ARTICLE

Beyond VIAF

Wikidata as a Complementary Tool for Authority Control in Libraries

Carlo Bianchini, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo

ABSTRACT

This paper aims to investigate the reciprocal relationship between VIAF® and Wikidata and their possible roles in the semantic web environment. It deals with their data, their approach, their domain, and their stakeholders, with particular attention to identification as a fundamental goal of Universal Bibliographic Control. After examining interrelationships among VIAF, Wikidata, libraries and other GLAM institutions, a double approach is used to compare VIAF and Wikidata: first, a quantitative analysis of VIAF and Wikidata data on personal entities, presented in eight tables; and second, a qualitative comparison of several general characteristics, such as purpose, scope, organizational and theoretical approach, data harvesting and management (shown in table 9). Quantitative data and qualitative comparison show that VIAF and Wikidata are quite different in their purpose, scope, organizational and theoretical approach, data harvesting, and management. The study highlights the reciprocal role of VIAF and Wikidata and its helpfulness in the worldwide bibliographical context and in the semantic web environment and outlines new perspectives for research and cooperation.

INTRODUCTION

In 2011, the Library Linked Data Incubator Group, a W3C working group with the aim "to help increase global interoperability of library data on the Web," published its final report. Two interrelated issues were tackled in that milestone report: what libraries can do for the semantic web and what the semantic web can do for libraries. Linked data is an important asset for libraries as the "use of identifiers allows diverse descriptions to refer to the same thing. Through rich linkages with complementary data from trusted sources, libraries can increase the value of their own data beyond the sum of their sources taken individually."[1] So linked data greatly contribute to library cataloguing work not just for description of resources but also for their proper identification.

On the other hand, libraries have always created and curated a significant amount of valuable information assets and library authority data for names and subjects to help reduce "redundancy of bibliographic descriptions on the Web by clearly identifying key entities that are shared across Linked Data. This will also aid in the reduction of redundancy of metadata representing library holdings."[2]

The report opened a new way of thinking about Universal Bibliographic Control (UBC), a "world-wide system for control and exchange of bibliographic information," (https://archive.ifla.org/ubcim/ubcim-archive.htm) the purpose of which is "to make universally

Carlo Bianchini (carlo.bianchini@unipv.it) is Associate Professor, Department of Musicology and Cultural Heritage, University of Pavia. **Stefano Bargioni** (bargioni@pusc.it) is Deputy Director, Library of the Pontifical University Santa Croce (Rome). **Camillo Carlo Pellizzari di San Girolamo** (camillo.pellizzaridisangirolamo@sns.it) is graduate student, Department of Classics, University of Pisa and Scuola Normale Superiore.

https://doi.org/10.6017/ital.v40i2.12959

and promptly available, in a form which is internationally acceptable, basic bibliographic data on all publications in all countries."[3]

Exchanging information and data requires standards, at both the national and international level, for description, identification, and data format. Nowadays, a pillar of UBC is VIAF® (the Virtual International Authority File), a worldwide project designed by a few national libraries and run by OCLC, which combines multiple name authority files with the goal "to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web [https://www.viaf.org/]." It "clusters together the various forms of names for an entity" and has become "a major source for authority control and is becoming the collective reference source at the international level."[4]

VIAF is a fundamental tool for the identification of entities (people, locations, works, and expressions) relevant for the bibliographic universe. Yet, as it is based on the harvesting of data from authoritative national libraries spread all over the world, it has a top-down approach: libraries and services that are not VIAF sources can only refer to VIAF, but not actively cooperate with it, and, for its nature, VIAF cannot admit user cooperation. Therefore, on a global scale, a very large number of local libraries are excluded, and their data, collections, and specificities are, too. Furthermore, since the design and development of VIAF at the beginning of the 21st century, the semantic web environment has hugely evolved, and libraries are more and more required to act in new directions and to explore new forms of cooperation.[5]

Illien and Bourdon maintain not only that libraries "must now be careful to keep up their own interoperability," but also that they "would be well-advised to keep up or enter into dialogue with the most influential communities in the Web of data—smoothing out their own disputes in the meantime."[6] Moreover, they believe that "building collaborative authority registries linked to standardized identifiers is one of the fundamental cornerstones of the new Universal Bibliographic Control."[7]

Also, Dunsire and Willer suggest that a "smart UBC should strive to support all those who wish to think globally and act locally, with a better mix of bottom-up and top-down methodologies" as far as the "attempts to implement UBC as a worldwide system for the control and exchange of bibliographic information using top-down methodologies have only partially succeeded at global scale."[8]

As a result, a better integration of libraries into the semantic web seems to require the involvement of a larger group of stakeholders—such as non-national agencies, museums, archives, and users—and the adoption of a complementary bottom-up approach.

A new global actor of the semantic web has both a bottom-up and a very inclusive approach: Wikidata. Wikidata is a freely available hosted platform that anyone—including libraries—can use to create, publish, and use Linked Open Data (LOD). Since 2012, many users have been involved in a bottom-up approach to identity management in Wikidata. Furthermore, interest in and experience with the use of Wikidata to publish LOD among GLAM (galleries, libraries, archives, and museums) institutions is constantly increasing.[9]

The Wikidata role as an important tool for the identification of entities of any kind—not just those of traditional importance to GLAM—has likewise been increasingly recognized in recent years.[10]

So, two worldwide identification tools, two different backgrounds, two opposite approaches. Are they mutually exclusive, or integrable? Is one of them sufficient for libraries' needs, or do libraries need both? Which stakeholders are best served by VIAF? Which are best served by Wikidata?

This paper investigates the reciprocal relationship between VIAF and Wikidata and of their possible specific roles in the semantic web environment with respect to their approach, their domain, and their stakeholders, with particular attention to identification as a fundamental goal of UBC.

Relationship between VIAF and Libraries

VIAF gathers a huge quantity of authority data from more than 50 sources, listed in the home page of the project (https://viaf.org). Millions of records coming from national libraries and other institutions are continuously processed using algorithms based on the matching of data and bibliographic relationships with the goal of creating clusters of names (figure 1).[11]

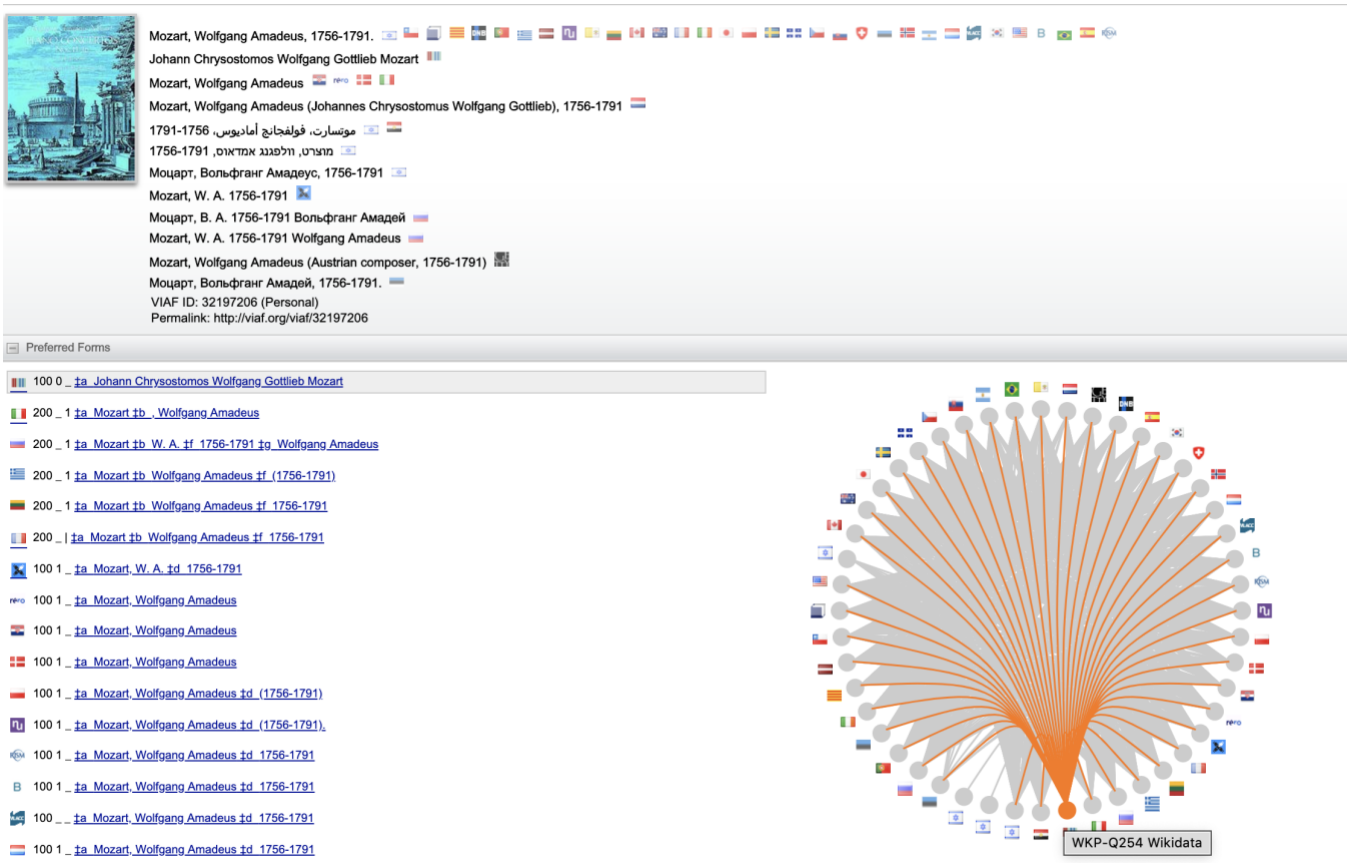

Figure 1. **VIAF cluster** for Wolfgang Amadeus Mozart

Clusters are usable in many services "to identify names, locations, works, and expressions while preserving regional preferences for language, spelling, and script" (https://www.oclc.org/en/viaf.html). Clusters may contain one or more IDs from VIAF sources. Furthermore, unique identifiers of clusters (a VIAF ID, e.g., https://viaf.org/viaf/7524651/) are freely reusable and reused by other institutions to add useful information to their catalogues, open up new paths of information for the end user, contribute local data to the linked data cloud, and much more.[12]

Data sources are selected and approved by the VIAF Council (see https://www.oclc.org/en/viaf/contributing.html), and may belong to two categories: VIAF Contributors, usually national LAM (libraries, archives, museums) agencies, admitted following very selective criteria; and Other Data Providers, i.e., "other selected sources (e.g., Wikipedia [*sic*]) that are not VIAF Contributor agencies."[13] Other Data Providers include ISNI and Wikidata (even if Wikidata is often confused with Wikipedia, as in the quotation above).[14] While Contributors are eligible to appoint a representative to the VIAF Council, Other Data Providers are not. So, VIAF is based on a rigid three-level hierarchical approach: VIAF, VIAF Contributors, and Other Data Providers.

