesting to note that ACM, while scoring worst at *Appropriate data (explicitly)*, has a notably high score for *Appropriate data (implicitly)*. This finding is consistent with a previous study (Nguyen, Dinneen, & Luczak-Roesch, 2019), which found that, in comparison with MAG, ACM DL has fewer affiliations containing explicit information about nationality. However, ACM DL also has a higher ratio of affiliations that can be disambiguated by applying string matching and Wikidata query (Nguyen, Dinneen, & Luczak-Roesch, 2020). This high ratio of implicit information compensates for the low ratio of affiliations containing explicit nationality information.

Our study also suggests that the quality of a domain-specific data source depends on that domain's nature (in Section 5.3.2). For the use of bibliographic data sources in general, other previous studies have also found that certain aspects differed across data sources, such as the average citation counts and the journal coverages (Huang et al., 2020). Our study's findings imply that the data bibliographic sources should be used for measuring IRC in domain-specific and multidisciplinary studies differently.

6.3. Results of Applying the Developed DQA Instrument

RQ3: Which data source(s) is/are most suitable for measuring IRC?

By ranking the data sources surveyed using the combined measure, we successfully validated the developed DQA instrument for IRC measurement. This DQA instrument provides the baseline for researchers to use and develop in their study with regard to assessing the quality of data sources used to measure IRC.

As mentioned above, data quality is commonly defined as “fitness for use” (Strong et al., 1997). For IRC measurement, we found that WoS is the most suitable choice among the data sources under the survey. However, the gap between the scores of WoS and the second highest quality data source (Dimensions) is quite small, as shown in Table 6, and importantly, accessing WoS entails a fee whereas accessing Dimensions is free. Therefore, Dimensions may be the top choice if *cost-effectiveness* were to be considered, a possibility which we consider in our concluding remarks. For use in a wider context than IRC measurement, results from other studies analyzing other uses of bibliographic data sources showed inconsistent ranking outcomes of bibliographic data sources. For example, Visser, Van Eck, and Waltman (2021) concluded that Scopus and WoS outperformed Dimensions and MAG regarding the quality of citation links in these data sources, while Singh et al. (2021) found that Dimension had more unique journals than Scopus, and WoS had the least number in terms of the journal coverage. In general, the results of other studies vary according to their focuses, and we cannot compare these results with our findings because the focus of IRC measurement is on the existence and quality of “country” and “time” information in the bibliographic data sources.

As journal coverage was an important aspect in informing the comprehensiveness of data sources (Martín-Martín et al., 2018), and significant differences in journal coverage were observed (Singh et al., 2021), this aspect could be used as an additional criterion for choosing suitable data sources. We can consider a broader approach for data assessment in which data quality, measured by our developed DQ framework, reflects the qualitative aspect while the coverage reflects the quantitative aspect of any data source. Another possible consideration is to include the journal coverage, which presents the number of unique journals covered in each data source, in an extended DQA framework. Prior studies of DQA for bibliographic data have not considered this aspect. For example, Zaveri et al. (2016) only included “sufficient scope (number of entities) and detail (number of properties applied)” in a given data source as a coverage metric for the dimension *Appropriate amount of data*. However, journal coverage is important in IRC measurement studies because their results may be different if data sources having different journal coverage are used in the studies. Among data sources performing equally at time and country disambiguation, those including more unique journals per year will reflect the image of the IRC activities more accurately. Also, the IRC network has changed over time (Wagner & Leydesdorff, 2005) so data sources covering a longer period of journals will give a more thorough image of the IRC activities. The present study assessed the fitness for use of the four data sources in a fixed period (1980–2017) because we did not have access to the whole coverage of all data sources under the survey. As journal coverage is important in IRC studies it can be further developed as another metric for IRC measurement.

