algorithms

MDPI

Article

How to Inspect and Measure Data Quality about Scientific Publications: Use Case of Wikipedia and CRIS Databases

Otmane Azeroual [1,*] and Włodzimierz Lewoniewski [2]

[1] German Centre for Higher Education Research and Science Studies (DZHW), 10117 Berlin, Germany
[2] Department of Information Systems, Poznań University of Economics and Business, 61-875 Poznań, Poland; wlodzimierz.lewoniewski@ue.poznan.pl
* Correspondence: azeroual@dzhw.eu

Received: 10 March 2020; Accepted: 24 April 2020; Published: 26 April 2020

Abstract: The quality assurance of publication data in collaborative knowledge bases and in current research information systems (CRIS) becomes more and more relevant by the use of freely available spatial information in different application scenarios. When integrating this data into CRIS, it is necessary to be able to recognize and assess their quality. Only then is it possible to compile a result from the available data that fulfills its purpose for the user, namely to deliver reliable data and information. This paper discussed the quality problems of source metadata in Wikipedia and CRIS. Based on real data from over 40 million Wikipedia articles in various languages, we performed preliminary quality analysis of the metadata of scientific publications using a data quality tool. So far, no data quality measurements have been programmed with Python to assess the quality of metadata from scientific publications in Wikipedia and CRIS. With this in mind, we programmed the methods and algorithms as code, but presented it in the form of pseudocode in this paper to measure the quality related to objective data quality dimensions such as completeness, correctness, consistency, and timeliness. This was prepared as a macro service so that the users can use the measurement results with the program code to make a statement about their scientific publications metadata so that the management can rely on high-quality data when making decisions.

Keywords: Wikipedia; current research information systems (CRIS); publications data; data quality; objective quality dimensions; research data processing; data management; data analysis; data measurement; completeness; consistency; correctness; timeliness; efficient decision-making

1. Introduction

Information and communication technologies are not the only ones that play an important role in all aspects of modern society [1]. But also the processing of electronic research data by the institutions. Research data are an essential part of the operational processes of scientific organizations. They also form the basis for decisions. In recent years, research data at the institution level has become accessible to researchers in many countries [2]. There is also an explosion of large data—various forms of establishment-level information that are typically created for business purposes [2]. In most facilities, incorrect research data only come to light in the current research information systems (CRIS) (The nomenclature for research information systems is more or less not standardized, including RIMS (Research Information Management System), RIS (Research Information System), RNS (Research Network System), RPS (Research Profiling System) or FAR (Faculty Activity Reporting). In this paper, the preferred term is CRIS (Current Research Information System) because it is widespread in European countries. A CRIS is a database for the collection, management and provision of research information

(e.g., publication data, personal data, project data, etc.)). The amount of data that accumulates in all phases of the research data cycle is growing, but should nevertheless be processed as quickly as possible. For this, the quality of the research data, as well as their usefulness and interpretation, is of enormous importance [3]. Because the data quality problems in the source data can reduce the data quality and thus the effective use of the data, the source data must be freed from data problems during the integration process [4]. So that an analysis can be carried out at all, data must be brought into the appropriate form, a process that is also referred to as data profiling [5]. There to finding data quality problems and cleansing the data, the term also includes testing during the integration of heterogeneous data sources. These can also be used data from publicly available data sources such as Wikipedia, where anyone can provide a reference in different ways in order to confirm the facts contained in the articles.

Nowadays Wikipedia is one of the most popular sources of knowledge in the world. This encyclopedia has over 51 million articles in over 300 language versions. Anyone can edit content on Wikipedia, so people with different education, experience and knowledge work on content on various topics. Authors of Wikipedia articles are not required to prove their competences in certain areas, we can expect that part of the content may be of low quality. Additionally, Wikipedia article in each language version is usually edited separately, so we can observe differences in the content on the same topic between languages [6].

Wikipedia community has created a grading scheme for articles that allows them to assess their quality based on specific criteria. One of the most important criteria is the presence of sources (references) in the articles, so readers can be able to confirm the information presented in the articles [7–10]. The sources also must meet certain quality criteria—they must be reliable, independent, published with a reputation for fact-checking and accuracy [11,12].

To assess the quality of references in Wikipedia, it is necessary to have information about it, such as author(s), title, place of the publication, special identifiers (such as ISBN, DOI) and others—we will call it metadata of the source (or metadata of the reference). Depending on the knowledge and experience of each Wikipedia user references can be described in various ways. So, the same reference can have different meta-data, including situations where it was entered incorrectly. Errors and incomplete metadata about Wikipedia references may impede their assessment as the sources of content. For instance, there are special tools for assessment of the scientific publications such as Altmetric [13], PlumX [14] and others. Such tools can assess publication based on a special identifier (mostly DOI). Other tools can assess the webpage source based on URL: SISTRIX Toolbox for SEO metrics [15], Nibbler, CheckTrust and others [16]. So, when some of the references described using the wrong metadata, it can be problematic to assess the quality of this source correctly. The goal of the paper is to describe a method to assess the quality of the metadata of the Wikipedia references. Moreover, in this paper, we showed the result of the assessment of the references on real data from Wikipedia in 2019.

2. Related Work

There are various studies related to analysis of the Wikipedia references. One of the studies showed that a large number of academic and peer-reviewed publications used as sources in this collaborative knowledge base [17]. Other works showed that Wikipedia preferred referenced with higher academic status and accessibility of journal [9,18]. There are also studies in a similar direction that concluded that Wikipedia can help assess the impact of scientific publications [19,20]. Moreover based on Wikipedia references it is possible to identify the most influential journals [21]. There is also research that showed how special identifiers can be used to find and unify similar sources with different metadata in Wikipedia [8]. For instance, it can be useful when Wikipedia users provide some mistakes in metadata, but also when they translate and provide titles of the publications on their local languages. It is also important to note that in Wikipedia there are special tools for adding and editing references in articles with limited metadata [22]. There are also publications, which take into account

the number of references to automatically assess the quality of information in Wikipedia articles [23], including approaches that used machine learning algorithms [24,25] or synthetic measure [26].

All previously mentioned studies did not analyze the quality of metadata of the Wikipedia references. One of the most relevant projects—WikiCite, which is an initiative to create a bibliographic database based on Wikidata [24]. The goals of this project include the improvement of citations in Wikipedia and other Wikimedia projects. However, not all citations from Wikipedia articles are extracted to Wikidata.

Assessment of the quality of the bibliographic metadata is a known problem. However, there are few existing studies that are focused on quality assessment of metadata in other resources. Based on literature research, the same environment is examined and the focus is on ensuring the data quality of metadata in institutional information systems such as CRIS (e.g., Pure, Converis, Symplectic Elements, etc.).

