Establishing Multi-script Name Headings in Finding Aids:

Library of Congress Versus the Virtual International Authority File

Justine Withers

San Jose State University

Abstract

The Virtual International Authority File (VIAF) aggregates authority records from libraries around the world. Does its use improve retrieval rates and differentiability when establishing name headings in an archive setting? This study compared retrieval rates from the Library of Congress and VIAF for personal names in a finding aid typical of the collections at the Stanford Hoover Institution Archives. Higher retrieval rates from the VIAF were anticipated, especially for the high concentration of Russian names in the collection. In reality, retrieval rates were similar. The VIAF interface offers advantages when searching for and differentiating headings.

Establishing Multi-script Name Headings in Finding Aids:

Library of Congress Versus the Virtual International Authority File

As the information world moves toward linked data and the Semantic Web, archives and their item-level records are at a disadvantage. The unique nature of their holdings limits the amount of pre-existing data that can be reused. In addition, the collection-level finding aid makes offering "actionable" chunks of information challenging.

Most archive items have a creator and a subject and authority records for these identities are a natural way to link resources (Mazzini & Rizzi, 2011). Library of Congress (LC) authority records are standard in the United States. LC name authorities have an authorized heading, differentiated from other names by middle initials, birth and death dates, or sometimes just guesswork. Alternate headings refer back to the main, authorized heading.

An alternative to Library of Congress headings is the Virtual International Authority File (VIAF), a service of OCLC from April, 2012, (OCLC, 2012a). The VIAF aggregates authority files from libraries across the world, including the Library of Congress, to create a "cluster record" for each identity. Its goals are to support linking memory institutions to the Semantic Web and allow users to search on and display their preferred form of a name. (OCLC, 2012b) Each VIAF identity cluster has a unique ID number and a stable URI.

When establishing name headings, does the VIAF offer a noticeable improvement over LC, especially in returning names from other countries and scripts, such as Cyrillic?

Statement of the Problem

The Hoover Institution Archives (HIA) at Stanford University are world-renown for its collection of "rare and unique material on political, economic, and social change in the modern

era," including collections on "imperial Russia, the USSR, and post-Soviet Russia" (Hoover Institution, 2012). Many of the original records in the Russia/Soviet/Commonwealth of Independent States (CIS) collection are of course in Russian. Names have been transliterated from Cyrillic to Roman script using the Library of Congress system. Some of the names in the original records were transliterated into Cyrillic—think surveillance reports on European revolutionaries by the Russian Imperial police—and have thus been roundtripped from Roman to Cyrillic and back. (J. Golden, personal communication, November 1, 2012).

Jill Golden, archivist at HIA, explained the archives processes in an interview (November 1, 2012). Finding aids are viewable online in PDF format and are searchable. Researchers must still visit the archive in person to view materials, most of which are on microfilm, to determine their usefulness and contents.

Finding aids at the Hoover Institute are in EAD (Encoded Archival Description) format published in XML using METS (Metadata Encoding and Transmission Standard). They must follow the guidelines of the Online Archive of California, where they are submitted to be shared with institutions from across the state. Searching on a name or subject will retrieve only the record for the collection and no further. As with most archives, digitizing the collection and linking to the items from the finding aid would be ideal. Currently, very little at HA is digitized. This may change depending on the priorities of new leadership. The current plan for improving accessibility is to make finding aids web-based with more access points based on LC headings. (J. Golden, personal communication, November 1, 2012).

The OAC guidelines for creating a finding aid in EAD require completing the *persname, famname, and corpname* fields from an standard authority name file, such as LC (California Digital Library, 2005).

The uniqueness of an archive's holdings exacerbates the problem of choosing authorized headings: uniform titles are of no use. Stevenson & Stevenson (2010) describe the challenge: "Archive data is by its nature incomplete and often sources are hidden and little known."