6.4. Additional Findings

Beyond answering the posed research questions, our study revealed additional insights. First, we made a “metadata crosswalk” between the FRBR model and bibliographic data sources. We applied this approach to select the relevant attributes for IRC measurement, and then to select the DQDs that could be evaluated with these attributes. This approach was useful to assess the relevance of DQDs for IRC measurement, in the context that we could not consider the frequency ranking of DQDs in the literature because just a few prior studies were researching the DQDs of bibliographic data. We realized that, although the FRBR model has been applied to distinguish a *work* (e.g., research) from its *manifestations* (e.g., many publications of the same research) in bibliographic studies (e.g., Bar-Ilan, 2010; Moed, Bar-Ilan, & Halevi, 2016), no previous research has described the mapping of publication data sources’ attributes to FRBR model’s entities. As our implementation of the “metadata crosswalk” categorized the attributes of the four bibliographic data sources under survey into appropriate FRBR model's entities, this map will also be useful for future studies in which researchers need to find references to make a publication-to-publication comparison or research-to-research comparison between these data sources. Second, a set of metrics was specifically proposed for IRC measurement. For example, we proposed the metric compact form of time and country information for the *Concise representation* dimension. *Concise representation* has been considered to be a subjective criterion in many prior studies (e.g., Caballero, Verbo et al., 2007; Naumann & Rolker, 2005), which means that the users' judgment determines this dimension's value. Our proposal is an attempt to determine the *Concise representation* dimension's value by a quantification method not involving human judgment. In our study, the metric's values for the *Concise representation* dimension varied across the data sources surveyed, and there were no relationships between this metric and the other dimensions' metrics. In other words, our proposed metric for the *Concise representation* dimension reflected the differences among data sources, and it reflected a separate aspect from the other metrics. Overall, our proposed set of metrics provides a practical baseline for future IRC measurement studies, which can simply reuse or develop this metric set for their own tasks.

6.5. Implications for IRC Measurement

This study's overall goal was to examine to what extent different bibliographic data sources are suited to measure IRC. We achieved it by steps, namely: There were seven particular DQDS found relevant for IRC measurement (*Accuracy*, *Appropriate amount of data*, *Completeness*, *Concise representation, Ease of Understanding, Relevancy*, and *Representational consistency*); of which six (Table S5, 'DQD' column) among the seven dimensions selected were successfully operationalized by practical metrics, five reflected differences in the data from primary sources for this task (*Completeness*, *Concise representation*, *Representational consistency*, *Appropriate amount of data*, and *Ease of Understanding*); and WoS is most suitable for measuring IRC.

Our study is critical because it contributes to understanding data quality for IRC measurement, which is a core but incomplete topic in IRC studies (Chen et al., 2019). We finalized a list of dimensions relevant to the task of IRC measurement, showed how the dimensions selected can be implemented with objectively computable metrics, and showed how the data sources were ranked for the task of IRC measurement. Either the DQA framework suggested, the operationalization method described or the ranking list of data sources presented in the present study can be used by other researchers in their IRC measurement studies. In light of the FRBR structure, we showed that different bibliographic data sources were organized differently. With this approach, the differences in the performance of the data sources surveyed can then be comprehensively compared, and the strengths and weaknesses of the four data sources surveyed can be easily identified.

Our study implies a methodological contribution in general. DQA is a tricky task because of the subjectivity of various parts in the DQA framework. The selection of DQs can be very subjective and task specific. Therefore, the assessments of bibliographic data are often irreproducible in IRC studies because the methods for selecting DQs are not described clearly. Our methodological contribution can be considered in the wider context of data quality, not just bibliographic data quality.