All the other national and local institutions, i.e., relevant national data producers that are not national agencies, cannot provide data to VIAF; instead, they are expected to benefit from the use of VIAF IDs after performing a reconciliation process of their own data with VIAF IDs. However, benefits could be not completely satisfactory in term of quality of data: while VIAF deals with "widely-used authority files," it can be supposed that the libraries of non-national agencies need authority data more relevant on a local or specialistic basis.

Lastly, while VIAF guidelines state that VIAF participants should periodically send updated data to VIAF, it is not clear when and how VIAF retrieves and collects data from Other Data Providers (https://www.oclc.org/content/dam/oclc/viaf/VIAF%20Guidelines.pdf).

Relationships between Wikidata and Academic, Research, and Public Libraries

Wikidata was launched in 2012 by the Wikimedia Foundation as the central storage of the structured data from all Wikimedia Foundation projects; it is "a freely available hosted platform that anyone—including libraries—can use to create, publish, and use LOD."[15]

Wikidata stores stable and common information about entities, i.e., items and properties, and interlinks between different Wikimedia projects, in a form compliant with the RDF model (see https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer). Additionally, Wikidata uses triples and enriches them with *qualifiers* and *references*.[16] Qualifiers allow adding specifications about the validity of a statement (start/end date, precision, obsolescence, series ordinal, etc.); references are fundamental to justify the data, i.e., to document the authority data creator's reason for choosing the name or form of name on which a controlled access point is based.[17]

Wikidata uses the software Wikibase (https://wikiba.se/), which is "an open-source software suite for creating collaborative knowledge bases" whose "data model prioritizes language independence and knowledge diversity."

The Wikibase open-source software, which is currently used by more than thirty institutions, supports federated SPARQL queries.[18] Wikibase's approach and characteristics are particularly interesting for the library world. Gemeinsame Normdatei (GND) created a working group with Wikimedia Deutschland in order to "debate whether Wikibase is suitable for the needs of existing authority files coming from libraries" (https://wiki.dnb.de/display/GND/Authority+Control+meets+Wikibase); in March 2020 it was stated that the cooperation "has proven successful" and the current aim is to "develop a Wikibase-based GND and put it into use" (https://wiki.dnb.de/pages/viewpage.action?pageId=167019461). Similarly, the Bibliothèque nationale de France (BnF) and the Agence bibliographique de l'enseignement supérieur (Abes) launched the joint French National Entities File (FNE), which in 2019 carried out "a Proof of Concept to investigate the feasibility of using the software

infrastructure of Wikibase to support the FNE."[19] A synthesis of the proof of concept, published in July 2020, mentioned, among the decisions taken, the choice to develop FNE to build on Wikibase (https://www.transition-bibliographique.fr/wp-content/uploads/2020/07/synthese-preuve-concept-fne.pdf). FNE is scheduled to be launched in the next few years (https://f.hypotheses.org/wp-content/blogs.dir/2167/files/2020/02/20200128_8_VersUnFichierNationalDEntites.pdf).

Even more interestingly, between 2017 and 2018, OCLC explored a linked data Wikibase prototype; the final report shows, among other results, that "the building blocks of Wikibase can be used to create structured data with a precision that exceeds current library standards" and that "to populate knowledge graphs with library metadata, tools that facilitate the import and enhancement of data created elsewhere are recommended [. . . and . . .] the pilot underscored the need for interoperability between data sources, both for ingest and export."[20]

In late 2019, the IFLA Wikidata Working Group was formed "to explore and advocate for the use of and contribution to Wikidata by library and information professionals, the integration of Wikidata and Wikibase with library systems, and alignment of the Wikidata ontology with library metadata formats such as BIBFRAME, RDA, and MARC" (https://www.ifla.org/node/92837).

On the Wikimedia side, in 2019 the LD4-Wikidata Affinity Group (LD4 stands for "linked data for") was created by Hilary Thorsen, Wikimedian in Residence at Stanford University, to understand "how the library can contribute to and leverage Wikidata as a platform for publishing, linking, and enriching library linked data" (https://wiki.lyrasis.org/display/LD4P2/LD4-Wikidata+Affinity+Group).

Libraries' interest in Wikidata is usually focused on LOD and semantic discovery, not on authority control: "Libraries may each use different, unique, or select identifiers and authority control methods for disambiguation. Increasingly, Wikidata is becoming an important tool for synchronizing across identifiers like Virtual International Authority File (VIAF) and ORCID identifiers. Integrating awareness of Wikidata and its uses for enhancing metadata and linked open data will help advance a more interconnected research web."[21]

Identification is a key issue both in bibliographic control and in the semantic web environment, as John Riemer noted: "Recent examination of the efforts involved in what we have historically called authority control in the PCC community has led us to the conclusion that the primary emphasis should be on identity management."[22] As a matter of fact, Wikibase and Wikidata's approach to authority control and bibliographic description is quite new: not only does the traditional distinction between authority and bibliographic data disappear in a Wikibase description, but Wikidata is to be considered firstly as an *identity management tool* for any kind of entity.[23]

Relationship between VIAF and Wikidata

The first attempt of cooperation between VIAF and Wikidata goes back to 2012, when Maximilian Klein and Alex Kyrios, Wikipedians in Residence at OCLC and the British Library, respectively, developed a project to integrate authority data from the VIAF with English Wikipedia biographical articles. The project successfully "added authority data to hundreds of thousands of articles on the English Wikipedia," but above all showed that "linking of data represents an opportunity for libraries to present their traditionally siloed data, such as catalogue and authority records, in more openly accessible web platforms."[24] At the time, Wikidata was taking its first steps, but later authority data were successfully transferred from English Wikipedia to Wikidata.

At present, the connection between Wikidata and VIAF is very strong. Both VIAF and Wikidata are founded on a strict authority control that is built on a few cataloguing principles. In particular, both apply the principle that the authorized access point "for the name of an entity should be recorded as authority data along with identifiers for the entity and variant forms of name."[25] In addition, Wikidata is a data provider in VIAF, while VIAF IDs are constantly recorded and updated in Wikidata items. At present, Wikidata has 8,304,947 personal items, out of which 2,061,046 items have a VIAF ID. Moreover, each month a Wikidata bot (https://www.wikidata.org/wiki/User:KrBot) updates links in Wikidata items to redirected VIAF clusters and removes links to abandoned VIAF clusters.

The relevance of VIAF to the Wikidata information ecosystem is evident in the visualization of external identifiers in the items: VIAF IDs, represented on Wikidata by property P214 (https://www.wikidata.org/wiki/Property:P214), are automatically sorted as the first external identifier, preceded by the group of ISO standards and followed by the group of VIAF sources.[26] Using specific gadgets, i.e., enhancements of the edit interface, Wikidata registered users can add to a specific item the IDs of single VIAF sources extracting them from the VIAF ID(s) present in the item.[27]

Unfortunately, there is no automatic reciprocity between VIAF and Wikidata: when a Wikidata item gets a link to a VIAF cluster, VIAF does not have an automated way to add a reciprocal link to the Wikidata item. Likewise, when a VIAF cluster gets a link to a Wikidata item, Wikidata has no automatic way to add a reciprocal link to the VIAF cluster.

Another very important aspect of the VIAF-Wikidata relationship is that Wikidata uploads data from VIAF only by voluntary work of Wikidata users; and this approach applies to national library data, and to any other data, too. When available, VIAF IDs are typically one of the most important elements used by users to decide the identity of a Wikidata item.

Wikidata Controls on VIAF

In Wikidata, the use of constraints—i.e., rules that check the appropriate use of a property (https://www.wikidata.org/wiki/Help:Property_constraints_portal)—enables easy discovery of possible inconsistencies in statements, both in data and in external identifiers. Weekly, a Wikidata bot (https://www.wikidata.org/wiki/User:KrBot2) updates the database reports containing the constraint violations for each property, so that users can check the issues and try to fix them. Users can also check constraint violations in real time using the appropriate queries linked in the talk page of each property. As far as to VIAF IDs, two types of constraint-violations are particularly relevant both for the data entry and for the present paper:

- "Single value" violations, i.e., one item has two or more VIAF IDs. This means that either one or more VIAF IDs are not to be related to the item, so that the non-pertinent VIAF IDs should be removed from the Wikidata item or that more VIAF IDs exist for the same real entity, so that all the existing VIAF IDs must be kept in the Wikidata item until VIAF merges them. An example of a merge performed by VIAF, maybe on the basis of the correspondent Wikidata item, can be found in Iulius Rufinianus (https://www.wikidata.org/wiki/Q28131664), where the eight distinct VIAF IDs contained in the Wikidata item on September 24, 2019, have now been merged (https://www.wikidata.org/w/index.php?title=Q28131664&oldid=1001570078); in April 2021, the Wikidata item for Alaricus I (https://www.wikidata.org/wiki/Q102371) contains

four VIAF IDs (but there were ten on June 29, 2020; https://www.wikidata.org/w/index.php?title=Q102371&oldid=1220309663).

- "Unique value" violations, i.e., two or more Wikidata items have the same VIAF ID. This violation means not only an error on the Wikidata side, but it could imply an error in VIAF too. In the former, either one or more Wikidata items have a non-pertinent VIAF ID, to be removed; or the same entity is referred to by one or more Wikidata items, to be merged. In the latter, the VIAF ID conflates two or more distinct entities in one cluster. An example of conflation is the cluster at https://viaf.org/viaf/57898554/, where the painter Herbert E. Abrams (1920–2003; https://www.wikidata.org/wiki/Q4117019) and the physician Herbert L. Abrams (1920–2016; https://www.wikidata.org/wiki/Q23665535) conflate. In that case, Wikidata users can report the VIAF conflation error in the proper Wikidata error-report pages.[28]

In most cases just a few weeks are required for VIAF to merge clusters regarding the same entity when Wikidata includes them in the same item, but solutions to cases of conflation are fixed more slowly. While updates to VIAF clusters and IDs are obviously necessary and welcome, they are somehow risky for VIAF Contributors, providers, and users that base the consistency of their data on VIAF. So, national libraries could import incorrect data into their IDs and Wikidata could import wrong national libraries IDs referring to different entities into the same Wikidata item. There is no evidence that the error-report pages created and updated by Wikidata users are being systematically taken into consideration by VIAF to solve its conflations.

