JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12749

VIAF and the linked data ecosystem

Nathan Putnam[a]

a) OCLC, http://orcid.org/0000-0002-3984-3035

Contact: Nathan Putnam, putnamn@oclc.org

Received: 30 April 2021; **Accepted:** 14 June 2021; **First Published:** 15 January 2022

ABSTRACT

This article reviews the founding, current state, and potential future of VIAF®, the Virtual International Authority File. VIAF consists of an aggregation of bibliographic and authority data from over 50 national agencies and infrastructures, systems that follow different cataloging practices and contain hundreds of languages. After a short history of the project, the results of surveys for implementers of linked data projects on the use of VIAF data and provides suggestions for future use and sustainability.

KEYWORDS

RDF; Library Linked Data; VIAF; History.

The Virtual International Authority File, known as VIAF, provides cultural heritage institutions and users with access to a combined, single international authority file with data from national libraries and infrastructures worldwide. VIAF contributors supply authority data that is matched, linked, and clustered with existing VIAF entities. VIAF allows researchers to identify names, locations, works, and expressions while preserving the regional language, spelling, and script preferences. There are more than 50 VIAF contributors from over 30 countries. The VIAF Council governs VIAF, which includes representatives from the contributor organizations and provides guidance on the policies, practices, and operations of VIAF.

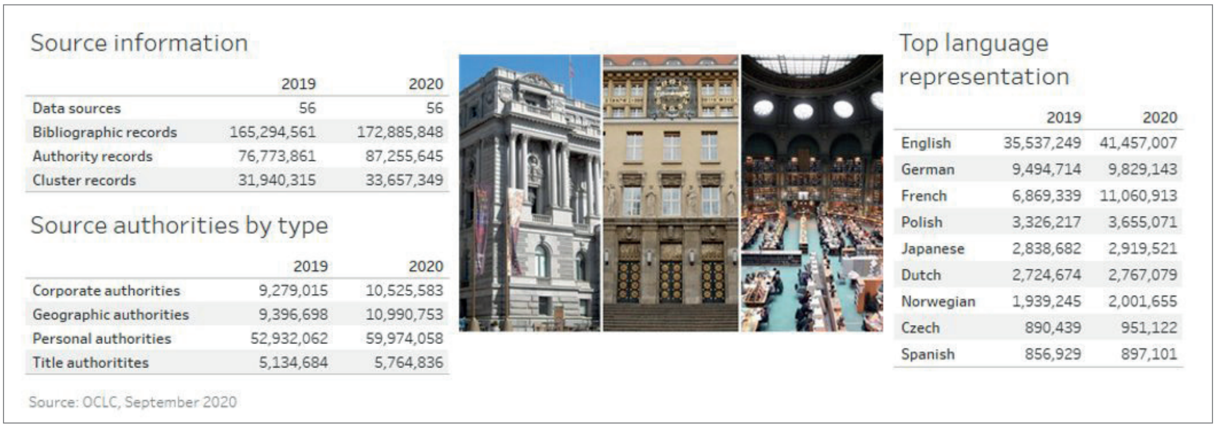
This article begins with a short history of VIAF, including some current statistical information. It then discusses VIAF within the larger linked data ecosystem through several surveys conducted by OCLC. It concludes with a discussion regarding OCLC's continued support for VIAF, its use and potential integration into OCLC's shared entity management infrastructure, and recommendations for further research and investigation.

VIAF history and current use

History

In April 1998, the United States Library of Congress (LC), the German National Library (Deutsche Nationalbibliothek, or DNB), and OCLC wanted to test linking to each other's authority records for personal names as a proof-of-concept project. In August 2003, the LC, the DNB, and OCLC formed the VIAF Consortium in a written agreement during the International Federation of Library Associations and Institutions (IFLA) conference in Berlin, Germany. In October 2007, the National Library of France (Bibliothèque nationale de France, or BnF) joined the consortium. After this, the four organizations, assuming the role of Principals, had joint responsibility for VIAF with the three libraries contributing authority and bibliographic content, while OCLC supported the software and infrastructure. Other organizations later joined the consortium as Contributors, providing source files and expertise to advance the state of VIAF. Due to the proof-of-concept success, the Principles and Contributors sought a suitable long-term organizational arrangement for VIAF. After considering several options, the Principals and Contributors agreed to transition VIAF to OCLC, which was completed in April 2012 (Murphy, 2012).

As of September 2020, VIAF receives data from 56 sources and includes 172 million bibliographic records, 87 million authority records, and 33 million cluster records. Records are clustered, i.e., grouped together, when they represent the same thing but with data from different sources. VIAF stores both the source record and the aggregated cluster record. Figure 1 provides detailed statistics on source information, authority records by type, and top language representation. The comparison year in the figure uses OCLC's fiscal year, which runs from July to June. Interestingly, the top three languages within VIAF are of the three Principal institutions, LC, DNB, and BnF. VIAF also contains a range of authority records from the sources, including 10.5 million corporate authorities, 10.9 million geographic authorities, approximately 60 million personal name authorities, and 5.7 million title authorities, totaling 87 million.