Our study produces some useful implications for IRC measurement as well. First, we proposed a complete DQA framework for IRC measurement. Therefore, our work supplies a reference for further studies of IRC measurement to easily choose suitable bibliographic data sources for their tasks. Other researchers can simply apply our selected DQDs, which were considered relevant to IRC measurement, or apply our corresponding developed metrics to automatically evaluate the DQDs they choose. Otherwise, further IRC measurement studies have to review the literature themselves, select relevant DQDs, and build corresponding metrics for each DQD selected to compare the quality of different bibliographic data sources. All of these steps take time and, therefore, put a heavy burden on the task of IRC measurement. Second, we built and applied specific metrics for each DQDS to measure the data quality for IRC measurement. For each DQD, we built relevant measures and corresponding algorithms to evaluate the values of this DQD. The methods will be shared online so that future work can reuse and develop them in IRC studies. As no similar study has been done previously, the shared methods can greatly contribute to the bibliometric community's development. Third, we proved that metrics result in different values across different data sources, depending on the data sources' certain aspects (e.g., how that data source is organized, collected, and provided). Metrics also result in different values across a data source's disciplines. In other words, the data quality of a domain-specific data source also depends on that domain's nature. Our findings confirmed that bibliographic data sources have discipline bias (presented in Section 5.3.2). These findings suggested that our designed DQA can be applied for IRC measurement studies, but the results will vary across data sources' subsets by discipline. Fourth, we proposed the "best" data source (among the four options reviewed) to measure IRC, either for just the domain of computing sciences or for all domains in general.

7. LIMITATIONS

Our study has identified some potential but not fully explored aspects in considering the data quality for IRC measurement. There are some limitations, as follows.

The first limitation is linked to the approach used to design the DQA framework for IRC measurement. In this approach, we assessed the relevance of each DQD by considering how it could be meaningfully applied to measure IRC using the attributes found (i.e., *time published* and *countries involved*) in the data sources surveyed. Consequently, the number of relevant DQDs selected was limited to only the DQDs that could be evaluated using the data sources surveyed. Some other DQDs would not have been excluded if the evaluations had been done with additional data sources. For example, information about the ranking of scientific journals can be used to access bibliographic data sources' reputations, and the *Reputation* dimension can then be considered for the DQA framework for IRC measurement. This limited selection of DQDs might not fully reflect the quality of data sources for IRC measurement, because DQDs may be more or less relevant depending on the IRC measurement task and its ultimate purpose.

Rather than assume and design for a specialized measurement task, we have picked dimensions that appear widely applicable for what all IRC measurement tasks have in common: quantifying collaboration across international borders. Nonetheless, other quality dimensions not considered here might be *very* appropriate for more specialized tasks. For instance, interlinking data may be necessary to examine the structure of national higher education systems and their organizational characteristics or the relevant national policies (Lepori, Barberio et al., 2013). For such a task, a DQD reflecting interlinking might be appropriate, especially to assess the ease or extent (i.e., possible ways) of combining data sources. Similarly, a DQD reflecting licensing might be relevant if the measurement task is concerned with the reusability of data and the reproducibility of the relevant publications. But these dimensions that may be useful for specific analyses are not easily measured via the approach we have proposed, which uses common kinds of bibliographic data and their related DQDs.

A possible improvement is assessing DQDs by not only how they can be applied in IRC measurement but also how they can be beneficial for IRC measurement in general. For instance, the assessment may include DQDs showing the ability to link to other data sources and, therefore, helping to check and improve the details of publications in one data source with data from another one. Another improvement is that the assessment may include DQDs that can be evaluated not just by the data source itself under investigation, but also by other data sources. For example, information about the extent to which the data source's content is highly regarded can be referred to from other data sources not considered for the DQA framework.