Recently, other issues in the use of VIAF as a source were raised when VIAF removed very important information about its cluster merging process, information that is no longer available to worldwide libraries and users. The VIAF data dump page (http://viaf.org/viaf/data) is refreshed monthly and, until April 2020, it included a *persist* file. For example, the February 2020 dump, viaf-20200203-persist-rdf.xml.gz, contained data about redirected clusters and—potentially—abandoned clusters as well. This information is essential to the prompt and safe synchronization of local data with VIAF clusters. In this dump, redirected clusters were described, for instance, as follows:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:p="http://viaf.org/viaf/abandonedViafRecord">
    <rdf:Description rdf:about="http://viaf.org/viaf/100035417">
        <owl:sameAs rdf:resource="http://viaf.org/viaf/67529853"/>
    </rdf:Description>
</rdf:RDF>
```

while any abandoned cluster (14,692,237 out of 24,030,176!) was erroneously described as follows:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:p="http://viaf.org/viaf/abandonedViafRecord"/>
```

This XML empty statement omits the specific information about the abandoned cluster. To obtain this invaluable information again, we filed a bug by email.[29] The decision taken was drastic: starting in May 2020, VIAF stopped including this information in its monthly dump, as stated at the bottom of the page itself.[30] As a result, the only recourse available to VIAF Contributors or any

other institution that would synchronize their authority records with VIAF identifiers is to rely on an external identification tool such as Wikidata!

MATERIALS AND METHODS

Any comparison between VIAF and Wikidata must consider their different content. VIAF contains personal name clusters, corporate name clusters, geographic name clusters, and work clusters, whereas Wikidata allows items to describe any kind of entity relevant in the universe of discourse of the users' data and irrespective of their bibliographic nature. Even if all kinds of VIAF clusters are relevant for bibliographic control, this study is limited to the analysis of personal name clusters in VIAF and of items having "instance of: human" (P31:Q5) in Wikidata, because they are largely the most represented in VIAF and they can be directly compared.[31] Some entities, such as mythological persons, legendary persons, etc., that are personal clusters in VIAF, are not treated as humans in Wikidata and belong to other instances (e.g., https://www.wikidata.org/wiki/Q95074).

A double approach was used to compare VIAF and Wikidata: First, data analyses of VIAF and Wikidata were performed, to compare VIAF clusters and Wikidata items and to investigate their reciprocal relationships (see the Data Analysis section). Second, a comparison of several general characteristics, such as scope, objectives, philosophy, authority control, and identification, was made based on respective websites and available literature to find and highlight differences and similarities.

Full VIAF dumps are available in native XML, RDF, MARC-21 XML, or ISO-2709 MARC-21 (http://viaf.org/viaf/data/). VIAF clusters were analyzed using an XML dump published on September 6, 2020 (http://viaf.org/viaf/data/viaf-20200906-clusters.xml.gz).

Full Wikidata dumps are available in XML, JSON, or RDF.[32] However, given the size of the entire dataset, it is much more convenient to create customized RDF dumps using the tool WDumper (https://wdumps.toolforge.org/). All the information (settings, dimension, and date of base dump) about dumps created using WDumper remains traced (https://wdumps.toolforge.org/dumps). Wikidata items were analyzed using a customized RDF dump updated to September 14, 2020 (https://wdumps.toolforge.org/dump/732). The customized dump contains all statements with non-deprecated values[33] present in items having both "instance of: human" (P31:Q5) in best rank and at least one value of "VIAF ID" (P214) in best rank.

Both dumps were parsed using three Perl scripts. Dumps and scripts were uploaded on Zenodo and are all available for analysis and reuse.[34] Perl scripts generate JSON data that are published on the HTML page http://catalogo.pusc.it/beyond_viaf/, where they are interpreted by JavaScript scripts in order to populate eight tables: three dedicated to VIAF (tables 1–3) and five to Wikidata (tables 4–8).

In order to select the statements to be analyzed in Wikidata items, three sets of relevant properties were found through three distinct SPARQL queries at the end of September 2020: VIAF members (table 5), authority controls related to libraries but not being VIAF members (table 6), and biographical dictionaries (table 7).[35] At the beginning of October 2020, another SPARQL query was performed to find all the personal items containing the authority controls related to libraries but not being VIAF members (table 6, column 4), without filtering the search to personal items having at least one value of "VIAF ID" (P214).[36]

DATA ANALYSIS: VIAF CLUSTERS AND WIKIDATA ITEMS

For this paper, two different versions of the data tables were produced: the first version, available at http://catalogo.pusc.it/beyond_viaf/, is a full, commented, and dynamic version of all the tables. Within that version, links to the acronyms (such as LC, DNB, SUDOC, etc.) of all the VIAF Contributors and Other Data Providers are available too. Static versions of these tables are included in this paper with commentary.

VIAF

VIAF has 22,099,715 personal clusters, half of which (50.90%; table 1, col. 2) are isolated clusters (i.e., they contain only one ID). The presence of isolated clusters is interesting because it means that those clusters are created based on data coming from just one source. What is more, the percentage of isolated clusters is much higher (71.19%; table 1, col. 12) if just VIAF Contributors are taken into account (i.e., excluding isolated clusters due to data from Other Data Providers, such as ISNI). It is worth noting that Other Data Providers can form isolated clusters, with the relevant exception of Wikidata (for which VIAF uses the acronym WKP), which never appears in isolated clusters (table 1, cols. 7 and 8).

Table 1. VIAF personal clusters by number of sources [adapted from http://catalogo.pusc.it/beyond_viaf/#tb1]

	1	2	3	4	5	6	7	8	9	10	11	12
No. of IDs in cluster	No. of personal clusters	Personal clusters (%)	No. of personal clusters (distinct sources)	Personal clusters (distinct sources) (%)	No. of personal clusters with ISNI	Personal clusters with ISNI (%)	No. of personal clusters with WKP	Personal clusters with WKP (%)	No. of personal clusters (ISNI and WKP excluded)	Personal clusters (excluding ISNI and WKP excluded) (%)	No. of personal clusters (based on contributors only)	Personal clusters (based on contributors only) (%)
0									136,216	0.62	399,645	1.81
1	11,249,563	50.90	11,256,242	50.93	131,880	1.17			15,839,017	71.67	15,733,035	71.19
2	5,777,125	26.14	5,774,003	26.13	3,865,370	66.91	652,260	11.29	2,773,048	12.55	2,664,443	12.06
3	1,832,439	8.29	1,830,689	8.28	1,432,743	78.19	334,149	18.24	1,077,636	4.88	1,060,891	4.80
4	910,532	4.12	910,125	4.12	783,258	86.02	205,296	22.55	615,344	2.78	607,859	2.75
5	583,014	2.64	583,658	2.64	534,772	91.73	159,547	27.37	405,750	1.84	401,371	1.82
6	406,902	1.84	407,234	1.84	385,913	94.84	128,279	31.53	290,631	1.32	287,093	1.30
7	299,300	1.35	299,710	1.36	289,538	96.74	104,668	34.97	217,298	0.98	215,216	0.97
8	226,283	1.02	226,631	1.03	221,378	97.83	86,648	38.29	168,154	0.76	166,753	0.75
9	176,307	0.80	176,688	0.80	173,604	98.47	73,486	41.68	132,340	0.60	130,762	0.59
10	139,375	0.63	139,761	0.63	137,881	98.93	63,487	45.55	105,205	0.48	104,451	0.47
11	110,264	0.50	110,686	0.50	109,304	99.13	55,505	50.34	83,904	0.38	82,890	0.38
12	88,684	0.40	88,851	0.40	88,047	99.28	49,956	56.33	64,664	0.29	63,931	0.29
13	70,531	0.32	70,574	0.32	70,054	99.32	45,021	63.83	49,115	0.22	48,146	0.22
14 or more	229,396	1.05	224,863	1.02	228,092	98.58	189,017	8.80	141,393	0.64	133,229	0.60
TOT	**22,099,715**	**100.00**	**22,099,715**	**100.00**	**8,451,834**	**38.24**	**2,147,319**	**9.72**	**22,099,715**	**100.00**	**22,099,715**	**100.00**

The total number of IDs present in VIAF clusters is 51,327,847 (table 2), distributed in 22,099,715 clusters; the most relevant Contributors include LC (7,266,628 IDs), DNB (5,677,731 IDs), SUDOC (3,278,189 IDs), and NTA (2,754,036 IDs), while the most relevant Other Data Providers are ISNI (8,455,814 IDs) and WKP (2,148,680 IDs) (table 2). Apart from LC and DNB, data about isolated clusters (table 2, col. 5) shows that the number of isolate clusters tends to slowly decrease over time and that clustering has improved: recently-added sources tend to have a higher share of isolated IDs. Another relevant figure is that sources in non-Latin alphabets usually have higher shares of isolated IDs.[37] So, a high number of isolated clusters may reveal a source that is partially in need to be gathered to existing clusters.

Table 2. VIAF personal clusters by source [adapted from http://catalogo.pusc.it/beyond_viaf/#tb2]

	1	2	3	4	5
VIAF contributor	No. of ids in personal clusters	No. of ids forming isolated personal clusters	Ids forming isolated personal clusters (%)	Ids forming isolated personal clusters on total isolated ids (%)	First year in VIAF
ISNI	8,455,814	131,880	1.56	1.17	2013
LC	7,266,628	2,119,073	29.16	18.84	2009
DNB	5,677,731	2,273,929	40.05	20.21	2009
SUDOC	3,278,189	855,167	26.09	7.60	2011
NTA	2,754,036	515,471	18.72	4.58	2012
WKP	2,148,680				2012
RERO	2,017,940	494,962	24.53	4.40	2010
BNF	1,810,665	210,674	11.64	1.87	2009
NUKAT	1,737,746	205,793	11.84	1.83	2009
BIBSYS	1,573,680	445,287	28.30	3.96	2012
J9U	1,546,097	192,049	12.42	1.71	2020
NII	1,499,092	298,078	19.88	2.65	2010
CAOONL	1,396,386	236,010	16.90	2.10	2018
PLWABN	1,345,430	408,045	30.33	3.63	2019
Other contributors	8,819,733	2,863,145	32.46	25.45	
TOT	**51,327,847**	**11,249,563**		**100.00**	

The histories of VIAF clusters, as contained in XML dumps, appear weird and incoherent. For example, many VIAF Contributors in their first year of appearance seem to have no additions and many removals (e.g., BAV row; for complete information see table 3 on the website at http://catalogo.pusc.it/beyond_viaf/#tb3). Incoherence is due to the absence of redirected and abandoned clusters in the data. Nevertheless, the histories allow us to reconstruct the year of first contribution of each source—an information otherwise unavailable—and to detect major changes in the data provided to VIAF by each source.[38]

Table 3. VIAF history of personal clusters by source [adapted from http://catalogo.pusc.it/beyond_viaf/#tb3]