Source information

	2019	2020
Data sources	56	56
Bibliographic records	165,294,561	172,885,848
Authority records	76,773,861	87,255,645
Cluster records	31,940,315	33,657,349

Source authorities by type

	2019	2020
Corporate authorities	9,279,015	10,525,583
Geographic authorities	9,396,698	10,990,753
Personal authorities	52,932,062	59,974,058
Title authoritites	5,134,684	5,764,836

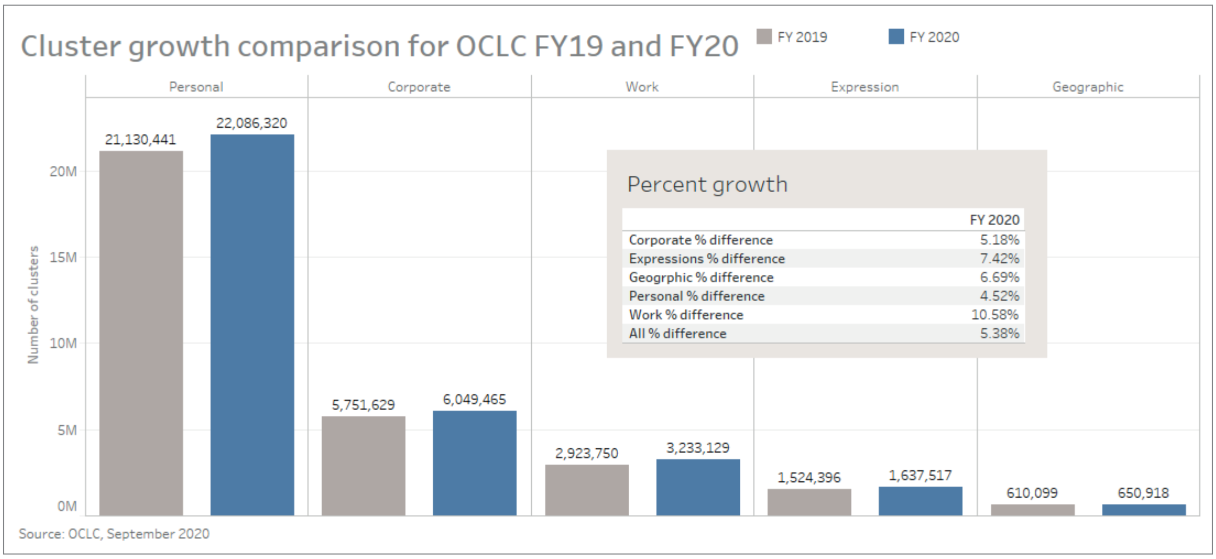
Top language representation

	2019	2020
English	35,537,249	41,457,007
German	9,494,714	9,829,143
French	6,869,339	11,060,913
Polish	3,326,217	3,655,071
Japanese	2,838,682	2,919,521
Dutch	2,724,674	2,767,079
Norwegian	1,939,245	2,001,655
Czech	890,439	951,122
Spanish	856,929	897,101

Source: OCLC, September 2020

Fig. 1. Source, type, and language representation

VIAF continually adds data from existing and new sources. As seen in Figure 2, data clusters continue to grow, and the cluster types for personal, corporate, work, expression, and geographic authority records. While there are considerably more personal name clusters within VIAF, OCLC believes that the other types will increase in importance as existing and new users consume the VIAF data.



Percent growth

	FY 2020
Corporate % difference	5.18%
Expressions % difference	7.42%
Geogrphic % difference	6.69%
Personal % difference	4.52%
Work % difference	10.58%
All % difference	5.38%

Source: OCLC, September 2020

Fig. 2. Cluster comparison between OCLC Fiscal Year 2019 and Fiscal Year 2020

International linked data survey for implementers

During the past seven years, OCLC Research conducted surveys on implementing various linked data tasks and uses. Building upon the interest of the OCLC Research Library Partners Metadata Managers Focus Group, OCLC Research conducted the first "International Linked Data Survey for Implementers" between 7 July and 15 August 2014. This followed updates to the original survey in 2015 and 2018. This article discusses the results regarding the use of VIAF data, but analyses and

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12749

results are available on the OCLC Research Linked data webpage (OCLC Research, n.d.). Interested persons can access the data directly on the OCLC Research linked data pages or through several articles written by Karen Smith-Yoshimura, including her discussion and analysis of the results[1].
Many institution types participated in the surveys including research libraries, national libraries, research institutions, library networks, governments, service providers[2], public libraries, museums, and a few classified as other. While research libraries continue to have many responders, Figure 3 shows the growing interest in different groups like national libraries, research institutions, and government institutions. Even on the lower end of the responder spectrum, public libraries and museums have seen a slight growth.