Another limitation of our study is that we operationalized the data dimensions by applying only computable metrics, compared to subjective ones (Rajan et al., 2019). Because not all metrics can be measured without human judgment, the number of metrics to measure each DQD was limited. Consequently, the evaluation for each of the dimensions in our DQA might not have reflected its definition properly as it should have done. One example is that the *Accuracy* dimension, although it implies various aspects (e.g., "correct, reliable, and certified free of error" by Wang & Strong [1996]), could not be measured for either "correct" or "free of error" by computable metrics. It is possible that, despite our findings, accuracy is a meaningful differentiator for the different data sources, and that our metric for accuracy was simply too narrow to capture the difference; only one metric—"the detection of malformed datatype"—was implemented for the *Accuracy* dimension. However, this is currently the only practical way to implement the metric. Without some "ground truth" data or human verification (impractical), determining the accuracy of data remains an open problem that prevents us knowing with more certainty whether accuracy differs and affects IRC measurement. Another example is the operationalization of the *Ease of Understanding* dimension. This dimension has been considered a subjective criterion (i.e., this dimension's value can only be determined by the users' rating, Naumann & Rolker, 2005). Our study attempted to measure it with two metrics: *Presence Relevant Vocabularies* (M8) and *Correct Spelling* (M9). Although these two metrics might somehow be necessary for the data to be clear, they might not be sufficient to ensure that the data could be easily comprehended (e.g., affiliation "university school" might contain relevant vocabularies, with all of these words correctly spelled, but the combination of them make no sense to readers). Therefore, the values measured by these two metrics might not entirely reflect the quality of the *Ease of Understanding* dimension by definition.

In addition, our computable metrics may not work as thoroughly as expected. Because the computable metrics were implemented with the assistance of available R packages, there might be some circumstances in which these metrics could not accurately reflect reality. For instance, the metric *correct spelling* tried to recognize geographical names included in the affiliation data before checking whether these data are fully checked for spelling. The *maps* package was used for this purpose. Because this package's database primarily includes world cities with a population greater than about 40,000, there is a possibility that some small towns or cities included in the affiliation data could not be identified by the metric *correct spelling.* Consequently, the metric *correct spelling* may wrongly evaluate some affiliations as incorrectly spelled if these affiliations include small towns or cities' names.

Furthermore, the lack of human involvement in evaluating the importance of DQDs for IRC measurement is also a limitation. Instead of weighting the importance of DQDs by interviews or surveys, our study applied the results of Wang and Strong (1996). In this study, they calculated the values indicating the importance of each DQD from data consumers' opinions. As the study by Wang and Strong was carried out many years ago, and the participants were data consumers in general, this study's results may not serve well for researchers in the IRC measurement domain at present. Although our approach simplified the burden of work involved, it is not the best way to get findings that could have been derived with the support of work from IRC measurement experts. Therefore, the application of our DQA might be less effective than that of a fully integrated DQA with evaluations by relevant humans.

Finally, the selection of data sources surveyed in this study is another limitation. Our study surveyed the data quality of two commercial bibliographic data sources (Dimensions and WoS), one open data source (MAG), and one specific domain data source (ACM DL). Although representing different types of bibliographic data sources, these four data sources do not cover all of the possible common data sources used in scientometrics. For example, we did not have access to Scopus and PubMed copies, which are also two important and commonly used sources for IRC measurement, so they were not included in this present study. As a result, our recommendations for bibliographic data sources could not be applied to the missing ones, and our study's application was limited to only the data sources surveyed.

8. CONCLUSION AND FUTURE WORK

In conclusion, our study aimed to find how well different bibliographic data sources are suited to measure international research collaboration and which dimensions (DQDs) of such sources are important in determining their suitability. Our work identified relevant DQDs from data quality literature and implemented corresponding computable metrics to build a framework for assessing data quality for IRC measurement. The designed instrument was then validated by applying it to four important bibliographic data sources. On the three multidisciplinary bibliographic data sources—Dimensions, MAG, and WoS—this application revealed that the measure of DQA depends on the nature of each discipline. Our findings also suggested WoS as the highest quality data source for IRC measurement studies. We also recommended the use of the second highest quality data source—Dimensions—if cost-effectiveness is considered. Our study filled the lack of DQA in IRC measurement by proposing a DQA framework for this task. In addition, the implementation of relevant DQDs in our study is shared online so that other researchers will be able to use them in future studies.