VIAF contributors	2009	2010	2011	2016	2017	2018	2019	2020
BAV	0 | -1404	0 | -1623	0 | -1083	0 | -2832	0 | -1386	0 | -665	0 | -1626	304691 | -15312
BNE	216606 | -1298	31163 | -2439	54103 | -3226	23359 | -3093	12045 | -1354	15608 | -810	17414 | -1605	8298 | -3023
BNF	790723 | -7702	114496 | -8323	62096 | -3035	113601 | -13334	66458 | -8918	70207 | -4317	99441 | -8024	46204 | -50473
DE663						98167 | -74	5663 | -198	34551 | -2131
DNB	1280061 | -13437	152706 | -8202	352193 | -6746	309164 | -211223	406865 | -751543	403126 | -8882	730778 | -20050	320668 | -49128
GRATEVE								195387 | -870
ICCU	13172 | -42	782 | -56	197 | -20	4169 | -1131	1416 | -535	4597 | -422	13932 | -808	2484 | -1182
ISNI				87146 | -13994	876108 | -13378	235274 | -3681	275733 | -5667	192908 | -12264
J9U								1545026 | -9798
LC	3975816 | -13998	402276 | -12673	271528 | -5917	232754 | -17991	246794 | -13152	203008 | -7226	409379 | -13836	202334 | -44955
LIH						474464 | -124	77262 | -938	20873 | -2700
LNB				15672 | -1359	13860 | -1054	10720 | -428	4433 | -841	18506 | -2421
LNL				4088 | -554	1257 | -80	90 | -32	1003 | -75	377 | -180
PERSEUS				88 | -145	72 | -51	38 | -26	59 | -43	186 | -141
SIMACOB						57703 | -9	5890 | -270	4645 | -725
WKP				148961 | -9345	130742 | -5875	225744 | -4950	235417 | -10782	602758 | -18187

Wikidata

Wikidata has 8,304,947 personal items and 2,061,046 of them contain a VIAF ID. Usually one or more VIAF sources are extracted from the VIAF ID(s), so that 1,905,470 personal items containing VIAF ID have at least one VIAF source ID (table 4, col. 1). Wikidata records IDs from a wide range

of other resources, such as non-VIAF bibliographic agencies and biographical dictionaries (investigated in these tables), but also encyclopedias and various online databases. Considering the 2,061,046 items containing a VIAF ID, 684,367 items contain only one VIAF source ID (table 4, col. 1), but only 353,710 items contain only one among VIAF sources IDs and non-VIAF sources IDs and biographical dictionaries IDs (table 4, col. 15); so, more than 300,000 items containing only one VIAF source ID have at least one non-VIAF source ID and/or one biographical dictionary ID.

Table 4. Wikidata personal items (pers. it.) by number of IDs [adapted from http://catalogo.pusc.it/beyond_viaf/#tb4]

	1	2	5	6	7	8	9	10	12	13	14	15	16
No. of IDs in item	No. of pers. it. (VIAF sources)	Pers. it. (VIAF sources) (%)	No. of pers. it. (non-VIAF sources)	Pers. it. (non-VIAF sources) (%)	Pers. it. (non-VIAF sources) vs pers. it. with non-VIAF sources (% on total)	No. of pers. it. (VIAF sources + non-VIAF sources)	Pers. it. (VIAF sources + non-VIAF sources) (%) on total of column 1	Pers. it. (VIAF sources + non-VIAF sources) (%) on total of column 8	No. of pers. it. (biographies)	Pers. it. (biographies) on total pers. it. (%)	Pers. it. (biographies) on total pers. it. with biographies (%)	No. of pers. it. (all sources)	Pers. it. (all sources) on total pers. it. with all sources (%) on total of column 1
1	684,367	35.92	372,299	19.54	67.65	599,007	31.44	31.27	663,099	34.80	91.37	353,710	18.56
2	370,238	19.43	99,428	5.22	18.07	385,623	20.24	20.13	55,692	2.92	7.67	492,119	25.83
3	231,097	12.13	39,652	2.08	7.21	243,225	12.76	12.70	5,973	0.31	0.82	305,270	16.02
4	154,956	8.13	18,535	0.97	3.37	160,576	8.43	8.38	821	0.04	0.11	195,339	10.25
5	114,113	5.99	9,201	0.48	1.67	117,405	6.16	6.13	121	0.01	0.02	125,620	6.59
6	90,711	4.76	4,855	0.25	0.88	90,866	4.77	4.74	30	0.00	0.00	95,713	5.02
7	73,506	3.86	2,652	0.14	0.48	72,489	3.80	3.78	16	0.00	0.00	76,065	3.99
8	58,417	3.07	1,430	0.08	0.26	58,904	3.09	3.08	2	0.00	0.00	61,834	3.25
9	42,288	2.22	818	0.04	0.15	46,134	2.42	2.41				49,551	2.60
10	28,192	1.48	519	0.03	0.09	34,515	1.81	1.80	1	0.00	0.00	38,758	2.03
11	18,321	0.96	289	0.02	0.05	26,056	1.37	1.36				29,257	1.54
12	11,934	0.63	202	0.01	0.04	19,089	1.00	1.00				22,383	1.17
others	27,330	1.41	433	0.01	0.07	61,491	3.18	3.18	0	0.00	0.00	74,847	3.88
TOT	**1,905,470**	**100.00**	**550,313**	**28.87**	**100.00**	**1,915,380**	**100.00**	**100.00**	**725,755**	**38.08**	**100.00**	**1,920,466**	**100.00**

VIAF and Wikidata: A Data Comparison

From a quantitative perspective, Wikidata personal items (8,304,947) are 37.58% of VIAF personal clusters (22,099,715), while Wikidata personal items having a VIAF ID (2,061,046) are 9.26%. IDs from VIAF sources present in Wikidata personal items containing VIAF ID (6,292,778; table 5, col. 3) are 12.91% of IDs present in VIAF personal clusters (48,740,933; table 5, col. 4).

In the authors' opinion, quantitative confrontation between VIAF and Wikidata must be carefully considered. It could be argued that is a noticeable disadvantage of Wikidata with respect to VIAF, but it would be right only from a bibliographic control perspective and the other side of the coin must be examined too. As Wikidata represents *any kind* of entity relevant for its users (libraries, archives, museums, and many other stakeholders), VIAF contains just over a third of Wikidata items (37%). Furthermore, a very large part of the personal entities represented in Wikidata (at present, more than 6,200,000, i.e., about 75%) cannot rely on VIAF for identification purposes (for example, because Wikidata personal items can also represent singers, lawyers, pilots, and so on). It can be concluded that VIAF can be considered just one specialized source, in the domain of the semantic web and with respect to the objectives of Wikidata.

Considering single VIAF sources, Wikidata surpasses VIAF by number of IDs only in two cases, PERSEUS (135.18%) and SIMACOB (102.17%) (table 5, col. 5). This is possible because Wikidata and VIAF gather different sets of data from both the sources; the former uses sets of data obtained by its users, while the latter uses only data sent by the contributor. All the other sources, because of the absence of systematic imports, are much rarer in Wikidata than in VIAF.

Table 5. Wikidata personal items (pers. it.) by VIAF source [adapted from http://catalogo.pusc.it/beyond_viaf/#tb5]

	1	2	3	4	5	6
VIAF contributor	Type	Wikidata property	No. of ids in Wikidata pers. it. with VIAF id	No. of ids in VIAF personal clusters	Ratio Wikidata /VIAF	First year in VIAF
ISNI	non library	P214	1,136,260	8,455,814	13.44	2013
DNB	library	P227	1,012,493	5,677,731	17.83	2009
LC	library	P244	983,206	7,266,628	13.53	2009
NTA	library	P1006	480,580	2,754,036	17.45	2012
SUDOC	library	P269	431,919	3,278,189	13.18	2011
BNF	library	P268	428,792	1,810,665	23.68	2009
NUKAT	library	P1207	423,734	1,737,746	24.38	2009
NKC	library	P691	322,325	791,187	40.74	2009
BNE	library	P950	156,569	463,783	33.76	2009
BIBSYS	library	P1015	124,567	1,573,680	7.92	2012
NDL	library	P349	89,537	931,728	9.61	2012
BAV	library	P8034	88,448	305,905	28.91	2009
JPG	non library	P245	83,697	221,101	37.85	2009
SIMACOB	library	P1280	69,752	68,269	102.17	2018
PERSEUS	non library	P7041	1,660	1,228	135.18	2013
others			459,239	13,403,243	3.43	
TOT			**6,292,778**	**48,740,933**		

Table 6 and table 7 show authority control in Wikidata living aside VIAF. Wikidata contains some non-VIAF sources (usually non-national libraries or groups of libraries which couldn't become VIAF Contributors); their IDs in personal items having VIAF ID (894,161) are the 86.04% of their IDs in all personal items (958,206; table 6, col. 4), meaning that Wikidata provides a clusterization for more than 64,000 IDs (6%) probably corresponding to non-existent VIAF clusters (table 6, totals).

Table 6. Wikidata personal items (pers. it.) by non-VIAF sources [adapted from http://catalogo.pusc.it/beyond_viaf/#tb6]

	1	2	3	4	5	6
Non-VIAF sources	Wikidata property	No. of ids in pers. it. with VIAF id	Ids in pers. it. with VIAF id (% of total)	Total no. of ids in pers. it. (with or without VIAF id)	Ratio between column 2 on column 4 (%)	Ids in pers. it. with or without VIAF (% of total)
CERL Thesaurus ID	P1871	223,111	24.95	229,888	97.05	23.99
Open Library ID	P648	178,952	20.01	183,512	97.52	19.15
NLA Trove ID	P1315	76,026	8.50	86,367	88.03	9.01
SHARE Catalogue author ID	P3987	74,905	8.38	75,134	99.70	7.84
SELIBR ID	P906	70,852	7.92	71,244	99.45	7.44
University of Barcelona authority ID	P1580	53,194	5.95	53,693	99.07	5.60
Pontificia Università della Santa Croce ID	P5739	50,188	5.61	50,539	99.31	5.27
Angelicum ID	P5731	41,871	4.68	42,126	99.39	4.40
NLP ID (unique)	P1695	32,746	3.66	32,848	99.69	3.43
HKCAN ID	P5909	18,203	2.04	18,137	100.36	1.89
CoBiS author ID	P7865	17,242	1.93	17,359	99.33	1.81
BVMC person ID	P2799	13,241	1.48	14,038	94.32	1.47
DBLP ID	P2456	9,655	1.08	42,907	22.50	4.48
National Library of Wales Authority ID	P2966	7,428	0.83	10,965	67.74	1.14
others		26,547	2.95	29,449	90.15	3.06
TOT		**894,161**	**100.00**	**958,206**	**93.32**	**100.00**

Table 7. Wikidata personal items (pers. it.) by biographical dictionary [adapted from http://catalogo.pusc.it/beyond_viaf/#tb7]

	1	2	3
Biographical source	Wikidata property	No. of ids in pers. it. with VIAF id	Ids in pers. it. with VIAF id (%)
Deutsche Biographie ID	P7902	600,458	75.38
Oxford Dictionary of National Biography ID	P1415	42,879	5.38
Austrian Biographical Encylopedia ID	P6194	19,446	2.44
Treccani's Dizionario biografico degli italiani ID	P1986	19,339	2.43
American National Biography ID	P4823	17,769	2.23
Spanish Biographical Dictionary ID	P4459	12,953	1.63
SIKART ID	P781	7,227	0.91
Internetowy Polski Słownik Biograficzny ID	P8130	5,462	0.69
Australian Dictionary of Biography ID	P1907	5,289	0.66
Dictionary of Swedish National Biography ID	P3217	5,198	0.65
others		60,589	7.61
TOT		**796,609**	**100.00**

In general the presence of IDs of biographical dictionaries (796,609 IDs in total) in 725,755 personal items having VIAF ID helps significantly in the definition of authoritative dates of birth and death (table 7, total of column 2 and table 4, total of column 12).