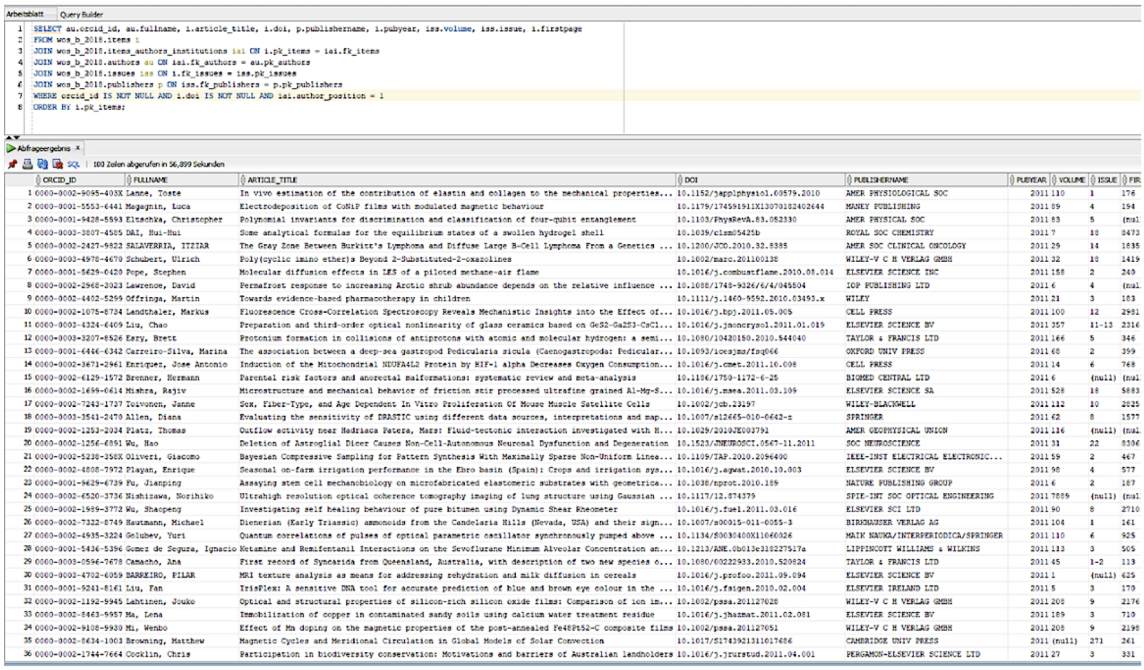
In research institutions, bad data quality can affect different areas. An improvement in data quality is therefore desirable and often meets with little resistance. However, some have found that improving data quality can be complex and time-consuming, but the opposite has to be analyzed and known only about the source of the data quality issues. This can be resolved using data profiling and data cleansing. Because only if the cause is remedied, a lasting improvement of the data quality can be achieved. For more details in this particular area, see the works from [5,27–30].

The collection of publication data in CRIS can lead to quality errors. If errors are already made when collecting the data, this can have many negative effects [30]. Spelling errors and typos, incorrect values or missing and inconsistent values are just a few examples of errors in collecting research information into CRIS [31]. To ensure data quality, continuous analysis of research information during its integration into CRIS is required [29]. For CRIS, high data quality is one of the main criteria that determine whether the project is successful and the resulting information is complete, correct and consistent.

3. Data Quality Challenges in References Data

If research information is incorrect, complete or consistent, it may result in significant organizational consequences. With increasing data volume and the number of source systems, it becomes increasingly difficult to meet the requirements for data acquisition and transformation processes. With both manual research information and automated data collection processes, increasing the amount of data can lead to more errors. The type and number of users can also have an impact on data quality.

In this section, we want to highlight the data quality issues from the reference data in Wikipedia and the publication data from Web of Science. For this purpose, methods are presented to make these problems of data quality recognizable and subsequently evaluable. Recently, the occurring data quality problems of the publication data from the Web of Science for the year 2018 were analyzed on behalf of the German Centre for Higher Education Research and Science Studies (DZHW) and the possible errors were categorized. Figure 1 illustrates the observational publication data from the Web of Science.



```
SELECT au.orcid_id, au.fullname, i.article_title, i.doi, p.publishername, i.pubyear, iss.volume, iss.issue, i.firstpage
FROM wos_b_2018.items i
JOIN wos_b_2018.items_authors_institutions iai ON i.pk_items = iai.fk_items
JOIN wos_b_2018.authors au ON iai.fk_authors = au.pk_authors
JOIN wos_b_2018.issues iss ON i.fk_issues = iss.pk_issues
JOIN wos_b_2018.publishers p ON iss.fk_publishers = p.pk_publishers
WHERE orcid_id IS NOT NULL AND i.doi IS NOT NULL AND iai.author_position = 1
ORDER BY i.pk_items;
```