For future work, some further developments can be carried out. Currently, we list seven DQDs as relevant to IRC measurement (by assessing how they can be applied in IRC measurement), in which six DQDs were operationalized (by applying only computable metrics). These DQDs were just a part of 15 dimensions in the conceptual framework of information quality proposed by Wang and Strong (1996). Consequently, some aspects of data quality were not considered in our study. Other studies in the future could examine some other dimensions to cover other aspects of data quality and better evaluate data quality for IRC measurement. One example is the *Reputation* dimension. By definition, *Reputation* implies two aspects (Wang & Strong, 1996). The first aspect is the *Reputation of the data source*. The implementation for this aspect needs a combination with other data source(s). For example, reputation scores about the data sources (Dimensions, MAG ...) from other studies, or at least reputation scores of lists of journals stored in the data sources should be available and used to calculate this dimension. The second aspect is the *Reputation of the data content*. This aspect can be calculated using some available attributes in the data sources. For example, information about each article's impact can be used as a baseline for that article's reputation. However, the implementation is somewhat complicated and this approach will introduce bias to the evaluation, as not all citations are endorsements, and the average citations in different disciplines are different (Huang et al., 2020). Another dimension that could be considered, for instance, is the

Interlinking dimension. *Interlinking* was considered an additional dimension to the *Accessibility* category in Wang and Strong's framework (Candela, Escobar et al., 2021; Zaveri et al., 2016). The implementation of this dimension in further studies could be done by detecting the existence of links to external data providers (Zaveri et al., 2016). The fact that information from bibliographic data sources (i.e., the discrete records) is not alone exhaustive of DQA suggests a need to standardize these metadata sources. Other potentially important aspects of sources, such as those attributable to the publishing industry (e.g., standard reports, or version control for the data), should also be encouraged as a means for the standardization of bibliographic data and their quality.

Human opinions will be used for both measuring the DQDs and weighting them in future work. For example, researchers in the domain of IRC measurement will be asked for their evaluations about how well each DQD performs on each data source, and how vital each DQD should be in the DQA framework. As the DQDs will be evaluated not only by computable metrics but also by other metrics with inputs from these experts, the results received in the evaluation will follow the definitions of these DQDs better. In addition, researchers' opinions in the domain of IRC measurement will be used to determine the weight of each DQD in the DQA framework. In total, the inclusion of both qualitative and quantitative assessments will help to increase the reliability of our DQA framework for IRC measurement.

Another area for future improvement is to develop the DQA framework into a data source evaluation framework with the inclusion of a measure for *cost-effectiveness*. As different data sources have different access fees (e.g., WoS has a fee to access it, whereas Dimensions does not), the opportunities for access are not the same for all researchers and institutions. Therefore, cost-effectiveness is likely an important criterion in the process of choosing data sources and should be included in a framework in future studies.

Our DQA framework will also be used to assess other bibliography data sources. As Scopus, PubMed, Crossref, and OpenCitations are the other major bibliographic data sources in quantitative science studies (Waltman & Larivière, 2020), these data sources will be considered to be included in our further study. From the recommendations of this future work, the researchers in the domain of IRC measurement will be able to assess and choose suitable bibliographic data sources for their studies.

Finally, the use of only joint research publications in IRC measurement has limitations because there are various types of outcomes as well, such as patents and joint research grants (Yuan, Hao et al., 2018). Therefore, the need to assess these relevant data sources' quality for measuring IRC will also be carried out in the future.

AUTHOR CONTRIBUTIONS

Ba Xuan Nguyen: Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Markus Luczak-Roesch: Conceptualization, Methodology, Resources, Supervision, Writing—review & editing. Jesse David Dinneen: Conceptualization, Methodology, Resources, Supervision, Writing—review & editing. Vincent Larivière: Data curation, Resources, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

No funding has been received for this work.

DATA AVAILABILITY

The data sets (ACM DL, Dimensions, WoS) used for analyses in the current study are not publicly available due to confidentiality clauses.

The data set MAG can be accessed at https://www.microsoft.com/en-us/research/project/open-academic-graph/.