A comparison between table 1, column 7, and table 2, row WKP (the acronym for Wikidata wrongly used by VIAF) shows that 2,147,319 clusters contain 2,148,680 WKP IDs; it means that, from a VIAF point of view, Wikidata duplicates are only 1,361. Furthermore, a comparison between the total and row 0 in table 8, col. 1, shows that 2,061,046 items contain at least one VIAF ID and that 2,037,638 items contain exactly one VIAF ID; so, items containing one or more VIAF duplicates are 23,408. As a result, it can be concluded that the percentage of duplicates in Wikidata is less than 0.01% and in VIAF is about 0.01%, so Wikidata is as trustworthy as VIAF.

VIAF and Wikidata not only are able to discover reciprocal duplicates, but also discover duplicates in VIAF sources, by a comparison between table 8, col. 3—containing the total number of the cases in which a VIAF source has at least one duplicate—and table 8, col. 5—containing the total number of the cases in which VIAF sources are duplicated. However, while duplicates recorded by VIAF are findable only by querying the monthly dumps using in-house–made programs, duplicates discovered by Wikidata are easily findable through SPARQL queries detecting single-value constraint violations.

Table 8. Wikidata personal items (pers. it.) by repeated VIAF sources and VIAF source IDs [adapted from http://catalogo.pusc.it/beyond_viaf/#tb8]

	1	2	3	4	5	6
No. of repeated VIAF source and non-VIAF sources	No. of pers. it. with VIAF id (repeated VIAF ids)	Items with repeated VIAF ids (%)	No. of pers. it. with VIAF id (repeated VIAF sources)	Items with repeated VIAF sources (%)	No. of pers. it. with VIAF id (repeated VIAF source ids)	Items with repeated VIAF source ids (%)
0	2,037,638	98.86	2,036,034	98.79	2,036,034	98.79
1	21,770	1.06	21,564	1.05	20,249	0.98
2	1,147	0.06	2,492	0.12	3,163	0.15
3	276	0.01	632	0.03	902	0.04
4	94	0.00	198	0.01	373	0.02
5	49	0.00	72	0.00	154	0.01
6	26	0.00	18	0.00	55	0.00
7	14	0.00	16	0.00	35	0.00
8	18	0.00	5	0.00	29	0.00
others	14	0.00	15	0.00	52	0.00
TOT	**2,061,046**	**100.00**	**2,061,046**	**100.00**	**2,061,046**	**100.00**

DISCUSSION

VIAF and Wikidata are quite different in their purpose, scope, organizational and theoretical approach, data harvesting and management.

A major difference between VIAF and Wikidata is in their purpose: on the one hand, VIAF aims to identify bibliographic entities and to connect authority data provided by selected Contributors (national libraries, cultural agencies, and other major institutions) and extracted from Other Data Providers (such as ISNI, RISM or DE663, Wikidata, etc.) through the creation of clusters by means of software. On the other hand, like ISNI, Wikidata focuses on both identification and description of entities and has the purpose of building collaboratively a database concerning the sum of all relevant knowledge—provided that each item complying with its notability criteria is accepted—using a crowdsourced approach (https://www.wikidata.org/wiki/Wikidata:Notability).

Another relevant difference between VIAF and Wikidata is their scope: while VIAF aims to identify a few selected types of entities already described within the bibliographic universe by national agencies, Wikidata aims to identify and describe any kind of entity of interest for the Wikidata community. Wikidata items may exist for any kind of entity and may contain a very broad range of data and of external identifiers. So, Wikidata can represent bibliographic data and entities—e.g., at present Wikidata records data for the 54% of all the bibliographic sources cited in Wikipedia entries—any other kind of entity provided for in VIAF (i.e., agents, works, expressions, and places), and any other entity defined by the FRBR-IFLA LRM model (e.g., manifestations, items, timespans, *nomens*, *res*, etc.), and by other models relevant for the GLAM universe (such as FRBRoo and CIDOC).[39] But it is open to any data model because it can also include any kind of entity *outside* the bibliographic or cultural heritage universe, as it is a knowledge base capable of containing any kind of statement on any entity users want to describe. In addition, for any kind of entity there is no minimum or maximum number of statements that must or can be added; as soon as an entity is clearly identified, it can be added to Wikidata. Moreover, when missing, new identifiers—and properties for description—can be proposed by anyone through property proposals and, if well defined, they are usually approved within two weeks (https://www.wikidata.org/wiki/Wikidata:Property_proposal). A broader scope is supposed to be much more convenient for users who wish to discover previously unknown links and information in the semantic web.

Organizational Model

Due to the VIAF top-down approach, data is completely managed by OCLC with no chance for common users or medium and small libraries or other institutions to directly improve VIAF clusters (e.g., by adding other data coming from their collections or from encyclopedias or online databases, merging duplicates, solving conflations, etc.). As the Wikidata approach is "to crowd-source data acquisition, allowing a global community to edit the data," data is curated directly by users interested in their creation and use.[40] So, in Wikidata, data is produced by volunteers, by means of semiautomatic or manual data harvesting from any desired and available source. Moreover, users' statistics show that authoritative data from national bibliographic agencies and other libraries, archives, and museums are normally uploaded by common users, not by librarians (or any other kind of institutional data curator).[41]

Identification Function

The theoretical approach differs too, both as to the form of the names and as to identification function. In VIAF, preferred and variant forms of names for persons are based on national cataloguing codes. Because national codes are different, VIAF is needed and works as a neutral hub of all the national preferred forms. Cataloguing rules can assure uniformity and univocity to the forms of the names of the entities within a national catalogue but are quite complicated to be understood and used by users. In Ranganathan's words, "the cataloguing conventions are on the surface quite contrary to what Mr. Everybody is familiar with."[42] In contrast, preferred forms in Wikidata are based on the international principles of the convenience of the user and common usage.[43] A clear example is the use of the direct form of name (Jane Doe) instead of the inverted form of name (Doe, Jane).

A different usage in the forms of names could be an issue for the integration of library metadata in Wikidata. In practice, however, it is not. First, there is no conflict between the Wikidata form and any other form from a theoretical point of view, as Wikidata form is already treated in VIAF as the preferred form within its specific context.[44] In addition to that, Wikidata accepts any library

identifier, so that any library-controlled form can be linked to a Wikidata item and vice versa. Furthermore, a Wikidata bot could be programmed to dump authorized and variant access points from national authority files and add them to the item labels and aliases.[45] Lastly, it could be argued that national cataloguing codes are compliant with the ICP principles and with the convenience of the user and common usage. But a remarkable difference is that while in national codes principles are applied *by cataloguers for users*, in Wikidata they are expressed directly *by the users* themselves.

As the identification function is a major feature of the semantic web, the different approach of VIAF and Wikidata to this issue must be underlined. As noted, "VIAF remains neutral towards differences in the cataloguing policy of its data contributors" and, for this reason, VIAF accepts all IDs provided by its sources, even when they are not clearly identifiable entities but are just labels (see for example https://viaf.org/viaf/307171748 or https://viaf.org/viaf/305052259).[46] On the contrary, Wikidata explicitly requires each item to refer to "a clearly identifiable conceptual or material entity" (second notability criterium; https://www.wikidata.org/wiki/Wikidata:Notability). As a consequence, many isolated clusters formed by VIAF on the basis of single Contributors' IDs related to not-clearly-identifiable entities are not acceptable in Wikidata and remain unlinked. Moreover, data on cluster duplication shows that identification in Wikidata is performed with the same quality level as in VIAF.

Clusters for identification purpose are created both in VIAF and Wikidata, but differently from VIAF, in Wikidata external identifiers—as all the other data—are not provided in a structured way by national libraries or other institutions (with very few exceptions); instead, identifiers are usually found and added by common users through web scrapers and after data cleaning. What is more, matches are not performed automatically, but semiautomatically (through tools such as OpenRefine or Mix'n'match (https://mix-n-match.toolforge.org/ and https://openrefine.org/) or manually. An enhanced feature of Wikidata in clusterization is the record of a wider variety of sources and relative IDs: due to its openness, Wikidata refers to VIAF and its sources, but also to any other library or cultural institution and to a large number of reference sources like encyclopedias and biographical dictionaries too (table 7). A wider variety of identification sources and manual work assure a higher level of identification.

Data Quantity

Data harvesting affects both quantity and quality of data. In VIAF, data are collected from periodical contributions of VIAF participants, with very large sets of data. Therefore, from a quantitative point of view, VIAF has a far larger number of people (22,099,715 personal clusters) in comparison with Wikidata (8,304,947 personal items).

Even though Wikidata was created in 2012, the number of personal items in Wikidata is currently only over a third (37%) of all VIAF personal clusters. Although quantities are not directly comparable due to the different universe to be described, in the last few years initiatives to enhance organized cooperation between libraries and Wikidata and to promote data production in Wikidata are increasing. A very high-quality initiative is supported by Cornell University, Harvard University, Stanford University, and the University of Iowa's School of Library and Information Science, in collaboration with the Library of Congress and the Program for Cooperative Cataloging (PCC). Their Linked Data for Production (LD4P) Wikidata project is "an in-depth exploration of how Wikidata could serve as a platform for publishing, linking, and enriching library linked data"

(https://www.wikidata.org/wiki/Wikidata:WikiProject_Linked_Data_for_Production). An additional example is the IFLA Wikidata Working Group that was formed "to explore and advocate for the use of and contribution to Wikidata by library and information professionals, the integration of Wikidata and Wikibase with library systems, and alignment of the Wikidata ontology with library metadata formats such as BIBFRAME, RDA, and MARC" (https://www.ifla.org/node/92837).