100 Zeilen abgerufen in 56,899 Sekunden

ORCID_ID	FULLNAME	ARTICLE_TITLE	DOI	PUBLISHERNAME	PUBYEAR	VOLUME	ISSUE	FIR
1 0000-0002-9095-493X	Lanne, Toste	In vivo estimation of the contribution of elastin and collagen to the mechanical properties...	10.1152/japplphysiol.00579.2010	AMER PHYSIOLOGICAL SOC	2011	110	1	176
2 0000-0001-5553-6441	Magagnin, Luca	Electrodeposition of CoNiP films with modulated magnetic behaviour	10.1179/174591911X13079182402644	MANEY PUBLISHING	2011	09	4	194
3 0000-0001-9428-5593	Eltschka, Christopher	Polynomial invariants for discrimination and classification of four-qubit entanglement	10.1103/PhysRevA.83.052330	AMER PHYSICAL SOC	2011	83	5	(nul
4 0000-0003-3007-4505	DAI, Hui-Hui	Some analytical formulas for the equilibrium states of a swollen hydrogel shell	10.1039/c1sm05425b	ROYAL SOC CHEMISTRY	2011	7	10	3473
5 0000-0002-2427-5822	SALAVERRIA, ITZIAR	The Gray Zone Between Burkitt's Lymphoma and Diffuse Large B-Cell Lymphoma From a Genetics ...	10.1200/JCO.2010.32.8385	AMER SOC CLINICAL ONCOLOGY	2011	29	14	1835
6 0000-0002-4970-4670	Schubert, Ulrich	Poly(cyclic imino ether)s Beyond 2-Substituted-2-oxazolines	10.1002/marc.201100138	WILEY-V C H VERLAG GMBH	2011	32	10	1419
7 0000-0001-5629-0420	Pope, Stephen	Molecular diffusion effects in LES of a piloted methane-air flame	10.1016/j.combustflame.2010.08.014	ELSEVIER SCIENCE INC	2011	158	2	249
8 0000-0002-2960-2023	Lawrence, David	Permafrost response to increasing Arctic shrub abundance depends on the relative influence ...	10.1088/1748-9326/6/4/045504	IOP PUBLISHING LTD	2011	6	4	(nul
9 0000-0002-4402-5299	Offringa, Martin	Towards evidence-based pharmacotherapy in children	10.1111/j.1460-9592.2010.03493.x	WILEY	2011	21	3	183
10 0000-0002-1075-8734	Landthaler, Markus	Fluorescence Cross-Correlation Spectroscopy Reveals Mechanistic Insights into the Effect of...	10.1016/j.bpj.2011.05.005	CELL PRESS	2011	100	12	2931
11 0000-0003-4324-4409	Liu, Chao	Preparation and third-order optical nonlinearity of glass ceramics based on GeS2-Ga2S3-CsCl...	10.1016/j.jnoncrysol.2011.01.019	ELSEVIER SCIENCE BV	2011	357	11-13	2316
12 0000-0003-3207-8826	Kery, Brett	Proteolysis formation in collisions of antiprotons with atomic and molecular hydrogen: a semi...	10.1080/10420150.2010.544040	TAYLOR & FRANCIS LTD	2011	166	5	346
13 0000-0001-6446-6342	Carreiro-Silva, Marina	The association between a deep-sea gastropod Pedicularia sicula (Caenogastropoda: Pedicular...	10.1093/icesjms/fsq066	OXFORD UNIV PRESS	2011	68	2	359
14 0000-0002-3671-2961	Enriquez, Jose Antonio	Induction of the Mitochondrial NDUFA4L2 Protein by HIF-1 alpha Decreases Oxygen Consumption...	10.1016/j.cmet.2011.10.008	CELL PRESS	2011	14	6	768
15 0000-0002-6129-1572	Bronner, Hermann	Parental risk factors and anorectal malformations: systematic review and meta-analysis	10.1186/1750-1172-6-25	BIOMED CENTRAL LTD	2011	6	(null)	(nul
16 0000-0002-1699-0414	Mishra, Rajiv	Microstructure and mechanical behavior of friction stir processed ultrafine grained Al-Mg-S...	10.1016/j.msea.2011.03.109	ELSEVIER SCIENCE SA	2011	528	18	5883
17 0000-0002-7243-1727	Toivonen, Janne	Sex, Fiber-Type, and Age Dependent In Vitro Proliferation Of Mouse Muscle Satellite Cells	10.1002/jcb.23197	WILEY-BLACKWELL	2011	112	10	2925
18 0000-0003-3541-2470	Allen, Diana	Evaluating the sensitivity of SPASTIC using different data sources, interpretations and map...	10.1007/s12665-010-0642-z	SPRINGER	2011	62	8	1577
19 0000-0002-1253-2024	Platz, Thomas	Outflow activity near Hadriaca Patera, Mars: Fluid-tectonic interaction investigated with H...	10.1029/2010JE003791	AMER GEOPHYSICAL UNION	2011	116	(null)	(nul
20 0000-0002-1256-6891	Wu, Hao	Deletion of Astroglial Dicer Causes Non-Cell-Autonomous Neuronal Dysfunction and Degeneration	10.1523/JNEUROSCI.0567-11.2011	SOC NEUROSCIENCE	2011	31	22	8336
21 0000-0002-5238-356X	Oliveri, Giacomo	Bayesian Compressive Sampling for Pattern Synthesis With Maximally Sparse Non-Uniform Linea...	10.1109/TAP.2010.2096490	IEEE-INST ELECTRICAL ELECTRONIC...	2011	59	2	467
22 0000-0002-4808-7972	Playan, Enrique	Seasonal on-farm irrigation performance in the Ebro basin (Spain): Crops and irrigation sys...	10.1016/j.agwat.2010.10.003	ELSEVIER SCIENCE BV	2011	98	4	577
23 0000-0001-9629-6739	Fu, Jianping	Assaying stem cell mechanobiology on microfabricated elastomeric substrates with geometrica...	10.1038/nprot.2010.189	NATURE PUBLISHING GROUP	2011	6	2	187
24 0000-0002-6520-3736	Nishizawa, Norihiko	Ultrahigh resolution optical coherence tomography imaging of lung structure using Gaussian ...	10.1117/12.874379	SPIE-INT SOC OPTICAL ENGINEERING	2011	7889	(null)	(nul
25 0000-0002-1989-3772	Wu, Shaopeng	Investigating self healing behaviour of pure bitumen using Dynamic Shear Rheometer	10.1016/j.fuel.2011.03.016	ELSEVIER SCI LTD	2011	90	8	2710
26 0000-0002-7322-6749	Hautmann, Michael	Dienerian (Early Triassic) ammonoids from the Candelaria Hills (Nevada, USA) and their sign...	10.1007/s00015-011-0055-3	BIRKHAUSER VERLAG AG	2011	104	1	141
27 0000-0002-2935-3224	Selzner, Yuri	Quantum correlations of pulses of optical parametric oscillator synchronously pumped above ...	10.1134/S0030400X11060026	MAIK NAUKA/INTERPERIODICA/SPRINGER	2011	110	6	925
28 0000-0001-5436-5396	Gomez de Segura, Ignacio	Ketamine and Remifentanil Interactions on the Sevoflurane Minimum Alveolar Concentration an...	10.1213/ANE.0b013e318227517a	LIPPINCOTT WILLIAMS & WILKINS	2011	113	3	565
29 0000-0003-0596-7678	Camacho, Ana	First record of Syncarida from Queensland, Australia, with description of two new species o...	10.1080/00222933.2010.520624	TAYLOR & FRANCIS LTD	2011	45	1-2	113
30 0000-0002-4702-6059	BARREIRO, PILAR	MRI texture analysis as means for addressing rehydration and milk diffusion in cereals	10.1016/j.profoo.2011.09.094	ELSEVIER SCIENCE BV	2011	1	(null)	625
31 0000-0001-9261-8161	Liu, Fan	IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the ...	10.1016/j.fsigen.2010.02.004	ELSEVIER IRELAND LTD	2011	5	3	179
32 0000-0002-1192-5945	Lahtinen, Jouko	Optical and structural properties of silicon-rich silicon oxide films: Comparison of ion im...	10.1002/pssa.201127028	WILEY-V C H VERLAG GMBH	2011	208	9	2176
33 0000-0002-8463-5957	Ma, Lena	Immobilization of copper in contaminated sandy soils using calcium water treatment residue	10.1016/j.jhazmat.2011.02.081	ELSEVIER SCIENCE BV	2011	189	3	719
34 0000-0002-9108-5939	Mi, Wenbo	Effect of Mn doping on the magnetic properties of the post-annealed Fe48Pt52-C composite films	10.1002/pssa.201127051	WILEY-V C H VERLAG GMBH	2011	208	9	2198
35 0000-0002-8434-1003	Browning, Matthew	Magnetic Cycles and Meridional Circulation in Global Models of Solar Convection	10.1017/S1743921311017686	CAMBRIDGE UNIV PRESS	2011	(null)	271	241
36 0000-0002-1744-7664	Cocklin, Chris	Participation in biodiversity conservation: Motivations and barriers of Australian landholders	10.1016/j.jrurstud.2011.04.001	PERGAMON-ELSEVIER SCIENCE LTD	2011	27	3	331

Figure 1. Example of publication data from the Web of Science database.

The following data quality issues categories were identified in the publication data from Web of Science:

1. Name change after marriage: If Vivian Braun publishes after her marriage as Vivian Mathis, it is not clear that it is the same person.
2. Common names: J. Donna, Johnny Donna or T. J. Donna—how many people are behind it? Or is it always the same person, only with other name forms?
3. Incorrect capture of umlauts and special characters in proper names: This is a transcription problem (for example Mueller, Müller, Muller or André-Léonard, Andre-Leonard).
4. Uncertainty in the assignment of surname and first name: Boris Johnson can be both a Mr. Johnson and a Mr. Boris, who has not set a comma between last name and first name.
5. Names of authors are written differently in different countries, e.g., in the Russian space "Дмитрий Менделеев" and in the European area "Dmitri Mendeleev".
6. Faulty multiple registration of institutions: An institution is recorded with different name forms.
7. Incorrect, incomplete and inconsistent collection and order of institutional information of the authors: The registration of institutional information is not done in the correct order. That is, if the order of the information does not correspond to the hierarchical structure of the associated organization. Example: "Institute for Database Systems and Information Systems—Technische Universität Berlin".
8. Erroneous separation: When reading in, various institutions are incorrectly separated (e.g., "and" not recognized and therefore two detected as one).
9. Duplicate detection of Digital Object Identifiers (DOIs): A DOI is assigned in different articles or a publication has different DOIs.

In such collaborative knowledge bases as Wikipedia metadata about sources can be provided in different ways depending on the skills and experience of the users. Additionally, there is no central editorial there and not all of the provided data are checked regularly. Therefore, we can expect that various problems related to the quality of the source metadata may appear in this open encyclopedia. As was mentioned before, there are different possibilities to place information about sources in Wikipedia articles. One of the shortest way to add reference is to put some basic information about the publication or URL address of the page, where we can see more information about this source.

However, in the case of URL address, Wikipedia readers must follow this link to see more information about this source. At the same time, basic information about the source without URL forces the reader to search for the source of its real existence. An additional problem with the description of the source can be connected with different formats of the citations (such as the American Psychological Association (APA), Modern Language Association (MLA), Harvard), when it presented as an unstructured text. So, in this case, it is difficult to automatically extract source metadata about author, title, publisher and others for analysis of correctness or completeness.

In Wikipedia, sources can be described using special citation templates different parameters depending on the source type and language version. Such parameters include authors, title, publisher, publication date, URL, access date, DOI, ISBN and others. Table 1 shows the number of references with citations templates that contains special identifiers in the top 55 most developed languages versions of Wikipedia. The most developed languages were selected based on article count and depth as proposed in [26].