The source code and data generated during the current study are available at https://doi.org/10.5281/zenodo.7016728 (Nguyen, Luczak-Roesch et al., 2022).

REFERENCES

Aksnes, D. W., Piro, F. N., & Rørstad, K. (2019). Gender gaps in International Research Collaboration: A bibliometric approach. *Scientometrics*, *120*(2), 747–774. https://doi.org/10.1007/s11192-019-03155-3

Anderson, C. J. (2010). Central limit theorem. In *The Corsini encyclopedia of psychology*. https://doi.org/10.1002/9780470479216.corpsy0160

Anuradha, K., & Urs, S. (2007). Bibliometric indicators of Indian research collaboration patterns: A correspondence analysis. *Scientometrics*, *71*(2), 179–189. https://doi.org/10.1007/s11192-007-1657-4

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics*, *83*(3), 809–824. https://doi.org/10.1007/s11192-009-0145-4

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, *52*(3), 365–377. https://doi.org/10.1023/A:1014254214337

Beaver, D. D., & Rosen, R. (1978). Studies in scientific collaboration. *Scientometrics*, *1*(1), 65–84. https://doi.org/10.1007/BF02016840

Bizer, C., & Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3199414

Bornmann, L., & Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLOS ONE*, *13*(5), e0197133. https://doi.org/10.1371/journal.pone.0197133, PubMed: 29791468

Burke, P. F., & Reitzig, M. G. (2007). Measuring patent assessment quality—Analyzing the degree and kind of (in)consistency in patent offices' decision making. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.880705

Caballero, I., Verbo, E., Calero, C., & Piattini, M. (2007). A data quality measurement information model based on ISO/IEC 15939. In *Proceedings of the 12th International Conference on Information Quality* (pp. 393–408).

Candela, G., Escobar, P., Sáez, M., & Marco-Such, M. (2021). A Shape Expression approach for assessing the quality of Linked Open Data in libraries. *Semantic Web*, 1–21. https://doi.org/10.3233/SW-210441

Chen, K., Zhang, Y., & Fu, X. (2019). International research collaboration: An emerging domain of innovation studies? *Research Policy*, *48*(1), 149–168. https://doi.org/10.1016/j.respol.2018.08.005

Choi, S. (2012). Core-periphery, new clusters, or rising stars? International scientific collaboration among 'advanced' countries in the era of globalization. *Scientometrics*, *90*(1), 25–41. https://doi.org/10.1007/s11192-011-0509-4

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, *35*(3), 370–381. https://doi.org/10.1016/j.socnet.2013.04.004

Downing, C., Temane, A., Bader, S. G., Hillyer, J. L., Christopher Beatty, S., & Hastings-Tolsma, M. (2021). International nursing research collaboration: Visualizing the output and impact of a Fulbright Award. *International Journal of Africa Nursing Sciences*, *15*, 100380. https://doi.org/10.1016/j.ijans.2021.100380

Flemming, A. (2010). Quality characteristics of linked data publishing datasources. *Master's thesis*, Humboldt-Universität of Berlin.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., … Barabási, A. L. (2018). Science of science. *Science*, *359*(6379). https://doi.org/10.1126/science.aao0185, PubMed: 29496846

Ge, M., Helfert, M., & Jannach, D. (2011). Information quality assessment: Validating measurement dimensions and processes. In *Proceedings of the 19th European Conference on Information Systems*.

Glänzel, W., & Schubert, A. (2001). Double effort = Double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, *50*(2), 199–214. https://doi.org/10.1023/A:1010561321723

Harder, R. H., Velasco, A. J., Evans, M. S., & Rockmore, D. N. (2015). Measuring verifiability in online information. *arXiv preprint*, arXiv:1509.05631. https://doi.org/10.48550/arXiv.1509.05631

Harter, S. P. (1997). *Online information retrieval: Concepts, principles, & techniques*. Academic Press.