Archives also face several challenges in joining the linked data world from a technical point of view: current schema do not have adequate vocabulary to describe archives; archival datasets are not "actionable;" the archival community has not offered sufficient guidance on meeting their needs; and linked data may ultimately prove inappropriate for "highly hierarchical data models." (Coyle & Bermès, 2011)

The lack of item-level records in archives also makes linking to specific resources tricky.

Literature review

Hickey (2012a) reports the VIAF contains approximately 30 million authority records culled from 24 authority files and is updated "almost every month." Racine, as quoted in OCLC (2012a), hopes that the "mutual consolidation" of records in the VIAF will encourage "multilingualism." In December, 2011, the VIAF had accumulated 114,119 visits from 153 countries (OCLC, 2011).

Tillett deemed the "data cluster," such as that in the VIAF, "probably the most practical approach" (2007, p. 358) because it could record all variants, languages, and scripts without international administration.

In a presentation by five users of the VIAF, three participants reported using the VIAF as a reference to verify and establish headings, usually in MARC records. One participant reported using the VIAF as a backend database for a union catalog of Arabic manuscripts. The fifth user explained the algorithmic matching used by the VIAF to create identity clusters and the responsibility of users to report errors. (OCLC, 2011)

Agenjo, Hernandez, and Viedma (2012) offer the Polymath Virtual Library (PVL) as a case study for creating access points for every version of the names in their collection: "Spanish, Hispano-American, Brazilian, and Portuguese polymaths from all times" (p. 803). They rely heavily on the Virtual International Authority File to create MARC21 authority records that they consider "digital aggregates." Names are used both to identify and contextualize resources. They note the VIAF's weakness in not identifying the language of its headings. The MARC21/RDA records can be easily downloaded as EAC-CPF.

Coyle & Bermès (2011) describe the unique needs and challenges of archives in the Linked Data environment. They argue that, although archives "hold unique works, thus---are unable to engage in the kind of metadata sharing that is common in libraries," they "could benefit in particular from linking to other sources of information that cover" archived items. They summarize the goals of several Linked Data use cases in archives: create "semantic connections" among datasets in various formats; allow "serendipitous discovery" through a greater number of access points; "gain greater visibility" by being able to share data among institutions worldwide; and improve best practices by "exchanging expertise" and "improving interoperability."

Of the eight case studies discussed in Coyle & Bermès, only one is using EAD/METS and they are converting those records to linked data in a CIDOC Conceptual Reference Model (CRM)/FRBRoo ontology (Fundación Marcelino Botín & Universidad de Cantabría, 2010).

Wisser (2011) describes the ability of the Encoded Archival Content—Corporate Bodies, Persons, and Families (EAC-CPF) format to accommodate the multiple identities and records that an entity might have, including multiple languages and scripts. It acts as an adjunct record to an EAD, providing encoded "descriptions of entities associated with the creation, use, and maintenance of archival material" (Dryden, 2010).

Mazzini & Ricci (2011) describe converting their authority records from the EAC-CPF format to an OWL ontology in order to open up the Istituto per i beni artistici culturali e naturali (IBC) archive in Italy to the Semantic Web.

At Harvard University, the EAC-CPF best practices recommend VIAF as a secondary source for establishing headings (Yearl, 2012).

In a study of "first time use" headings at an academic library, Van Pulis (2006) found that only two-thirds of author names had matching authority records in OCLC. She argues that "variants in OCLC are a significant problem" and that an increase in international sources will exacerbate the problem (p. 564). Van Pulis' study was an influence on this study's methodology, leading to the expectation of even fewer matches from an archival collection.

The many languages and scripts contained in HIA records adds more complexity. Diekema (2012) considers the subject of multi-lingual digital libraries to be "understudied" (p. 175).

El-Sherbini and Chen (2011) surveyed librarians and end users of several non-Roman alphabets, who reported inconsistent Romanization and a preference for using their own language and script. One librarian specifically mentioned "good subject access to Slavic materials may be lacking because language expertise in cataloging is lacking" (p. 470).