Table 1. Number of references with particular identifier in Wikipedia articles in various language versions. Source: own calculations in November 2019.

lang	arxiv	doi	isbn	issn	jstor	pmc	pmid	oclc
ar	7698	104,720	93,441	16,282	3505	15,581	71,194	5425
de	5132	66,030	216,212	38,389	1281	4513	17,866	2716
en	144,030	1,867,773	3,167,551	465,982	138,776	318,113	887,354	243,087
es	2376	115,305	311,940	73,463	5224	13,301	60,670	23,403
fr	10,850	129,779	459,534	96,769	3860	3711	31,264	51,462
it	1376	74,779	112,148	8187	1521	6362	43,363	10,082
ja	9080	89,714	309,612	22,748	1998	9354	37,150	8352
nl	25	7603	16,582	1468	148	1115	4887	191
pl	2289	124,713	418,603	64,675	1113	6396	47,455	23,985
pt	3765	75,128	158,321	32,182	2608	6528	31,635	10,566
ru	10,429	99,207	470,442	61,274	1783	6561	34,620	3798
sv	989	863,598	79,911	8695	243	1146	5907	2815
uk	3909	38,352	59,289	24,710	740	2555	14,133	2282
vi	7309	64,362	75,875	10,051	1961	9643	37,330	3527
zh	10,454	92,789	284,148	22,072	2341	9787	42,929	10,163
others	25,146	477,271	1,037,157	125,333	14,734	70,629	327,728	40,624
Total	**244,857**	**4,291,123**	**7,270,766**	**1,072,280**	**181,836**	**485,295**	**1,695 485**	**442,478**

Based on extracted data it is possible to get the information that was not directly provided in the citation templates. For example, based on special identifiers we can obtain the data about the publishers of each reference using other open bibliographic databases such as Crossref [32]. Thus, after extraction of metadata for over 1 million unique scientific publications with DOI numbers from over 40 million Wikipedia articles in considered languages, we found the top 20 most common publishers of the Wikipedia scientific references:

1. Elsevier BV (16.08%),
2. Wiley (11.12%),
3. Springer Science and Business Media LLC (6.04%)
4. Springer Nature (5.48%)
5. Oxford University Press (OUP) (5.03%)
6. Informa UK Limited (4.95%)
7. American Chemical Society (ACS) (4.27%)
8. JSTOR (2.70%)
9. SAGE Publications (2.48%)
10. Oxford University Press (2.37%)
11. Proceedings of the National Academy of Sciences (1.79%)
12. American Association for the Advancement of Science (AAAS) (1.73%)
13. IOP Publishing (1.48%)
14. Public Library of Science (PLoS) (1.34%)
15. BMJ (1.19%)
16. Ovid Technologies (Wolters Kluwer Health) (1.08%)
17. Cambridge University Press (CUP) (1.04%)
18. The Royal Society (0.87%)
19. University of Chicago Press (0.86%)
20. American Psychological Association (APA) (0.83%)

Citation templates can have different names depending on the source which we want to mention as a reference in Wikipedia article: journal, book, conference etc. For example, to give a reference to a book, in English Wikipedia we can use "Cite book" template [33]. However in German Wikipedia to cite the same source we need to use another template name—"Literatur" [34]. Figure 2 shows

the different names of the Wikipedia citation templates related to scientific publications in various languages. Additionally, each of the citation templates in each language version of Wikipedia can have its own set of permitted parameter names that can be used to describe the reference. Even the first and last name of each author of the source can be provided separately. Figure 3 showed the most popular parameters in the Wikipedia citations templates related to scientific publications in English Wikipedia (related figures for other language versions can be found in the supplementary web page [35]).

Figure 2. Wikipedia citation templates which describes scientific publications in different languages. Font size shows frequency of use. Source: own study in November 2019.

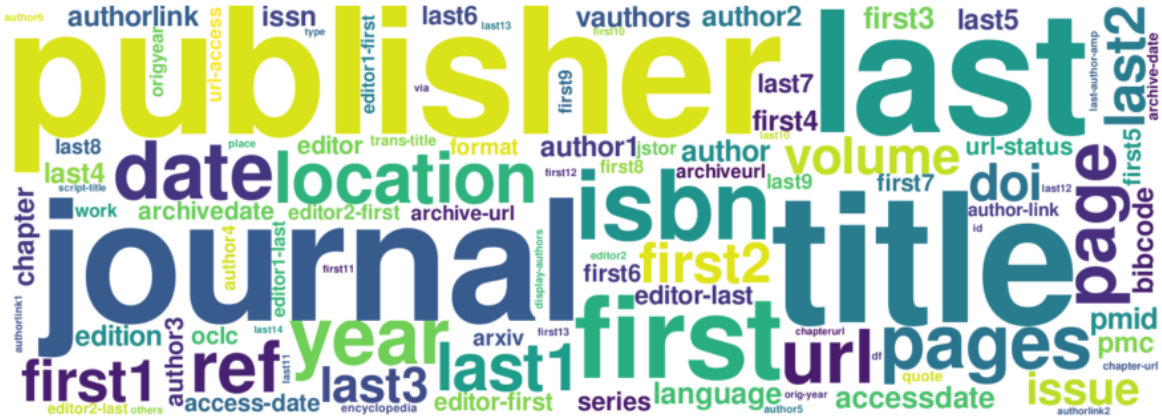


Figure 3. The most popular parameters in the Wikipedia citation templates related to scientific publications in English Wikipedia. Font size shows the frequency of use. Source: own study in November 2019.

Citation templates are more convenient for machine processing compared to unstructured text with source metadata. However, there may be problems with the quality of data in these templates, such as:

1. Incompleteness of the metadata. Not all parameters are filled by users in some cases.
2. Mismatches between parameters. For example, users can provide the DOI number related to other publications then it was provided in the URL parameter of the template.
3. Wrong value of the parameters such as title, authors and other parameters comparing to publisher database.
4. Invalid data format. Some parameters must have a special structure to be shown correctly in the Wikipedia article. For example, the "year" parameter cannot contain letters.
5. Uncertainty in the assignment of surname and first name. Sometimes parameter "first" consists of the last name of the authors, and conversely "last" parameter consists of the first name.

6. Redundancy of parameters values. For example in the parameter about the journal-title sometimes contains additional information, which must be placed in another parameter: journal = Scientific Data volume 3, article number: 150075.
7. Non-existent URL. This may be related to an incorrect value entry or removal of the destination web page after some time. Incorrect order of the publication authors.

4. Quality Analysis of References Data in Wikipedia and CRIS

In organizational database systems, there are usually very many data sources available. As an example, we will consider the reference data in Wikipedia as a data source. From the importance of data quality in organizational database systems, arises the question of how the quality of the reference data can be analyzed before it integrates into the information system and leads to a good decision. For that, it is necessary to examine the notion of data quality and introduce methods using the DataCleaner (https://datacleaner.org/) to analyze the data quality of reference data in Wikipedia.

We understand the concept of data quality both from the point of view of the provider of a data source and from the point of view of the user of the data source. The definition of data quality states that reference data in Wikipedia should be suitable for the purpose for which it was collected and generated.

Data quality is multi-dimensional and context-dependent [36]. It can not be defined by a single criterion, but from four different quality criteria (such as correctness, completeness, consistency and timeliness) together [3]. Only if these are considered together can the quality of the reference data be meaningfully described. According to [27] the quality of data must often be defined as the suitability of the data to be used for certain required usage goals, which must be error-free, complete, correct, up-to-date and inconsistent so that users can get better results.