Hatakenaka, S. (2008). New developments in international research collaboration. *International Higher Education*, *50*. https://doi.org/10.6017/ihe.2008.50.7998

Heckman, J. J. (2005). 1. The scientific model of causality. *Sociological Methodology*, *35*(1), 1–97. https://doi.org/10.1111/j.0081-1750.2006.00164.x

Heinrich, B., Klier, M., & Kaiser, M. (2009). A procedure to develop metrics for currency and its application in CRM. *Journal of Data *and Information Quality*, *1*(1), 1–28. https://doi.org/10.1145/1515693.1515697

Heinrich, B., & Klier, M. (2010). Assessing data currency—A probabilistic approach. *Journal of Information Science*, *37*(1), 86–100. https://doi.org/10.1177/0165551510392653

Hennemann, S., Wang, T., & Liefner, I. (2011). Measuring regional science networks in China: A comparison of international and domestic bibliographic data sources. *Scientometrics*, *88*(2), 535–554. https://doi.org/10.1007/s11192-011-0410-1

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of linked data conformance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3198962

Huang, C. K. K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, *1*(2), 445–478. https://doi.org/10.1162/qss_a_00031

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional Requirements for Bibliographic Records: Final Report (UBCIM Publications, New Ser., v. 19)* (Reprint 2013 ed.). De Gruyter. https://doi.org/10.1515/9783110962451

Jacsó, P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. *Online Information Review*, *33*(2), 376–385. https://doi.org/10.1108/14684520910951276

Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. (1999). Architecture and quality in data warehouses: An extended repository approach. *Information Systems*, *24*(3), 229–253. https://doi.org/10.1016/S0306-4379(99)00017-4

Kim, J., Kim, H., & Diesner, J. (2014). The impact of name ambiguity on properties of coauthorship networks. *Journal of Information Science Theory and Practice*, *2*(2), 6–15. https://doi.org/10.1633/JISTaP.2014.2.2.1

Lepori, B., Barberio, V., Seeber, M., & Aguillo, I. (2013). Core–periphery structures in national higher education systems. A cross-country analysis using interlinking data. *Journal of Informetrics*, *7*(3), 622–634. https://doi.org/10.1016/j.joi.2013.03.004

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, *10*(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017

Närman, P., Holm, H., Johnson, P., König, J., Chenine, M., & Ekstedt, M. (2011). Data accuracy assessment using enterprise architecture. *Enterprise Information Systems*, *5*(1), 37–58. https://doi.org/10.1080/17517575.2010.507878

Naumann, F., & Rolker, C. (2005). *Assessment methods for information quality criteria*. Humboldt-Universität zu Berlin.

Nguyen, B. X., Dinneen, J. D., & Luczak-Roesch, M. (2019). Enriching bibliographic data by combining string matching and the Wikidata knowledge graph to improve the measurement of international research collaboration. *arXiv preprint*, arXiv:1905.13226. https://doi.org/10.48550/arXiv.1905.13226

Nguyen, B. X., Dinneen, J. D., & Luczak-Roesch, M. (2020). A novel method for resolving and completing authors' country affiliation data in bibliographic records. *Journal of Data and Information Science*, *5*(3), 97–115. https://doi.org/10.2478/jdis-2020-0020

Nguyen, B. X., Dinneen, J. D., & Luczak-Roesch, M. (2022). Research topics in the international research collaboration measurement domain. *Data Science and Informetrics*, *2*(1), 1–9.