Even so, Wikidata is still very far from having a structured workflow to ingest data from national or local libraries, museums, and archives. In fact, while the projects mentioned above are mainly dedicated to explaining to the public of librarians and institutions why Wikidata is important and how to contribute to it, there are still very few projects which are mainly dedicated to the concrete massive synchronisation of data between library and bibliographic data and Wikidata. In fact, they also require a relevant effort in the manual cleaning of discrepancies and oddities emerging from the synchronisation. Relevant exceptions are the National Library of Wales[47] and the Biblioteca europea di informazione e cultura, where significant work has been done to synchronise respective databases of authors (and of other types of entities) with Wikidata.[48]

Data Quality

Data quality also needs to be analyzed in detail. Even if data from national libraries are authoritative and of high quality, as a virtual file VIAF neither has nor produces its own data. Consequently, VIAF data does not always remain authoritative because errors can be both inherited and added, and clusters can be duplicated. The issue is well known by ISNI, that "whenever necessary [. . .] splits and merges data coming from VIAF, and even applies protection to data that has been fixed manually."[49] As shown in table 2 and table 8, VIAF clusters are subject to isolation and duplication when they are created and to many changes and updates when they are maintained. So, even if VIAF collects a huge amount of authoritative data and creates clusters of IDs, VIAF users can not always safely and continuously rely on them. Data flows just in one direction (from national libraries to VIAF), VIAF deletes and rebuilds clusters without giving priority to the stability of one cluster over another, and, after April 2020, VIAF no longer makes available to users a record of its changes.[50] On the contrary, Wikidata data is always under strict control of any user, as its structure is designed to trace any minimum change to its data. Every single addition or deletion is documented, not just to easily recover eventual vandalism, but also to support any decision with clear evidence. Any stakeholder can exactly know if, how, when, and why data changed, in any moment.

What is more, from a qualitative point of view, Wikidata seems to offer a better solution for the recording of authority data than VIAF. First, it can store a wider variety of data about a person in a more semantic way. Not only is it possible in Wikidata to express preferred and variant forms of the name, related names, works, co-authors, publication statistics, and other data about the person—like in VIAF—but all these data are all expressed in a semantic way. For example, whereas in VIAF "Bach, Anna Magdalena" is just a related name of Johann Sebastian Bach, in Wikidata she is recorded and qualified as the person who married the musician. Thanks to that different approach, Wikidata can represent and show Bach's full genealogic tree (https://magnus-toolserver.toolforge.org/ts2/geneawiki/?q=Q1339). As Adamich noted, "building graphs from bibliographic entities is really about making the data machine readable and understandable. It is about making the data web enabled. In terms of translation, linked data opens up a whole new world over our MARC entrapment."[51]

Quality is enhanced by matching methods too; whereas VIAF matches identities by an algorithm based on explicit identifiers or string matching (such as the forms of the name, dates, and bibliographic relationships),[52] Wikidata matches are usually decided by a human, the user, or (in the case of semiautomatic imports) at least checked *a posteriori* by a human after some time. The higher precision of manual over automatic matching is recognized also in VIAF Guidelines.[53] Furthermore, as seen above, notability requires that, when clear identification is impossible, no item must be created in Wikidata.

Data Maintenance and Usability

Data quality relies also on maintenance. Comparison between Wikidata items and VIAF clusters shows a very small but constant presence of errors to be fixed in both (around 0.01%), even if it is impossible to determine with certainty whether VIAF uses Wikidata error pages. Issues on fixing VIAF errors directly by VIAF Contributors were already noted: "While clustering anomalies can be handled by VIAF itself, reporting errors found in source data of VIAF partners raise problems related to the efficiency of the notification workflows. At this point, involvement of VIAF partners themselves in the process is needed."[54] On the other hand, in Wikidata anyone can edit items, add new data or delete mistakes, merge items, fix various issues, and so on, on the fly. Due to its openness, Wikidata may also suffer from vandalism, but it has its own solutions.[55] Along with this, data receive special attention to their accuracy and reliability because they are uploaded and maintained by users that are direct stakeholders. For this reason, in Wikidata, references to bibliographical or biographical sources and to Other Data Provider IDs such as any national and international identification system are suggested, promoted, and carefully examined. Moreover, there is a commitment to monitor the consistency of VIAF clusters. The ability of Wikidata to identify inconsistent VIAF clusters and the fact that VIAF isolated clusters can be reduced at least by 30%[56] by referring to identifiers from Wikidata and Other Data Providers, are the best demonstration of the quality of its data and of the importance of the Other Data Providers in VIAF clusterization.

As to the usability of data, the internal search of VIAF lacks more than basic functions: the only available filter allows to limit results to clusters having one specific source; on the contrary, filtering searches for clusters having and/or not having a specific group of sources or to clusters having more or less sources would be very useful, especially in order to find duplicates. In contrast, Wikidata has a SPARQL query service which returns results based on the current status of the database and its internal search can integrate some of the functions of the query service, allowing to look for items having and/or not having specific statements (https://www.wikidata.org/wiki/Special:Search).[57] Considering cases in which VIAF and Wikidata discover potential duplicates in their sources, VIAF has no page dedicated to listing cases of (supposedly) duplicate IDs from its sources, while Wikidata easily allows to find cases in which single sources have (supposedly) duplicate IDs through constraint violations[58] and appropriate SPARQL queries.

A Comparison Table

A comparison table was built to compare scope, role, system, and functions between VIAF and Wikidata, inspired by and adapted from a VIAF vs ISNI comparison.[59]

Table 9. Comparison between and complementarity of VIAF and Wikidata features

Feature	VIAF	Wikidata
Scope	• Persons • Organizations • Works • Expressions • Locations	• Any kind of VIAF entity • Any "res" of IFLA LRM • Any entity of CIDOC • Any other non-GLAM entity • Any entity in the universe of discourse
Software	• Unknown	• Wikibase[60]
Data. Person entity properties	• Preferred form of name, based on national cataloguing rules • Very rich variant forms of name, identified by national agencies variant forms • Sources	• Preferred form of name (label) based on convenience of the user and common usage[61] • Variant forms of name (aliases), organized by languages and scripts[62] • Sources (as statements and references and with qualifiers)
Data. Quantity (persons)	• Number of clusters: 33,656,281 (Sept. 2020) • Number of personal clusters: 22,099,715 (Sept. 2020)	• Number of entities: 90,260,081 (Oct. 2020) • Number of personal items: 8,304,947 (Oct. 2020) • Number of personal items with VIAF ID: 2,061,046 (Sept. 2020)
Data. Harvesting	• Data are provided by authoritative national bibliographic agencies	• Data are added through massive semiautomatic imports and/or manually by any interested user
Data. Quality	• Data are granted by authoritative national bibliographic agencies	• Data are controlled by any directly interested user, based on data from VIAF, available bibliographic agencies, and other authoritative bibliographic sources
Data. Other entities properties	• ISBN, titles, dates included in the cluster	• Any kind of property applicable to an entity can be used (multimedia included)[63]

Feature	VIAF	Wikidata
	• Dates, genre, bibliographic references from sources, xlinks, etc. • Properties are unchangeable	• All statements admit references, which are strongly recommended in some cases • Unavailable properties can be freely added through a process of property proposal[64]
Data. Dates	• Dates are extracted from authority and bibliographic records using a parsing technique; calendars and precision are not available[65]	• Dates are imported semiautomatically from various sources or filled in manually; different calendars are available and further statements can be made through qualifiers[66]
Data. Vandalism	• No vandalism: data are editable only by OCLC	• Everyone can edit, but items which are frequently vandalized can be temporarily or permanently protected from the edits of unregistered users[67]
Data. Fixing errors, deduplicating, or unmerging clusters/items	• Suggestions and requests via email • Asynchronous • Presumably, automated processes and human interventions • VIAF rebuilds clusters and does not give priority to the stability of one cluster over another[68]	• Everyone can edit[69] • Instantaneous • Probable errors (constraint-violations) are detected in an automated way (by bots and through queries) • Pages with lists of probable errors (constraint-violations) are freely available and constantly updated in an automated way (by bots)[70]
Data. License	• All public data (license: http://opendatacommons.org/licenses/by/1.0/)	• All public data (license: https://creativecommons.org/publicdomain/zero/1.0/deed.it)
Role	• Create clusters • Ingest authority records from VIAF Contributors and Other Data Providers (included WKD and ISNI) • Publish and diffuse VIAF IDs and data	• Create items with a worldwide recognized and standard identifier • Interlink items with any available external identifier • Ingest data from VIAF, from VIAF Contributors, and Other Data Providers (e.g., ISNI)

Feature	VIAF	Wikidata
		● Allow to create and maintain on Toolforge free tools—e.g., Mix'n'match—to ingest external identifiers[71] ● Manage library, bibliographic, and non-library and non-bibliographic linked data ● Publish and diffuse Wikidata IDs and data
Organizational model	● OCLC service, guided by VIAF Council of participating institutions ● Hierarchical, top-down ● Membership on request and subordinated to approval ● Largely limited to national bibliographic agencies	● Wikimedia project ● Distributed, bottom-up ● Everyone can take part in the project[72] ● Open to any bibliographic or non-bibliographic institution (national, large, medium, and small)
System. Website	● Interface only in English language	● Interface in nearly any language and script; new ones can be added ● Online facilities (end user input; edit online facilities for end user) ● Login enhances users' experience (by gadgets and scripts)
System. Updating	● Periodical (asynchronous) ingestions	● Continuous, instantaneous, free updates
System. Versioning	● History is included in each present cluster and for abandoned clusters ● History is inaccessible in redirected clusters	● Page history available in each item and for redirected items ● For deleted items, history is accessible only to administrators
Long-term preservation policy	● OCLC maintains the hosting, software, and data for VIAF[73]	● Wikimedia Foundation maintains the hosting, software, and data for Wikidata[74]

Feature	VIAF	Wikidata
Notifications to stakeholders	• Notifications to be sent to data providers	• Notifications are sent to end users and contributors
Display, search, and download	• In multiple formats: xml and json, including justlinks.json; • Basic search interface • Clusters are listed without clear ranking rule • Integrating monthly dumps • API endpoint[75] • Before April 2020, by monthly dump with persist links; after, monthly dumps without persists links	• In multiple formats: json, php, n3, ttl, nt, rdf, jsonld, html[76] • Search interface[77] • API endpoint[78] • SPARQL query endpoint[79] • Dumps[80], also customizable[81] • See https://www.wikidata.org/wiki/Help:About_data
Linked data and SRU	• Linked data • SRU[82] (search and browse indexes, using CQL syntax; output formats are XML or HTML)	• Linked data
Interoperability. Local	• Local institution can only reconcile VIAF IDs to their own data • As changes are made by VIAF, synchronization must be periodically performed by sources and local institutions	• Full reconciliation, upload, and synchronization of local IDs on Wikidata and vice versa • Dedicated tools: Mix'n'match • Other tools: OpenRefine • Bots • Manually

CONCLUSION

Main VIAF and Wikidata features and personal entities data were analyzed and compared in this study to focus on analogies and differences, and to highlight their reciprocal role and helpfulness in the worldwide bibliographical context and in the semantic web environment.