High data quality can simplify integration on the one hand. On the other hand, quality issues become apparent in the course of data integration when comparing multiple data sources [37]. Data quality problems (such as duplicates, incomplete information, incorrect information, null values, etc.) during integration can be controlled by the process of data analysis and this is referred to as data profiling [5]. Data profiling is responsible for collecting as much information as possible about the data, making it easier to identify potential sources of error. In addition, data profiling analyzes the attributes at the instance level and captures as many metadata as possible. the attribute name, data type, value ranges, unique key, patterns, and domains [5].

Based on our practical example of reference data in Wikipedia with a subset of randomly selected 12,334 records, we performed data profiling using the DataCleaner tool to analyze and improve the quality problems of the reference data. With a tool like DataCleaner improves the data quality of existing data in a sustainable way and reduces the duplicates and redundancies within individual databases but also across an entire database. Furthermore, DataCleaner reduces the effort and cost of editing the data. The process of DataCleaner consists of three steps:

1. check data and identify errors
2. validate data
3. correct mistakes

Our example includes three columns *citation_id*, *parameter_name* and *parameter_value* and for that they have analyzed their data structure and data contents and used this analyzer as follows (see Figure 4):

- Completeness analyzer: Reference data can be checked with completeness analyzer, if all required fields are completely filled out.
- Unique Key: Unique key analysis is the finding of null values or duplicates of reference data.
- Character set distribution: This analysis checks and maps the text characterization of reference data to the corresponding affinity, e.g., Arabic, Latin, etc.

- Pattner finder: Possible patterns of reference data can be detected and resolved using pattern analysis and this could e.g., date format, e-mail addresses, etc. [5].
- Value Distribution: The value distribution can be used to identify all the values of a specific column and to examine which rows belong to specific values.

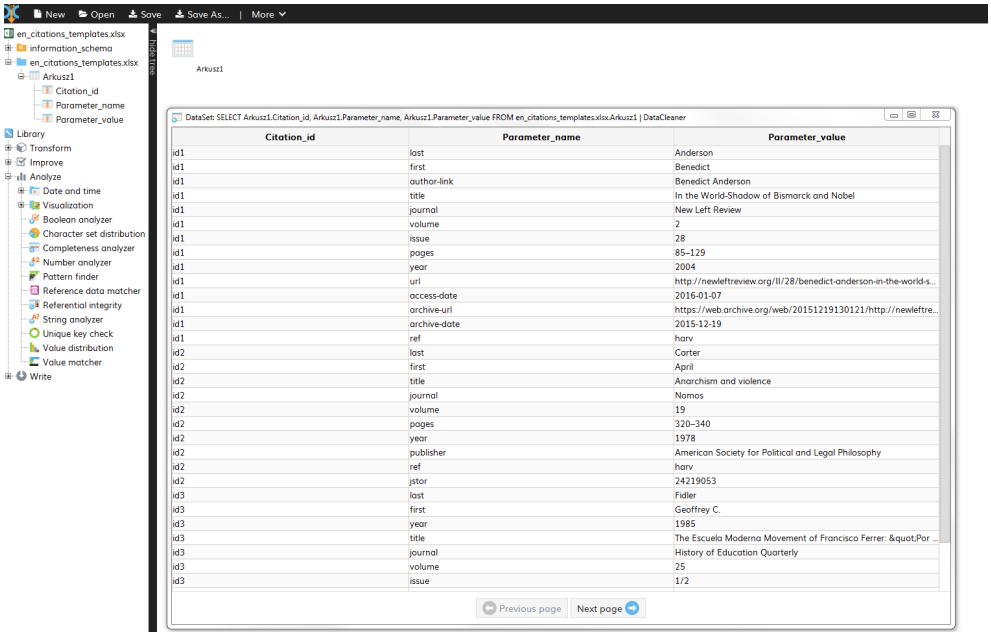


Figure 4. Example of the analyzed reference data in Wikipedia.

The results of our quality analysis for the reference data in Wikipedia with this analyzer are clearly shown in Figures 5 and 6. The findings are analyzed in tabular or graphical form.

Completeness Analyzer returns the number and percentage of NULL values for each column of the data set. Only explicit NULL values are taken into account, i.e., the absence of a value in a column. In our example, there are 119 incomplete records and this is a signal for non-meaningful data. Next, we used Unique Key Analysis, which determined that the 7336 attributes have a Unique Key among 12,333 rows. It is, therefore, necessary to create a rule that does not duplicate all entered values nor contains NULL values.

Character set distribution is useful to gain insight into the international aspects of our data. Here the question is answered: Can all our data be read and understood? Pattern finder discovers and identifies patterns or representations in our example by analyzing the attributes (see Figure 5). The values are searched for possible patterns and identified and correlated with the filtered-out patterns [5].

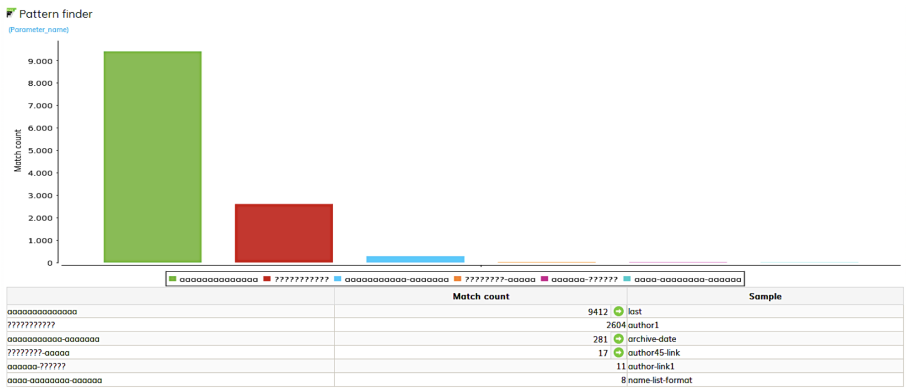


Figure 5. Pattern finder.

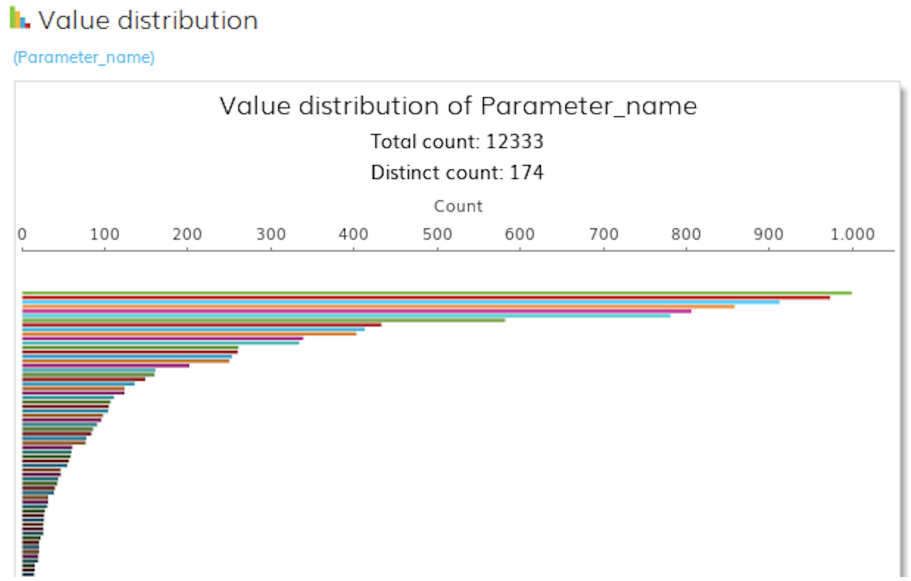


Figure 6. Value distribution.