Nguyen, B. X., Luczak-Roesch, M., & Dinneen, J. D. (2019). Exploring the effects of data set choice on measuring international research collaboration: An example using the ACM digital library and Microsoft Academic Graph. *arXiv preprint*, arXiv:1905.12834. https://doi.org/10.48550/arXiv.1905.12834

Nguyen, B. X., Luczak-Roesch, M., Dinneen, J. D., & Larivière, V. (2022). Assessing the quality of bibliographic data sources for measuring international research collaboration. *Zenodo*. https://doi.org/10.5281/zenodo.7016728

Olensky, M. (2015). Data accuracy in bibliometric data sources and its impact on citation matching. *Doctoral dissertation*. Humboldt-Universität zu Berlin (Germany). Retrieved April 1, 2020, from https://edoc.hu-berlin.de/dissertationen/olensky-marlies-2014-12-17/PDF/olensky.pdf

Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: Re-discovering the ecosystem of scientific information. *El Profesional de la Información*, *27*(2), 420–431. https://doi.org/10.3145/epi.2018.mar.21

Peters, M. A. (2006). The rise of global science and the emerging political economy of international research collaborations. *European Journal of Education*, *41*(2), 225–244. https://doi.org/10.1111/j.1465-3435.2006.00257.x

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

Rajan, N. S., Gouripeddi, R., Mo, P., Madsen, R. K., & Facelli, J. C. (2019). Towards a content agnostic computable knowledge repository for data quality assessment. *Computer Methods and Programs in Biomedicine*, *177*, 193–201. https://doi.org/10.1016/j.cmpb.2019.05.017, PubMed: 31319948

Redman, T. C., & Godfrey, B. A. (1996). *Data quality for the information age*. Artech House Publishers.

Schmoch, U., & Schubert, T. (2008). Are international co-publications an indicator for quality of scientific research? *Scientometrics*, *74*(3), 361–377. https://doi.org/10.1007/s11192-007-1818-5

Shen, Z., Ma, H., & Wang, K. (2018). A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations* (pp. 87–92). https://doi.org/10.18653/v1/P18-4015

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., … Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246). https://doi.org/10.1145/2740908.2742839

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820–1833. https://doi.org/10.1002/asi.22695

Strotmann, A., & Zhao, D. (2015). An 80/20 data quality law for professional scientometrics? In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of the 15th International Conference of the International Society for*

Scientometrics and Informetrics. https://www.issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2015/

Thelwall, M., & Kousha, K. (2017). ResearchGate versus Google Scholar: Which finds more early citations? *Scientometrics*, *112*(2), 1125–1131. https://doi.org/10.1007/s11192-017-2400-4

Van Holt, T., Johnson, J. C., Moates, S., & Carley, K. M. (2016). The role of datasets on scientific influence within conflict research. *PLOS ONE*, *11*(4), e0154148. https://doi.org/10.1371/journal.pone.0154148, PubMed: 27124569

Visser, M., Van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112

Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics*, *62*(1), 3–26. https://doi.org/10.1007/s11192-005-0001-0

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the network of global science: Comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, *1*(2), 185–208. https://doi.org/10.1504/IJTG.2005.007050

Waltman, L., & Larivière, V. (2020). Special issue on bibliographic data sources. *Quantitative Science Studies*, *1*(1), 360–362. https://doi.org/10.1162/qss_e_00026

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Xiao, Y., Lu, L. Y., Liu, J. S., & Zhou, Z. (2014). Knowledge diffusion path analysis of data quality literature: A main path analysis. *Journal of Informetrics*, *8*(3), 594–605. https://doi.org/10.1016/j.joi.2014.05.001

Yuan, L., Hao, Y., Li, M., Bao, C., Li, J., & Wu, D. (2018). Who are the international research collaboration partners for China? A novel data perspective based on NSFC grants. *Scientometrics*, *116*(1), 401–422. https://doi.org/10.1007/s11192-018-2753-3

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, *7*(1), 63–93. https://doi.org/10.3233/SW-150175

Zhou, P., Zhong, Y., & Yu, M. (2013). A bibliometric investigation on China–UK collaboration in food and agriculture. *Scientometrics*, *97*(2), 267–285. https://doi.org/10.1007/s11192-012-0947-7

Zhu, H., & Wu, H. (2011). Quality of data standards: Framework and illustration using XBRL taxonomy and instances. *Electronic Markets*, *21*(2), 129–139. https://doi.org/10.1007/s12525-011-0060-4