VIAF is a major international initiative to address the challenge of reliably identifying bibliographic agents on the web, by means of authoritative data based on national cataloguing codes and coming from the national libraries involved in the UBC program. Moreover, VIAF is a pillar of the identification process that users enact within Wikidata. Still, the comparison emphasized a few relevant issues in VIAF's approach, designed more than twenty years ago: a very selective policy of inclusion of its sources—Contributors and Other Data Providers—and to their participation to the governance, that prevents a worldwide openness of the project to non-national libraries and cultural institutions; an obvious neutrality toward data coming from its

Contributors, even when data are not compliant with the identification requirements of the semantic web; troubles in correct clustering of IDs (duplicate clusters to be merged and conflated clusters to be split), and a one-way flow of data due to its top-down approach that prevents a quick and cooperative workflow to identify and fix errors; the ability to identify only a narrow range of entities (i.e., mainly bibliographic entities, but not even all those provided by IFLA LRM).

On the other side, the semantic web has offered new important tools and chances to libraries, archives, museums and other cultural institutions, and their data are recognized as a relevant asset for building the backbone of the semantic web as to the control of entities of bibliographic and cultural interest. After eight years of existence, Wikidata is playing a relevant role in the publication, aggregation, and control of bibliographic and non-bibliographic information in the semantic web too. It is more and more indicated as a hub for identifiers in the semantic web.[83]

Wikidata depends on VIAF for a large part of the identification work of its items on VIAF and VIAF's preeminent role in Wikidata is acknowledged by its primary position in the identifiers section of the data of each item. For this reason, the Wikidata community constantly monitors the consistency of VIAF clusters and continuously updates lists of errors present in them. On the other hand, if VIAF is undoubtedly very useful to the Wikidata community, Wikidata can support the consistency of VIAF clusters. The Wikidata informational ecosystem is much larger and wider, can be built by any interested institution and person, and its identification function can count also on the authority work of national and non-national libraries excluded from the VIAF environment, and on authoritative non-bibliographical reference sources too.

This study opens some research perspectives. Analysis was limited to data about personal entities, as this kind of entity was the only one directly comparable, while further research is wanted to possibly extend the analysis to other kinds of entities. Moreover, more research should be devoted to the investigation of the treatment of special categories of persons and their names, such as mythological and legendary characters, ancient Greek and Latin authors, kings, queens, popes, saints, and so on, as VIAF Guidelines[84] themselves declare among VIAF's typical problems the clusterization of such names (and they often get five or more VIAF IDs in Wikidata). A further line of research should consider the relevance of the clusterization of encyclopedias and other reference sources in the identification process within Wikidata. Lastly, isolated clusters would need more consideration; as a matter of fact, in this study they were used as a clue of relatively recent uploads in VIAF, but LC and DNB show a high rate of isolated clusters too (maybe due to the richness of their collections and metadata). More research on isolated clusters could help to describe with more precision the possible role of non-national libraries and institutions and of their locally rich collections in identifying lesser-known agents (not just persons) in a worldwide perspective.

From analyzed data and direct comparison, it can be concluded that VIAF and Wikidata can be constantly improved through reciprocal comparison, which allows discovery of errors in both. VIAF and Wikidata are two relevant tools for the authority control in the semantic web and they each have a specific role to play and different stakeholders. Unfortunately, as opposed to the relationship between VIAF and ISNI, at present no aspect of VIAF-Wikidata interoperability is discussed between the managing structures of both systems, on a regular or irregular basis.

While Wikidata appears to be more reliable with regards to the identification process, its most significant weakness consists in its unorganized and unplanned crowdsourced data acquisition,

even if based at present on about 11,500 active editors.[85] Furthermore, the Wikidata community still lacks the constant support and cooperation of institutional data curators such as librarians, archivists, and museum curators. Many current projects are mainly dedicated to explaining to the potential institutional stakeholders the importance and the usefulness of Wikidata for their institutional missions, but there are still too few projects devoted to massive synchronization of data from institutional silos to Wikidata. But, as soon as these initiatives reach a critical mass, Wikidata will become the real global hub of the web of data.

ACKNOWLEDGEMENTS

All the authors have cooperated in the redaction and revision of the article. Nevertheless, each author has mainly authored specific sections and subsections of the article:

- Stefano Bargioni: Data Analysis; VIAF; Wikidata; VIAF and Wikidata: A Data Comparison.
- Carlo Bianchini: Introduction; Discussion; Organizational Model; Identification Function; Data Quantity; Data Quality; Data Maintenance and Usability.
- Camillo Carlo Pellizzari di San Girolamo: Relationship between VIAF and Libraries; Relationship between Wikidata and Academic, Research, and Public Libraries; Relationship between VIAF and Wikidata; Wikidata Controls on VIAF; Materials and Methods; Conclusion.

All authors contributed to A Comparison Table. The authors wish to thank the anonymous reviewer whose suggestions helped to improve and enrich the paper, and the editor for his helpful edits.

ENDNOTES

[1] Thomas Baker et al., *Library Linked Data Incubator Group Final Report,* sec. 2 (W3C Incubator Group, October 25, 2011), http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

[2] Baker et al., *Library Linked Data*.

[3] Dorothy Anderson, *Universal Bibliographic Control. A Long Term Policy—A Plan for Action* (Munchen: Verlag Dokumentation, 1974), 11.

[4] Anila Angjeli, Andrew Mac Ewan, and Vincent Boulet, "ISNI and VIAF: Transforming Ways of Trustfully Consolidating Identities," in *IFLA WLIC 2014* (IFLA 2014 Lyon, IFLA, 2014), 2, http://library.ifla.org/985/1/086-angjeli-en.pdf.

[5] Rick Bennett et al., "VIAF (Virtual International Authority File): Linking the Deutsche Nationalbibliothek and Library of Congress Name Authority Files," *International Cataloguing and Bibliographic Control* 36, no. 1 (2007): 12–18; Barbara B. Tillett, *The Bibliographic Universe and the New IFLA Cataloging Principles : Lectio Magistralis in Library Science = L'universo bibliografico e i nuovi principi di catalogazione dell'IFLA : Lectio Magistralis di biblioteconomia* (Fiesole (Firenze): Casalini libri, 2008), 14–15, http://digital.casalini.it/9788885297814; "VIAF. Connect Authority Data across Cultures and Languages to Facilitate Research," OCLC, 2020, https://www.oclc.org/en/viaf.html.

[6] Gildas Illien and Françoise Bourdon, "A la recherche du temps perdu, retour vers le futur: CBU 2.0" (paper, IFLA WLIC 2014, Lyon, France, 2014), 13–14, http://library.ifla.org/956/.

[7] Illien and Bourdon, "A la recherche," 15.

[8] Gordon Dunsire and Mirna Willer, "The Local in the Global: Universal Bibliographic Control from the Bottom Up" (paper, IFLA WLIC 2014, Lyon, France, 2014), 11, http://library.ifla.org/817/.

[9] Luca Martinelli, "Wikidata: La Soluzione Wikimediana Ai Linked Open Data," *AIB Studi* 56, no. 1 (March 2016): 75–85, https://doi.org/10.2426/aibstudi-11434; Jesús Tramullas, "Objetos culturales y metadatos: hacia la liberación de datos en Wikidata," *Anuario ThinkEPI* 11 (2017): 319–21, https://doi.org/10/ghbj63; Xavier Agenjo-Bullón and Francisca Hernández-Carrascal, "Wikipedia, Wikidata y Mix'n'match," *Anuario ThinkEPI* 14 (2020), https://doi.org/10/ghbj6t; Claudio Forziati and Valeria Lo Castro, "The Connection between Library Data and Community Participation: The Project SHARE Catalogue-Wikidata," *JLIS.it* 9, no. 3 (2018): 109–20, https://doi.org/10/ggxj9n; Adrian Pohl, "Was Ist Wikidata Und Wie Kann Es Die Bibliothekarische Arbeit Unterstützen?," *ABI Technik* 38, no. 2 (2018): 208, https://doi.org/10/ghbj6w; *ARL White Paper on Wikidata: Opportunities and Recommendations* (The Association of Research Libraries, 2019), https://www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf; Regine Heberlein, "On the Flipside: Wikidata for Cultural Heritage Metadata through the Example of Numismatic Description" (paper, IFLA WLIC 2019, Libraries: Dialogue for Change, session 206: Art Libraries with Subject Analysis and Access, Athens, Greece, August 28, 2019), http://library.ifla.org/2492/1/206-heberlein-en.pdf.

[10] *ARL White Paper on Wikidata*, 27–30; Theo van Veen, "Wikidata: From 'an' Identifier to 'the' Identifier," *Information Technology and Libraries* 38, no. 2 (2019): 72–81,

https://doi.org/10/ghbj62; Hilary Thorsen, "LD4P: Linked Data for Production: Wikidata as a Hub for Identifiers" (slideshow presentation, June 11, 2020), https://docs.google.com/presentation/d/1jWz3_nCf5rdd-7ejETGlfv99UV2PnD1v/edit?usp=embed_facebook.

[11] Tillett, *The Bibliographic Universe*, 15.

[12] Open Data Commons Attribution License (ODC-By) v1.0 (as stated in http://viaf.org/viaf/data/).

[13] "VIAF Admission Criteria," OCLC, 2020, https://www.oclc.org/content/dam/oclc/viaf/VIAF%20Admission%20Criteria.pdf.

[14] The description of Wikidata source in http://viaf.org/viaf/partnerpages/WKP.html seems to refer to Wikipedia before the existence of Wikidata. The same acronym WKP reflects this anachronism, whereas ISNI correctly uses WKD. Anyway, this description, as well as many others, requires an update.

[15] Stacy Allison-Cassin and Dan Scott, "Wikidata: A Platform for Your Library's Linked Open Data," *Code4Lib Journal* 40 (May 4, 2018), https://journal.code4lib.org/articles/13424.

[16] Carlo Bianchini and Pasquale Spinelli, "Wikidata at Fondazione Levi (Venice, Italy): A Case Study for the Publication of Data about Fondo Gambara, a Collection of 202 Musicians' Portraits," *JLIS.it* 11, no. 3 (September 15, 2020): 24.

[17] IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), *Functional Requirements for Authority Data: A Conceptual Model* (München: K. G. Saur, 2009), 46, https://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf. For qualifiers, see https://www.wikidata.org/wiki/Help:Qualifiers; for references see https://www.wikidata.org/wiki/Help:Sources.

[18] Partial lists are linked from https://wikibase-registry.wmflabs.org/wiki/Main_Page.

[19] See https://www.transition-bibliographique.fr/fne/french-national-entities-file/; the Proof of Concept is available at https://github.com/abes-esr/poc-fne.

[20] Jean Godby et al., *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage* (Dublin OH: OCLC Research, 2019): 8, https://doi.org/10.25333/faq3-ax08.

[21] IFLA, "Opportunities for Academic and Research Libraries and Wikipedia" (discussion paper, 2016), 10, https://www.ifla.org/files/assets/hq/topics/info-society/iflawikipediaopportunitiesforacademicandresearchlibraries.pdf.