Brewer (2009) describes the confusing standards for Romanizing Cyrillic languages that make research difficult. If words are *transcribed,* the sounds are re-created in the Latin alphabet, creating the possibility of many allowable interpretations. If they are *transliterated,* Cyrillic letters are converted to Latin letters following a standardized scheme. However, multiple schemes exist and the Library of Congress system used by most North American institutions is not used widely elsewhere in the world.

Kiebuzinski (2012) argues that access is more important than preservation and archives must be more pro-active in providing multi-language, multi-script access points.

Before the VIAF was viable, a team at the Queens Borough Public Library in New York built a Perl-based program to "de-transliterate" information in MARC records back to the original Cyrillic (Jacobs, Summers, & Ankersen, 2004). The writers point out that the conversion program relies on the strict transliteration rules to ensure accuracy, the opposite case from using VIAF.

Research Questions

The Hoover Institution Archives, like all archives, want to increase access to their holdings. Digitization of items is one crucial step. However, even if a collection is digitized, how will interested parties find it if the names in it do not match other authority records or typical search terms? Does access to the VIAF provide better results in determining established

headings? Does the VIAF provide a reliable source of the original Cyrillic version of Russian names? How can HIA best utilize VIAF and link up to the Semantic Web, and is it worth the effort?

Hypothesis 1: Name retrieval will be greater from the VIAF than from LC alone, because of the availability of Russian names from European libraries, including the Russian State Library.

Hypothesis 2: The availability of headings in Cyrillic would offer a sizable benefit from using the VIAF.

Methodology

Jill Golden recommended the Nikolaevsky Collection for a dataset. It is a manageable size and contains a good mix of notable names and lesser known figures.

From the finding aid for the Nikolaevsky Collection, I gathered 162 personal names, from both the series' headings and descriptions. I searched on each name in the Library of Congress (LC) Authority Headings (authorities.loc.gov) and Virtual International Authority File (VIAF) (viaf.org) in two phases:

1. "Active Search": Enter last name and first initial to retrieve as many results as possible. Ascertain correct heading using dates, publication titles, and information from outside sources such as Wikipedia. If unsuccessful, search under alternate names, including pseudonyms, maiden names, and sections of hyphenated names. This process most closely represents the usual process of determining an established heading.

2. "Exact Match": Enter the name as given in the finding aid to mimic automated retrieval from a spreadsheet.

In both cases, if the correct heading was found, the result was marked as positive. In addition, I recorded the availability of Cyrillic versions of names from the VIAF.

Results

When actively differentiating headings and searching for alternate versions, the LC and VIAF returned the same number of correct headings. Seven of them were listed in LC but not established. Exact match searches in the VIAF returned 8.6% more headings than LC.

Table 1

Number and Percentage of Headings Returned for Active Search and Exact Match

Source	Active Search	Exact Match
Library of Congress	107 (66%)	83 (51%)
VIAF	107 (66%)	97 (60%)

Note. n=162

H_1 proved correct only for exact match searching. H_2 was not proven.

Some surprising anomalies arose. The Nikolaevsky Collection finding aid lists documents from Lev Davydovich Trotskii, a famous name in Russian history, of course. In an exact name search, no results will be returned from LC or VIAF (Figures 1 and 2). The LC authorized heading is *Trotsky, Leon, 1879-1940* (Figure 4), following the convention of using the most familiar version of a name in English (as they also use *Leo Tolstoy* instead of the transliterated *Lev Tolstoĭ*). The VIAF offers Троцкий, Лев Давидович, 1879-1940, but the only transliteration available is *Trockij, Lev Davydovič* (Figure 3).