Based on Figure 6, we used our example to find out which values represent most of the value distribution profiles and which rows belong to specific values. In summary, DataCleaner is an open-source application for data analysis, profiling and cleaning. With its help, the activities can monitor and manage their data quality. Data analysis by DataCleaner helps to evaluate and detect errors and deficiencies in the reference data, even before they are processed with the data, they are adopted in a database system. Only when the analyzed reference data are clean and free of defects can it be used to make high-quality decisions that add real value to the organization.

5. Quality Dimensions of the Source Metadata in Wikipedia and CRIS

In order to make the quality of the reference data measurable, certain quality dimensions must be assigned to the data. Current information technology has a wide range of properties that subdivide data quality into measurable areas. Table 7 below shows the most frequently mentioned data quality dimensions in the literature.

The quality dimensions differ according to subjective, objective and request-specific dimensions, which from the point of view of the user should reflect a structure of the requirements for the data.

- Subjective quality dimensions can only be assessed by the user. These include e.g., interpretability, relevance, reputation, simplicity of presentation, comprehensibility of presentation.
- Objective quality dimensions can, for the most part, be collected automatically and obtained from a data source, such as is possible when measuring the number of emergency zero values. This requires querying the data source in order to have values for processing. Objective quality criteria count e.g., completeness, timeliness, objectivity, correctness, security, consistency.
- The request-specific quality dimensions change from request to request, such as the latency, which depends, among other things, on the time of day and the complexity of the request.

At the core of the measurement of data quality, all quality dimensions (see Figure 7) point in the same direction and we have not compared them with each other because they want to achieve high data quality through constant adjustments to the system. In order for the quality of data to be and remain good, it is necessary to measure and monitor this quality.

System	**1. Accessibility**	The extent to which data are available or easily and quickly retrievable
	2. Ease of manipulation	The extent to which data is easy to manipulate and apply to different same format
Content	**3. Reputation**	The extent to which data are trusted or highly regarded in terms of their source or content
	4. Correctness	The extent to which data is correct and reliable
	5. Objectivity	The extent to which data are unbiased (unprejudiced) and impartial
	6. Believability	The extent to which information is regarded as true and credible
	7. Consistency	To what extent a data record may not have any contradictions in itself and with other data records.
Presentation	**8. Interpretability**	The extent to which data are in appropriate language and units and the data definitions are clear
	9. Concise representation	The extent to which data are compactly represented without being overwhelming (i.e., brief in presentation, yet complete and to the point)
	10. Ease of understanding	The extent to which data are clear without ambiguity and easily comprehended
Use	**11. Relevancy**	The extent to which data are applicable and helpful for the task at hand
	12. Appropriate amount of data	The extent to which the quantity or volume of available data is appropriate
	13. Completeness	The extent to which data are of sufficient breadth, depth, and scope for the task at hand
	14. Value-added	The extent to which data are beneficial and provide advantages from their use
	15. Timeliness	The extent to which the age of the data is appropriate for the task at hand

Figure 7. Data quality dimensions [36].

As part of the survey carried out in the papers [30,31], the relevant dimensions for checking and measuring the data quality in CRIS were examined and thereby determined what is of particular importance for institutions. For the respondents, the correctness, completeness, consistency and timeliness of the research information are very important. For this reason, this paper focuses only on the four objective data quality dimensions in the context of reference data in Wikipedia and CRIS. Therefore, data quality dimensions are selected that must be measurable on the one hand, and on the other hand, identified by users as particularly important in practice [31]. These are described in the following subsections with their simple metrics as follows. For each metric, we can achieve a degree from 0% to 100%.

5.1. Completeness

Data is complete if it is not missing and available at the specified times in the respective process steps. It is essential to determine against which amount the completeness is tested. Metadata of the reference completeness shows how many parameters have appropriate value among all defined or the most important parameters in the citation template.

$$Q_{completeness} = 1 - (\frac{Number\_of\_incomplete\_units}{Number\_of\_checked\_units}) \quad (1)$$

5.2. Correctness

Data is correct and error-free if it matches reality. In this case, the value of the parameters in the citation template must be compared to the primary source of the metadata (e.g., from the publisher website).

$$Q_{correctness} = 1 - (\frac{Number\_of\_incorrect\_data\_units}{Total\_number\_of\_data\_units}). \quad (2)$$

5.3. Consistency

Data is presented consistently if it is consistently mapped in the same way. Metadata of the references can be provided manually by different users in various language versions of Wikipedia, therefore often used citations can have a lower value of this measure compared to other sources.

$$Q_{consistency} = 1 - (\frac{Number\_of\_inconsistent\_units}{Number\_of\_consistency\_checks\_performed}) \quad (3)$$

5.4. Timeliness

Data is up-to-date if it reflects the actual property of the described object in a timely manner. References in the Wikipedia articles that were provided a relatively long time ago can have a lower value of this measure compared to recent ones.

$$Q_{timeliness}(W, A) = exp(-decline(A) \cdot age(W, A)). \quad (4)$$

6. Measures Modeling

The measurement of four described before objective data quality criteria will be explained in the following subsections and showed how data quality of the reference data in Wikipedia and CRIS can be measured. The aim was to present four metrics for the data quality dimensions, which enable an objective, targeted and largely automated measurement on different levels of aggregation (e.g., attribute values, tuples, etc.). Finally, there is an example of this measurement as written Python source code in pseudocode to measure the completeness, correctness, consistency and timeliness of the reference data before it is integrated into the CRIS. The CRIS employee can import his internal or external data source as a file at Python, copy the code provided and execute it as a script. This code then calculates the degree of completeness, correctness, consistency and timeliness for the user in order to have the most objective judgment possible. The program code including package will be available on the following website (https://github.com/OtmaneAzeroualDZHW/Forschungsinformationssysteme) and can be downloaded.

6.1. Measurement of Completeness

In order to measure completeness, we must analyze check if each parameter of the citation has some value. So, in the algorithm input we have records with certain number parameters such as author(s), title, DOI number, publication type, publication year. If some of the parameter values are empty ("NULL") we decrease completeness of this record by 0.25. When the algorithm iterates each record, it also counts how many values of the specific parameter in the column are empty, so in the end, we can also get information about completeness on each parameter. For instance, we can find how complete are citations in the entire dataset for DOI numbers.

In Algorithm 1 *records* means all records to be checked, while *records_num* is number of records. Each record contains fields that are related to different metadata: name, publication title, DOI number, date of publication etc. The $field\_emvalues$ variable counts the number of the empty values within the selected field, while $empty\_values$ means the number of empty fields of the current record. Fields that must be checked marked as fields, while the number of the fields as $fields\_num$.

Algorithm 1: Pseudocode of completeness measurement for source metadata

```
for record in records do
    for field in fields do
        if field is NULL then
            empty_values ++
            field_emvalues[field] ++
        end
    end
    record_completeness = 1 - empty_values/fields_num
    all_incomplete_values increment by empty_values
end
for field in fields do
    if field in field_emvalues then
        field_completeness = 1 - field_emvalues[field]/records_num
    else
        field_completeness = 1
    end
end
dataset_completeness = 1 - all_incomplete_values/(fields_num * records_num)
```

6.2. Measurement of Correctness

To measure the correctness of the citation metadata we need some point of reference to the so-called "golden standard", which is another dataset with complete and correct data. Therefore in the input to our algorithm, we must have two datasets: with analyzed metadata and dataset to which the first will be compared. As with completeness, the algorithm compares all parameters within each record and also within each parameter in the whole dataset. It is important to note that value can be incorrect due to changes in the parameter over time (such as the last name of the author). In that case, we need to check if there were other correct values before in "golden standard".

In Algorithm 2 *standard* is the golden standard database with metadata about all existing publications.