[22] John Riemer, "The Program for Cooperative Cataloging & a Wikidata Pilot" (slideshow presentation, June 16, 2020), slide 5, https://docs.google.com/presentation/d/1NpkAQdGGft1Wi2vX0zgMtIxwXWjPq96NtXx4MmyXFFI/edit#slide=id.p.

[23] Godby et al., "Creating Library Linked Data," 8.

[24] Maximilian Klein and Alex Kyrios, "VIAFbot and the Integration of Library Data on Wikipedia," *Code4Lib Journal* 22 (October 14, 2013), https://journal.code4lib.org/articles/8964.

[25] IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code, *Statement of International Cataloguing Principles (ICP)* (Den Haag: IFLA, 2016), para. 5.3.

[26] https://www.wikidata.org/wiki/MediaWiki:Wikibase-SortedProperties#IDs_with_datatype_%22external-id%22; ISNI (P213, https://www.wikidata.org/wiki/Property:P213) is presently sorted after VIAF instead of in the ISO section because it is considered primarily as a VIAF source.

[27] Epìdosis, *Viaf e Wikidata.mpg*, 2020, https://commons.wikimedia.org/wiki/File:VIAF_e_Wikidata.mpg; a list of gadgets is available at https://www.wikidata.org/wiki/Wikidata:VIAF/cluster#Gadgets.

[28] The main error-report page is https://www.wikidata.org/wiki/Wikidata:VIAF/cluster/conflating_entities; its subpage https://www.wikidata.org/wiki/Wikidata:VIAF/cluster/conflating_specific_entries is designed for collecting "easy" cases of conflation, when only a few members of a cluster should be moved elsewhere, while the cluster is substantially sane.

[29] Moreno Hayley, email to author, March 23, 2020. To the question if data about abandoned clusters would have been maintained, the VIAF answered, "We recognize that the data in the file was not usable. VIAF is in a period of transition and it was decided that we could not at this time fix the file so it has been removed from the list of available downloads."

[30] The statement read: "The persist-rdf.xml file has been removed and will no longer be available," accessed October 23, 2020.

[31] Angjeli, Mac Ewan, and Boulet "ISNI and VIAF," 3.

[32] https://dumps.wikimedia.org/wikidatawiki/; instructions and a list of kinds of data dumps are available at https://www.wikidata.org/wiki/Wikidata:Database_download.

[33] A general explanation of ranks is available at https://www.wikidata.org/wiki/Help:Ranking. Here is a small summary: values of statements can be ranked in three ways, "preferred," "normal" (default), and "deprecated"; the expression "values with non-deprecated rank" includes all values with preferred rank or normal rank; the expression "values with best rank" includes only values with preferred rank or normal rank, with this condition: if the same statement has two or more values and at least one of them has preferred rank, values with normal rank aren't counted; if there aren't values with preferred rank, all values with normal rank are counted.

[34] VIAF and Wikidata dumps, together with the scripts, were published on Zenodo at https://doi.org/10.5281/zenodo.4457114.

[35] The queries can be performed using the following links: VIAF members: https://w.wiki/i5J; authority controls related to libraries but not being VIAF members: https://w.wiki/i5K; biographical dictionaries: https://w.wiki/i5N.

[36] The query can be performed using the following link: https://w.wiki/i5p.

[37] It could be because they are probably more difficult to cluster, but in some cases also because they represent infrequently described entities.

[38] As suggested by the reviewer, more removals than additions may be a clue of a cleanup project.

[39] Pat Riva, Patrick Le Boeuf, and Maja Zumer, *IFLA Library Reference Model*, draft (Den Haag: IFLA, 2017), https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla_lrm_2017-03.pdf; Nick Crofts et al., "Definition of the CIDOC Conceptual Reference Model," version 5.0.4, ICOM/CIDOC CRM Special Interest Group, 2011, http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html; Chryssoula Bekiari et al., eds., *FRBR Object-Oriented Definition and Mapping from FRBRER, FRAD and FRSAD,* version 2.0 (International Working Group on FRBR and CIDOC CRM Harmonisation, 2013), http://old.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V2.0_draft_2013May.pdf; Lydia Pintscher, Lea Lacroix, and Mattia Capozzi, "What's New on the Wikidata Features This Year," YouTube video, October 26, 2020, truocolo, https://www.youtube.com/watch?v=EbXdZK54GrU.

[40] Denny Vrandečić and Markus Krötzsch, "Wikidata: A Free Collaborative Knowledgebase," *Communications of the ACM* 57, no. 10 (September 23, 2014): 80, https://doi.org/10/gftnsk.

[41] For a general statistic see http://wikidata.wikiscan.org/users; for a statistic about the VIAF property see https://bambots.brucemyers.com/NavelGazer.php?property=P214; changing the id of the property at the end of the URL allows exploring other property statistics.

[42] Shiyali Ramamrita Ranganathan, *Reference Service*, 2nd ed., Ranganathan Series in Library Science 8 (Bombay: Asia Publishing House, 1961), 74.

[43] IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code, *Statement of International Cataloguing Principles (ICP)*, 5, https://www.ifla.org/publications/node/11015.

[44] Wikidata does have a guideline for a preferred label, and its choice is based on users' convenience (https://www.wikidata.org/wiki/Help:Label, par. 1.2) as required by International Cataloguing Principles (2016). As to the choice of the Wikidata label in a specific language, VIAF does not show any clear principle, while the authors believe that it would be preferable to use the English ("en") label, whenever available. See IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code, *Statement of International Cataloguing Principles (ICP)*.

[45] For example, in September it was done for NKC using OpenRefine (sample edit: https://www.wikidata.org/w/index.php?title=Q520487&diff=1269046867&oldid=1266870464).

[46] Angjeli, Mac Ewan, and Boulet, "ISNI and VIAF," 9.

[47] Simon Cobb (https://www.wikidata.org/wiki/User:Sic19) became Wikidata Visiting Scholar in 2017 (https://en.wikipedia.org/wiki/User:Jason.nlw/Wikidata_Visiting_Scholar).

[48] Federico Leva and Marco Chemello, "The Effectiveness of a Wikimedian in Permanent Residence: The BEIC Case Study," *JLIS.It* 9, no. 3 (September 2018): 141–47, https://doi.org/10.4403/jlis.it-12481.

[49] Angjeli, Mac Ewan, and Boulet, "ISNI and VIAF," 11.

[50] Andrew Mac Ewan, "ISNI, VIAF and NACO and Their Relationship to ORCID, discussion paper for PCC Policy Committee, 4 November," 2013, 2, http://www.loc.gov/aba/pcc/documents/ISNI%20PoCo%20discussion%20paper%202013.docx.

[51] Tom Adamich, "Library Cataloging Workflows and Library Linked Data: The Paradigm Shift," *Technicalities* 39, no. 3 (May/June 2019): 14.

[52] OCLC, *VIAF Guidelines*, rev. July 16, 2019, 2, https://www.oclc.org/content/dam/oclc/viaf/VIAF%20Guidelines.pdf.

[53] OCLC, *VIAF Guidelines*, 5. "When VIAF is unable to algorithmically match some of the source authority records with each other, they can be manually pulled together into a single cluster using an internal table."

[54] Angjeli, Mac Ewan, and Boulet, "ISNI and VIAF," 16.

[55] Stefan Heindorf et al., "Vandalism Detection in Wikidata," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM '16 (New York, NY: Association for Computing Machinery, 2016), 327–36, https://doi.org/10/gg2nmm; Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli, "Building Automated Vandalism Detection Tools for Wikidata," in *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion (Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017), 1647–54, https://doi.org/10/ghhtzf.

[56] See table 1, col. 1 vs col. 9; it should be noted that col. 9 considers only non-VIAF sources and biographical dictionaries, but Wikidata also links to encyclopedias and other online databases.

[57] For example, people not having VIAF id but having ICCU id (https://tinyurl.com/y6hbtjuo); instructions about the internal search are available at https://www.mediawiki.org/wiki/Help:Extension:WikibaseCirrusSearch.

[58] https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations.

[59] Angjeli, Mac Ewan, and Boulet, "ISNI and VIAF," 16.

[60] https://www.mediawiki.org/wiki/Wikibase/DataModel.

[61] "The label is the most common name that the item would be known by" (https://www.wikidata.org/wiki/Help:Label). See also IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code, *Statement of International Cataloguing Principles (ICP)*, 5., https://www.ifla.org/publications/node/11015.

[62] Bots exist to create more and more variant forms based on matching properties, such as date of birth (P569) and date of death (P570), and to import variant forms of names from national authority files. See, for example, https://www.wikidata.org/w/index.php?title=Q5669&diff=611600491&oldid=608231160.

[63] https://www.wikidata.org/wiki/Help:Data_type.

[64] https://www.wikidata.org/wiki/Wikidata:Property_proposal.

[65] Jenny A. Toves and Thomas B. Hickey, "Parsing and Matching Dates in VIAF," *Code4Lib Journal,* 26 (October 21, 2014), https://journal.code4lib.org/articles/9607; Stefano Bargioni, "From Authority Enrichment to AuthorityBox : Applying RDA in a Koha Environment," *JLIS.It* 11, no. 1 (2020): 175–89, https://doi.org/10/gg66rq.

[66] https://www.wikidata.org/wiki/Help:Dates.

[67] See Heindorf et al., "Vandalism Detection in Wikidata."

[68] See Mac Ewan, "ISNI, VIAF and NACO."

[69] See https://www.wikidata.org/wiki/Help:Merge, https://www.wikidata.org/wiki/Help:Split_an_item, and https://www.wikidata.org/wiki/Help:Conflation_of_two_people.

[70] Complete list at https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations (e.g., https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations/P214).

[71] https://admin.toolforge.org/; see also Xavier Agenjo-Bullón and Francisca Hernández-Carrascal, "Registros de autoridades, enriquecimiento semántico y Wikidata," *Anuario ThinkEPI* 12 (2018): 361–72, https://doi.org/10/ghbj6z.

[72] https://www.wikidata.org/wiki/Wikidata:Property_proposal.

[73] https://www.oclc.org/en/viaf.html.

[74] https://www.wikidata.org/wiki/Wikidata:Introduction.

[75] https://platform.worldcat.org/api-explorer/apis/VIAF.

[76] https://www.wikidata.org/wiki/Special:EntityData; see also https://www.wikidata.org/wiki/Wikidata:Database_download.

[77] https://www.wikidata.org/wiki/Special:Search.

[78] https://www.wikidata.org/w/api.php.

[79] https://query.wikidata.org/.

[80] https://dumps.wikimedia.org/wikidatawiki/.

[81] https://wdumps.toolforge.org/.

[82] https://www.oclc.org/developer/develop/web-services/viaf/authority-source.en.html.

[83] van Veen, "Wikidata."

[84] See "Typical problems" in VIAF Guidelines: https://www.oclc.org/content/dam/oclc/viaf/VIAF%20Guidelines.pdf.

[85] Pintscher, Lacroix, and Capozzi, "What's New."