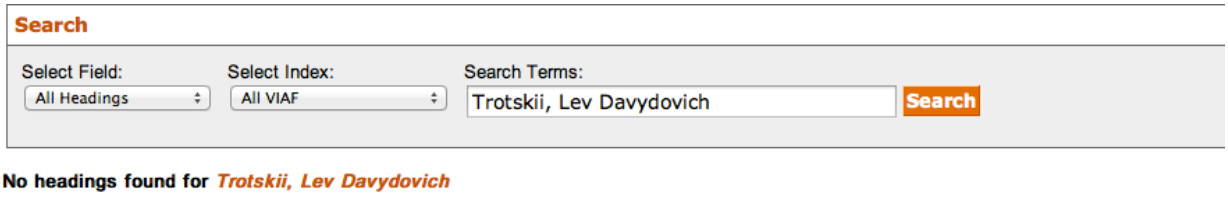

Figure 1. VIAF results for exact name search

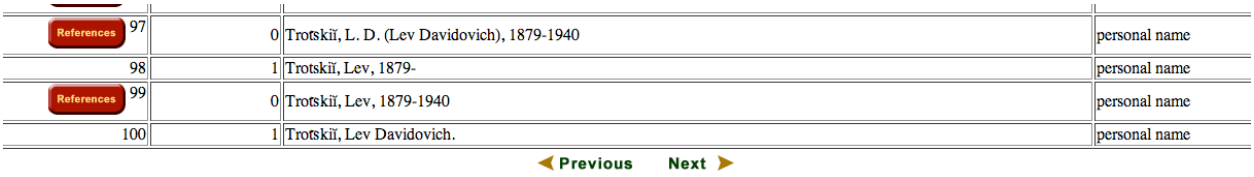

Figure 2. LC results for exact name search

Figure 3. VIAF identity cluster

HEADING: Trotsky, Leon, 1879-1940
000 03035cz a2200685n 450
001 1889400
005 20090711071350.0
008 830913nl acannaabn |b aaa
010 __ **|a** n 79054261
035 __ **|a** (OCoLC)oca00287399
040 __ **|a** DLC **|b** eng **|c** DLC **|d** DLC **|d** MdU **|d** NNC **|d** InU **|d** NIC **|d** OCoLC **|d** DLC-R **|d** DLC-R **|d** DLC-R **|d** OCoLC **|d** WU
100 1_ **|a** Trotsky, Leon, **|d** 1879-1940
400 1_ **|a** Bronshteĭn, Lev Davidovich, **|d** 1879-1940
400 1_ **|a** Trotski, Leo, **|d** 1879-1940
400 1_ **|w** nna **|a** Троцкий, сѣргѣй, Lev, **|d** 1879-1940
400 1_ **|a** Trotzki, Lee, **|d** 1879-1940
400 1_ **|a** Trotzky, Leon, **|d** 1879-1940
400 1_ **|a** Trotzky, N., **|d** 1879-1940
400 1_ **|a** Trotzkij, Leo, **|d** 1879-1940
400 1_ **|a** Trockij, Lev, **|d** 1879-1940
400 1_ **|a** Trotsky, L. D., **|d** 1879-1940
400 1_ **|a** Trocĭki, **|d** 1879-1940
400 1_ **|a** TÊ»o-lo-ssu-chi, **|d** 1879-1940
400 1_ **|a** Trotskiĭ, N., **|d** 1879-1940
400 1_ **|a** Trotsky, Lev Davidovitch, **|d** 1879-1940
400 1_ **|a** TÊ»o-lo-tzÊ»u-chi, **|d** 1879-1940
400 1_ **|a** Troĭtski, Leon, **|d** 1879-1940
400 1_ **|a** Bronstein, Leĭon, **|d** 1879-1940
400 1_ **|a** Троцкий, сѣргѣй, L. D. **|q** (Lev Davidovich), **|d** 1879-1940
400 1_ **|a** Trocki, Lev, **|d** 1879-1940
400 1_ **|a** Troĭ,tski, Leĭ,oĭ,ng, **|d** 1879-1940
400 1_ **|a** Torotsukiĭ,,, **|d** 1879-1940
400 1_ **|a** Torokkiĭ,,, **|d** 1879-1940
400 1_ **|a** Torotsukii, **|d** 1879-1940