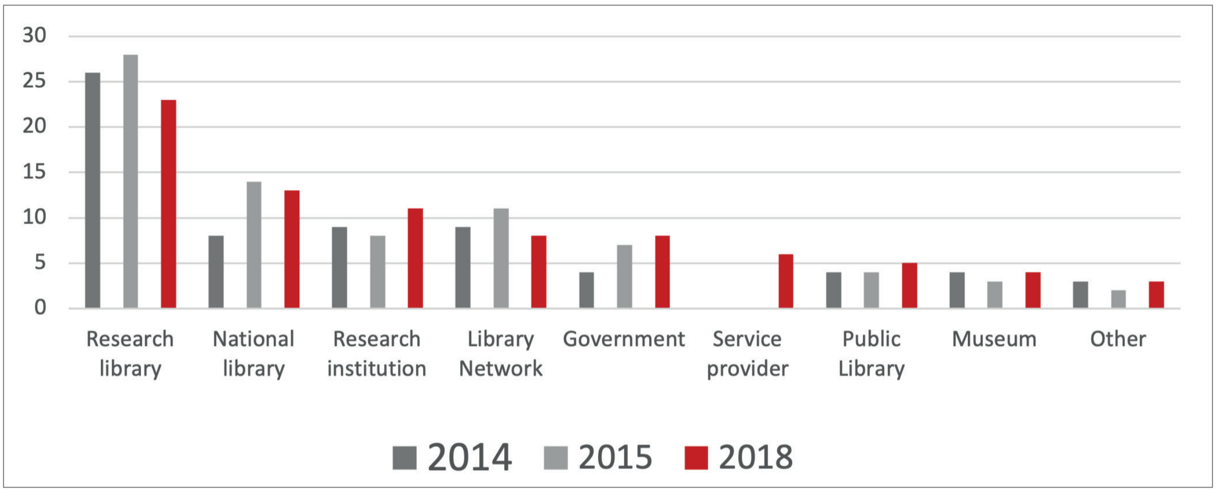


Fig. 3. Reponding insitutions by type

As the ecosystem matures and adoption increases, the participation of these groups will continue to grow. While there was a slight drop in the number of institutions answering how long they have had a linked data project or service in production, the number of projects and the time they have remained active continue to grow. 75% of the linked data projects/services described in 2018 are in production, slightly higher than the 67% reported in 2015. 40% of the linked data projects/services described in 2018 have been in production for more than four years.
The 2018 survey highlights the top seven linked data implementations. "Most used" is measured by the average number of requests per day, with all services reporting over 100,000 requests per day. All eight services have also been in production for more than four decades and include:

- American Numismatic Society's nomisma – a thesaurus of numismatic concepts[3]
- Bibliothèque nationale de France's data.bnf.fr – provides access to the BnF's collections and is a hub among different resources[4]

[1] See https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-survey.html for a complete listing of Karen's publications and presentations
[2] Service providers responded only to the 2018 survey
[3] http://nomisma.org/
[4] https://data.bnf.fr/

- Europeana – an aggregation of metadata for digital objects from museums, archives, and audiovisual archives across Europe[5]
- Library of Congress Linked Data Service – provides access to over 50 vocabularies[6]
- National Diet Library's NDL Search – provides access to bibliographic data from Japanese libraries, archives, museums, and academic research institutions[7]
- North Rhine-Westphalian Library Service Center (hbz) Linked Open Data service – provides access to bibliographic resources, libraries and related organizations, and authority data[8]
- OCLC's Virtual International Authority File (VIAF) – an aggregation of over 50 authority files from different countries and regions[9]

Figure 4 shows the top ten linked data sources consumed by the 2018 survey respondents compared to 2015. The count of respondents in the 2018 and 2015 surveys was the same, 69 and 68, respectively. Six of the ten sources dropped between the 2015 and 2018 surveys while the other four grew. The most considerable change was in the increased use of Wikidata. And even though VIAF dropped between the two surveys, it is still ranked relatively high, coming in second after the Library of Congress's ID service.