Algorithm 2: Pseudocode of correctness measurement for source metadata

```
for record in records do
    if record not in standard then
        record_correctness = 0
        for field in fields do
            field_incorrect[field] ++
            incorrect_values ++
        end
    else
        for field in fields do
            if field != standard[record][field] then
                field_incorrect[field] ++
                incorrect_values ++
            end
        end
    end
    record_correctness = 1 - incorrect_values/fields_num all_incorrect_values increase by incorrect_values
end
for field in fields do
    if field in fields then
        field_correctness = 1 - field_incorrect[field]/records_num
    else
        field_correctness = 1
    end
end
dataset_correctness = 1 - all_incorrect_values/(fields_num * records_num)
```

6.3. Measurement of Consistency

Consistency can be measured by taking into account at least two datasets with citation metadata. Here we do not have "golden standard", so the algorithm compares records between selected datasets and detects any discrepancies between parameter values. Similar to completeness and correctness, the Algorithm 3 measure consistency at the level of each record and at the level of specific parameter (column).

Algorithm 3: Pseudocode of consistency measurement for source metadata

```
for current_dataset in [dataset1, dataset2] do
    if current_dataset == dataset1 then
        comparison_dataset = dataset2
    else
        comparison_dataset = dataset1
    end
    for record in current_dataset do
        if record not in comparison_dataset then
            record_consistency = 0 for field in fields do
                field_inconsistent[field] ++
                inconsistent_values ++
            end
            else
                for field in fields do
                    if field != comparison_dataset[record][field] then
                        field_inconsistent[field] ++
                        inconsistent_values ++
                    end
                end
            end
        record_consistency = 1 - inconsistent_values/fields_num
        all_inconsistent_values increase by inconsistent_values
    end
    for field in fields do
        if field in field_incorrect then
            field_consistency = 1 - field_inconsistent[field]/records_num
        end
        else
            field_consistency = 1
        end
    end
    current_dataset_consistency = 1 - all_inconsistent_values/(fields_num * records_num)
end
```

6.4. Measurement of Timeliness

Some of the values of the parameters in citation metadata can be out-of-date. To measures timeliness we need to know, what are the parameters, than can be changed during the time. So the algorithm takes into account only selected parameters, such as the e-mail address of the author, affiliation of the authors, author name etc. In order to measure timeliness correctly, as an input algorithm must have a dataset with all possible values of the parameter of each record with citation metadata.

In Algorithm 4 *standard* is the golden standard database with metadata about all existing publications. Due to the fact, that not all fields can be changed during the time, the algorithm only check selected fields that are presented as $fields\_to\_check\_num$ in pseudocode. In case, when the record is not in the golden standard database, it is not possible to measure timeliness. $field\_values$ is a dictionary with a specific value of the filed as a key, and the value as a date of first a. The decay rate of the data value presented as *decline* and it is a static value that depends on $field\_name$. Age of the $field\_value$ market as *age*. Assessment for a particular field in each record takes $timfield$ variable.

Algorithm 4: Pseudocode of consistency timeliness for source metadata

```
for record in records do
    if record not in standard then
        record_timeliness = NULL
        norecords ++
    else
        timrec = 0 for field_name, field_value in fields_to_check do
            if field_value not in standard[record][field_name][field_values] then
                notassessed_values ++
            else
                if field_value != standard[record][field_name][last_id] then
                    timfieald = exp(decline[field_name]) * age(field_value)
                    assessed_values ++
                else
                    timfieald = 1
                    assessed_values ++
                end
            end
            timrec increase by timfieald
        end
        if assessed_values > 1 then
            record_timeliness = 1 - timrec/assessed_values
        else
            record_timeliness = NULL
        end
    end
    if record_timeliness != NULL then
        timeliness_sum increase by record_timeliness
        assessed_records ++
    end
end
dataset_timeliness = timeliness_sum/assessed_records
```

7. Discussion

To continuously analyze and measure all data quality problems of the reference data in Wikipedia and publication data in CRIS, in addition to this correction of accidentally conspicuous data errors [27,28]:

1. The identification and elimination of the respective sources of error,
2. Constant monitoring of the data and its quality,
3. Measures (such as pro-active measures) to prevent another mistake and
4. The regular control of new data errors.

It is not enough to clean up the research information only when it is integrated into CRIS, but it is necessary to communicate errors to the appropriate places so that they are already fixed in data sources.

This not only prevents the occurrence of the same data quality problems the next time the research information is loaded from the respective system into the the CRIS, but sensitizes the responsible IT person, e.g., during manual data entry.

In order to check the frequency of change of the research information from common values for publication data and reference data in Wikipedia, two aspects should be considered and this is explained as follows:

1. Trustworthiness of the data sources: Whether research information in the CRIS can achieve high quality depends on the quality of the respective data sources. The question to answer is whether a data source is trustworthy at all and whether it has a high-quality database from the outset that is easily overlooked. However, if data quality is serious, the data sources should also rate their trustworthiness [37].
2. Selection of research information: In order to be able to analyze the research information, it must first be sensibly selected with regard to the intended use. The right choice of research information is the first step towards high data quality. Which research information should be selected for integration depends on the individual needs of the scientific organization.

In order to assess the timeliness of the source the metadata we need to know, which of the values of the available parameters in the Wikipedia citation templates can be changed over time. For instance, there is no parameter related to contact details of the author (e.g., email address), at the same time there is an author name, publisher, URL address and other fields that can be compared with the related historical values. Additionally, each of the historical values must have an assigned date that shows when this value was updated according to the real state.

8. Conclusions

Our paper illustrated how data quality can be analyzed and measured in the context of scientific references in multilingual Wikipedia and CRIS. Systematically building the data quality problems were first explained in practice and performed the analysis using DataCleaner. Furthermore, the data quality was measured by the four objective metrics and we show how they can be measured using pseudocode. We also adjust the code for Python programming language.

So far, the results of our proposed solution could not be compared with the results of the other existing related solutions, as this does not exist in the literature. The solution presented here offers a novelty in research since the measurement of the objective criteria has never been programmed with Python to diagnose and evaluate the quality of metadata from the scientific publications in Wikipedia and CRIS. It is advisable to carry out these measurements continuously because quality measurement of data is not a one-time action, but rather has to be viewed as a permanent task. For this purpose, employees have to be made aware of data quality and motivated to produce it. This is the only way to ensure long-term and sustainable better data quality.

The results of our paper suggest that data analysis uses data profiling to examine, analyze, and summarize reference data in Wikipedia and CRIS. This provides a clear overview that allows organizations to better understand problems, risks, and general data quality trends. Data profiling enables organizations to gain and exploit important data-based insights, e.g., predictive decision-making [5].

For an optimal assessment and measurement of the data quality of reference data in Wikipedia and CRIS, the corresponding four data quality dimensions (correctness, completeness, consistency and timeliness) are required. As Azeroual [3] says, according to its measurement results, with these four objective dimensions in the area of CRIS "*the four selected dimensions enable an objective, effective and largely automated measurement within the CRIS. Their metrics have proven to be relatively easy to measure. In addition, these represent a particularly representative presentation of the reporting for the CRIS users and lead to an improved basis for decision-making. In this respect, the review of data quality must always be done with special regard to their context*".

In each language version of Wikipedia, we can meet with different names of the templates that describe the scientific references. Each of them can have its own set of permitted parameters and we found which of them are commonly used - title, first name, last name, special identifiers (DOI, ISBN, ISSN, JSTOR) and others. In this work, we showed how often special identifiers are used in over 40 million Wikipedia articles in different languages. Based on special identifiers we can obtain additional data about scientific publications from other databases. Thus we showed the top 20 most popular publishers in Wikipedia scientific references. Moreover, we can compare completeness, correctness, consistency and timeliness of metadata when we know how to clearly identify the same publications in different databases.