Figure 4. LC Authority Record

Retrieval rates seem to depend on the vagaries of transliteration and historical influences. Russian revolutionaries and writers commonly took up pseudonyms. Vladimir Il'ich Lenin (born Ulyanov) is an obvious example. The VIAF offered more success in identifying pseudonyms. For instance, *Kolyshko, I. (Iosif), 1862-1938* is in the finding aid as *Kolyshko-Baian, Iosif Iosifovich,*—Baian is his pseudonym. The VIAF was able to provide headings for an exact match search while LC did not. However, the revolutionary and journalist G.D. Lindov (born Leĭteĭzen) has the LC authorized heading *Lindov, G.* An exact match search will bring up the heading for his pseudonym, although the user will have to go to the previous page because searching on "Lindov, G.D." causes LC to skip "Lindov, G." Petr Abramovich Bronshteĭn has also been enshrined under his pseudonym, *Garvi,* in the Library of Congress.

Discussion

The VIAF did not return any more identities than LC. There are, so far, no hidden figures from history available only in the VIAF. Its benefits lie in its interface, which allows a searcher to see available options as they fill in the Search field, and its greater flexibility in offering alternate versions of names.

Ed Summers at the Library of Congress analyzed the VIAF RDF dataset in May, 2012. He found 27,046,631 total links: 8,325,352 were from LC NACO, 327,455 from the National Library of Israel (the primary source of headings in Cyrillic), and only 997 from the Russian State Library (still in test status). When the Russian State Library contributes more records to the VIAF, a wider range of Russian names may become available, especially in Cyrillic.

In the meantime, HIA should consider using the VIAF to find and verify headings. If they can find a technical solution, they could use the VIAF and other linked data sources to provide more access and sharing of resources, mindful of the advantages and limitations described by Coyle and Bermès (2011).

Could HIA consider putting the VIAF ID in the Indexing terms of the EAD, so at least it is recorded if not actionable? (California Digital Library, 2005)

The <authorizedForm> and ><alternateForm> elements in EAC-CPF provide space to record alternate headings from “multiple, controlled vocabularies” (Wisser, 2001, p. 171). The VIAF ID can also be recorded in the <sources> element (Encoded Archival Context Working Group of the Society of American Archivists and the Staatsbibliothek zu Berlin, 2010).

The SNAC prototype aggregates finding aids, including those from the OAC, and derives EAC-CPF records from them (SNAC, 2010). Through this roundabout method, HIA records are becoming accessible to the Semantic Web.

Even the OCLC does not expect the VIAF to be the only unique ID used in the information community. Hickey (2011b) identified issues of conflicting conventions and concepts of identity, inconsistent data, different rules for creation and differentiation of identities, and varying priorities in using only one identification scheme. OCLC expects VIAF data to be combined with other schema and are involved themselves with two other identity systems, ISNI and ORCID.

An unaddressed challenge is verifying the accuracy of transliteration in the first place. Without handling each item directly, researchers must trust the transliteration of an archivist many years ago.

Conclusion

Although VIAF does not yet return many Cyrillic headings, it does offer a wider variety of transliterations and stylings, increasing the chances of finding a particular name. Although archives like at the Hoover Institution may not have the resources to fully digitize and link collection records, the VIAF is at least a useful resource for establishing traditional authority headings. Ventures like the SNAC prototype might serve as a useful workaround to increase linked data access to HIA holdings.

Future research could compare the number of headings in a set of identity clusters with alternate headings available in LC for the same set to more specifically measure VIAF benefits,

if any. A similar study could be conducted when more Russian State Library records are present in the VIAF.

References

Agenjo, X., Hernández, F., & Viedma, A. (2012). Data aggregation and dissemination of authority records through Linked Open Data in a European context. *Cataloging & Classification Quarterly, 50*(8), 803–829.

Brewer, M.M. (2009). Romanization of Cyrillic script: Core Competencies and basic research strategies for Slavic students, scholars, and educators. *Slavic & East European Information Resources, 10*, 244–256.