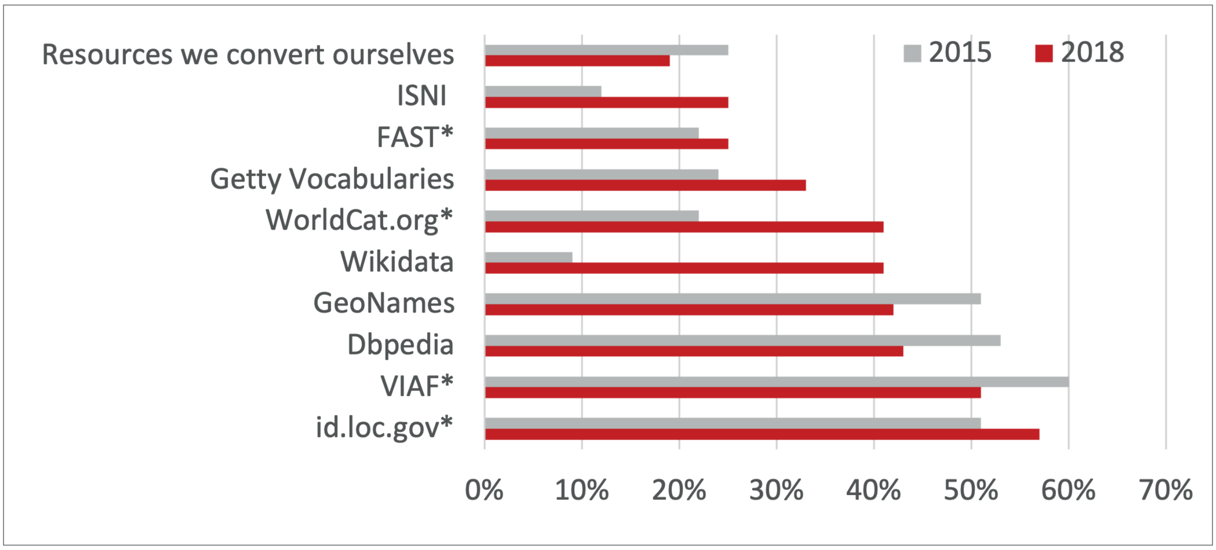


Fig. 4. A comparison of linked data sources consumed in 2015 and 2018

[5] https://www.europeana.eu/
[6] https://id.loc.gov/
[7] https://iss.ndl.go.jp/
[8] http://lobid.org/
[9] http://viaf.org/

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12749

The potential for VIAF

VIAF continues to be supported by OCLC and, as discussed earlier, continues to add new sources and data. The ongoing success of VIAF for various consumers, including OCLC, will depend on greater integration into the linked data environment. Success includes transforming VIAF from a primarily MARC-based system to native RDF and integrating with RDF services and support.

With financial support from the Andrew W. Mellon Foundation, OCLC is building a shared entity management infrastructure for library linked data. When completed in December 2021, this infrastructure will include authoritative descriptions of several types of entities, including works and persons, and will be enhanced and managed by the library and OCLC. Connections to other external vocabularies will place library collections in a broader context across the web.[10]

VIAF plays an integral role in the entity infrastructure, especially during the infrastructure's initial development phase. The grant-funded portion using VIAF entities to connect person entities within the infrastructure. During the first six months of data loading, selected VIAF clusters had connections to either WorldCat® works or Wikidata entities. The key to the initial phase was that the entities had built-in relationships with other entities that provided an enriched experience. The second six-month checkpoint continued the enrichment by adding additional entities. The second six-month checkpoint, which ended December 2020, included personal name entities from VIAF and work entities from WorldCat FRBR clusters. While not part of the grant requirements, it also had place entities from a separate linked data pilot for OCLC CONTENTdm®[11] with data from GeoNames, a database of geographical place names[12].

Areas for further investigation

The existing VIAF infrastructure continues to meet the goals of the original Principals and current Contributors. As with any library data project, continued usefulness will require change.

Two key areas to help determine the future of VIAF include running the implementers survey in the coming year and continued integration within the entity infrastructure. The implementer survey would indicate the continued use of VIAF within the larger linked data ecosystem and the probable and continued growth of Wikidata. Implementing VIAF into the infrastructure will help ensure stability and continuity as the ecosystem moves from record-based description to graph-based. Note that OCLC remains committed to providing a level of free access to those that wish to use the VIAF data regardless of in which ecosystem it finds itself. Additional areas for consideration include continued work with Wikidata partners to find solutions to challenges and the ever-on-going issues revolving around data quality and integrity.

[10] More information on the entity management infrastructure can be found at oc.lc/sharedentitymgmt.

[11] More information on the CONTENTdm Linked Data Pilot can be found at oc.lc/transform-linked-data.

[12] https://www.geonames.org/

References

Murphy, Bob. 2012. Virtual International Authority File service transitions to OCLC; contributing institutions continue to shape direction through VIAF Council. 4 April. Accessed January 24, 2021. https://worldcat.org/arcviewer/7/OCC/2015/03/19/H1426803137790/viewer/file1365.html.

OCLC Research. n.d. Linked data from OCLC Research. Accessed January 24, 2021. https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-survey.html.