Finally, it can be said that the data quality requirements depend on the organizations and in particular on the users of the data. With the help of the procedure described in this paper, quality problems are recognized and corrected automatically at an early stage. In this way, the data quality is continuously monitored and improved. Deviations or errors are quickly recorded across systems, localized in a targeted manner and thus remedied more cost-effectively. In order to avoid the mostly cost-intensive reactive measures, a holistic data quality management process is required. This makes it possible to introduce and permanently guarantee quality in facilities as an overall target for data.

Author Contributions: O.A. and W.L. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roztocki, N.; Soja, P.; Weistroffer, H.R. The role of information and communication technologies in socioeconomic development: towards a multi-dimensional framework. *Inform. Tech. Dev.* **2019**, *25*, 171–183.
2. Bloom, N.; Lemos, R.; Sadun, R.; Scur, D.; Reenen, J.V. International Data on Measuring Management Practices. *Am. Econ. Rev.* **2016**, *106*, 152–156.
3. Azeroual, O.; Saake, G.; Wastl, J. Data measurement in research information systems: metrics for the evaluation of data quality. *Scientometrics* **2018**, *115*, 1271–1290.
4. Calì, A.; Calvanese, D.; De Giacomo, G.; Lenzerini, M. Data Integration under Integrity Constraints. In *Advanced Information Systems Engineering. CAiSE 2002*; Pidduck, A.B., Ozsu, M.T., Mylopoulos, J., Woo, C.C., Eds.; Springer: Berlin, Germany, 2002; Volume 2348, pp. 262–279.
5. Azeroual, O.; Saake, G.; Schallehn, E. Analyzing data quality issues in research information systems via data profiling. *Int. J. Inform. Manag.* **2018**, *41*, 50–56.
6. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Relative quality and popularity evaluation of multilingual Wikipedia articles. *Informatics* **2017**, *4*, 43.
7. Haigh, C.A. Wikipedia as an evidence source for nursing and healthcare students. *Nurse Educ. Today* **2011**, *31*, 135–139.
8. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Analysis of references across Wikipedia languages. In *Information and Software Technologies. ICIST 2017;* Damaševičius, R., Mikašytė, V., Eds.; Springer: Cham, Switzerland, 2017; Volume 756, pp. 561–573.
9. Nielsen, F.Å. Scientific citations in Wikipedia. *arXiv* **2007**, arXiv:0705.2106. Available online: https://arxiv.org/pdf/0705.2106.pdf (accessed on 18 November 2019).
10. Viégas, F.B.; Wattenberg, M.; McKeon, M.M. The hidden order of Wikipedia. In *Online Communities and Social Computing. OCSC 2007;* Schuler, D., Ed.; Springer: Berlin, Germany, 2007; Volume 4564, pp. 445–454.
11. Luyt, B.; Tan, D. Improving Wikipedia's credibility: References and citations in a sample of history articles. *J. Am. Soc. Inf. Sci. Tec.* **2010**, *61*, 715–722.
12. English Wikipedia. Wikipedia: Verifiability. Available online: https://en.wikipedia.org/wiki/Wikipedia:Verifiability (accessed on 15 November 2019).

13. Costas, R.; Zahedi, Z.; Wouters, P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *J. Am. Soc. Inf. Sci. Tec.* **2015**, *66*, 2003–2019.
14. Champieux, R. PlumX. *J. Med. Libr. Assoc.* **2015**, *103*, 63.
15. Lewoniewski, W.; Härting, R.C.; Węcel, K.; Reichstein, C.; Abramowicz, W. Application of SEO metrics to determine the quality of Wikipedia articles and their sources. In *Information and Software Technologies. ICIST 2018;* Damaševičius, R., Vasiljevienė, G., Eds.; Springer: Cham, Switzerland, 2018; Volume 920, pp. 139–152.
16. Redkina, N. Library Sites as Seen through the Lens of Web Analytics. *Automat. Doc. Math. Ling.* **2018**, *52*, 91–96.
17. Ford, H.; Sen, S.; Musicant, D.R.; Miller, N. Getting to the source: Where does Wikipedia get its information from? In Proceedings of the 9th International Symposium on oPen Collaboration, Hong Kong, China, 5–7 August 2013; ACM: New York, NY, USA, 5 August 2013; pp. 1–10.
18. Teplitskiy, M.; Lu, G.; Duede, E. Amplifying the impact of open access: Wikipedia and the diffusion of science. *J. Am. Soc. Inf. Sci. Tec.* **2017**, *68*, 2116–2127.
19. Evans, P.; Krauthammer, M. Exploring the use of social media to measure journal article impact. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 374.
20. Shuai, X.; Jiang, Z.; Liu, X.; Bollen, J. A comparative study of academic and Wikipedia ranking. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital libraries, Indianapolis, IN, USA, 22–26 July 2013; ACM: New York, NY, USA, 22 July 2013; pp. 25–28.
21. Jemielniak, D.; Masukume, G.; Wilamowski, M. The Most Influential Medical Journals According to Wikipedia: Quantitative Analysis. *J. Med. Internet. Res.* **2019**, *21*, e11429.
22. English Wikipedia. Help: Citation tools. Available online: https://en.wikipedia.org/wiki/Help:Citation_tools (accessed on 15 November 2019).
23. Lewoniewski, W. Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia. In *Business Information Systems Workshops. BIS 2018;* Abramowicz, W., Paschke, A., Eds.; Springer: Cham, Switzerland, 2018; Volume 339, pp. 619–633.
24. Warncke-Wang, M.; Cosley, D.; Riedl, J. Tell me more: an actionable quality model for Wikipedia. In Proceedings of the 9th International Symposium on Open Collaboration, Hong Kong, China, 5–7 August 2013; ACM: New York, NY, USA, 5 August 2013; pp. 1–10.
25. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Quality and importance of Wikipedia articles in different languages. In Proceedings of the International Conference on Information and Software Technologies, Druskininkai, Lithuania, 13–16 October 2016; Springer: Berlin, Germany, 2016; pp. 613–624.
26. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. *Computers* **2019**, *8*, 60. doi:10.3390/computers8030060.
27. Azeroual, O.; Abuosba, M. Improving the data quality in the research information systems. *Int. J. Comput. Sci. Inf. Secur.* **2017**, *15*, 82–86.
28. Azeroual, O.; Saake, G.; Abuosba, M. Data quality measures and data cleansing for research information systems. *J. Digit. Inform. Manag.* **2018**, *16*, 12–21.
29. Azeroual, O.; Saake, G.; Abuosba, M. ETL Best Practices for Data Quality Checks in RIS Databases. *Informatics* **2019**, *6*, 10.
30. Azeroual, O.; Schöpfel, J. Quality issues of CRIS data: an exploratory investigation with universities from twelve countries. *Publications* **2019**, *7*, 14.
31. Azeroual, O.; Saake, G.; Abuosba, M.; Schöpfel, J. Quality of Research Information in RIS Databases: A Multidimensional Approach. In *Business Information Systems. BIS 2019;* Abramowicz, W., Corchuelo, R., Eds.; Springer: Cham, Switzerland, 2019; Volume 353, pp. 337–349.
32. Crossref. Main Page. Available online: https://www.crossref.org/ (accessed on 23 November 2019).
33. English Wikipedia. Template: Cite book. Available online: https://en.wikipedia.org/wiki/Template:Cite_book (accessed on 2 December 2019).
34. German Wikipedia. Vorlage: Literatur. Available online: https://de.wikipedia.org/wiki/Vorlage:Literatur (accessed on 2 December 2019).
35. Data.Lewoniewski.info. The most popular parameters in Wikipedia citation templates related to scientific publications. Available online: http://data.lewoniewski.info/bis2020/ (accessed on 15 November 2019).

36. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inform. Syst.* **1996**, *12*, 5–33.
37. Batini, C.; Scannapieco, M. *Data and Information Quality: Dimensions, Principles and Techniques*; Springer: Berlin, Germany, 2016.