California Digital Library. (2005, April). OAC best practices guidelines for EAD (OAC BPG EAD). Retrieved from http://www.cdlib.org/services/dsc/contribute/docs/oacbpgead_v2-0.pdf

Coyle, K. & Bermès, E. (2011, September 11). Cluster archives. Retrieved from http://www.w3.org/2005/Incubator/lld/wiki/Cluster_Archives

Diekema, A.R. (2012) Multilinguality in the digital library: A review. *The Electronic Library, 30*(2), 165–181.

Dryden, J. (2010). A structure standard for archival context: EAC-CPF is here. *Journal of Archival Organization, 8,* 160–163. doi: 10.1080/15332748.2010.513325

El-Sherbini, M. & Chen, S. (2011). An assessment of the need to provide non-Roman subject access to the library online catalog. *Cataloging & Classification Quarterly, 49*(6), 457–483.

Encoded Archival Context Working Group of the Society of American Archivists and the Staatsbibliothek zu Berlin. (2010). EAC-CPF tag library draft. Retrieved from http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e6909

Fundación Marcelino Botín & Universidad de Cantabria. (2010, November 1). Use case ontology of Cantabria's cultural heritage. Retrieved from http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Ontology_of_Cantabria%27s_Cultural_Heritage

Hickey, T. (2012, July 2a). Identifiers for names. Retrieved from http://outgoing.typepad.com/outgoing/2012/07/identifiers-for-names.html

Hickey, T. (2012, July 6b). VIAF and other IDs. Retrieved from http://outgoing.typepad.com/outgoing/2012/07/viaf-and-other-ids.html

Hoover Institution. (2012). Library and Archives (About tab). Retrieved from http://www.hoover.org/library-and-archives

Jacobs, J.W., Summers, E., & Ankersen, E. (2004). Cyril: Expanding the horizons of MARC21. *Library Hi Tech, 22*(1), 8–17. doi: 10.1108/07378830410524459

Kiebuzinski, K. (2012). Samizdat and dissident archives: Trends in their acquisition, preservation, and access in North American repositories. *Slavic & East European Information Resources, 13*(1), 3–25.

Mazzini, S., and Ricci, F. (2011). EAC-CPF ontology and linked archival data. Proceedings of the *First International Workshop on Semantic Digital Archives,* September 29, 2011, Berlin.

OCLC. (2011, December 12). VIAF show and tell webinar. Retrieved from http://oclc.org/research/events/webinar.html

OCLC. (2012a, April 4). Virtual International Authority File service transitions to OCLC; contributing institutions continue to shape direction through VIAF Council. Retrieved from http://www.oclc.org/news/releases/2012/201224.htm

OCLC. (2012b). VIAF Virtual International Authority File. Retrieved from http://www.oclc.org/viaf

SNAC (Social Networks and Archival Context Project). (2010, December). Prototype. Retrieved from http://socialarchive.iath.virginia.edu/prototype.html

Stevenson, A. & Stevenson, J. (2010, October 22). Use case LOCAH. Retrieved from http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_LOCAH

Summers, E. (2012, May 15). Diving into VIAF. Retrieved from http://inkdroid.org/journal/2012/05/15/diving-into-viaf/

Tillett, B.B. (2007). Numbers to identify entities (ISADNs—International Standard Authority Data Numbers). *Cataloging & Classification Quarterly, 44*(3–4), 343–361.

Van Pulis, N. (2006). "First time use" (FTU) name headings, authority control, and NACO. *Library Management, 27*(8), 562–574.

Wisser, K.M. (2011). Describing entities and identities: The development and structure of Encoded Archival Context–Corporate Bodies, Persons, and Families. *Journal of Metadata, 11*(3–4), 166–175.

Yearl, M.K.K. (2012, August 22). Best Practices (Final). Retrieved from https://wiki.harvard.edu/confluence/pages/viewpage.action?pageId=50528611