Creating Multi-lingual Access Points in Archive Materials

Justine Withers

San Jose State University

Abstract

Millions of documents are squirreled away in archives around the world. If enough time and resources are available to create it, a finding aid, perhaps digital but most likely paper, is the only way to ascertain the contents of a collection. The challenges are multiplied when a collection contains multiple languages and is of interest to speakers of many languages. Among possible solutions, linking finding aids to multiple access points available from the Semantic Web could help.

Creating Multi-lingual Access Points in Archive Materials

In 1992, the new Russian government opened its archives and made available, among many resources, the records of the Communist International (Comintern), the worldwide Communist Party in operation from 1919 to 1943. Finally, in 2003, the online collection was released: an inventory of 55 million pages in more than 90 languages; 1.2 million digital scans of most frequently used documents; 220,000 records. (Doorn-Moissenko, 2005). To make the records available to a global audience, the United States Library of Congress coordinated the conversion of approximately 175,000 personal names from the Cyrillic to the Latin alphabet. Many of the names were "first time use" headings. The process involved transliterating the names from Russian, converting the transliteration to the proper spelling in the appropriate language (e.g. Khuan to Juan), sending the names out to 167 scholars in 54 countries to approve the list, and then incorporating changes in the database. Even after this arduous process, about half of the users, those with little or no Russian, have no way to search the text itself, because the content notes are still only available in Russian. (Bachman, 2005)

There has got to be a better way, right? While archives might never have the resources to fully catalog every item, linking names in finding aids to international authority files could lessen work and increase access.

Providing Context in Archives

While explaining the history and philosophy of Encoded Archival Content—Corporate Bodies, Persons, and Families (EAC-CPF), Wisser (2011) emphasizes the importance of separating contextual from bibliographic information. EAC-CPF aims to accommodate the

multiple identities and records that an entity might have, including multiple languages and scripts. Separating contextual and bibliographic information also eases automated search and retrieval, record sharing, and incorporation of outside sources, including authority files (Szary, 2005). Harper and Tillett (2007) strongly argue for libraries and archives participating in building a trustworthy Semantic Web, citing Miller (2004).

What's in a Name?

Snyman and van Rensburg (1999) created a prototype for an International Standard Author Number (ISAN), modeled after International Standard Book Numbers (ISBN). They argued for creating multiple access points over a single name authority. They hoped their prototype would address common challenges with disambiguating names: variations in presentation, shared names, changed names, redundancies in bibliographic and authority records, and time and expense better spent in other endeavors. Tillett, a strong advocate for an international authority file, repeated many of Snyman and van Rensburg's concerns when she explored the advantages and challenges of an international authority number (2007). In quickly sweeping aside the usefulness of single authority headings ["people need to have names they can read, in languages and scripts they can read" (p. 349)], she contrasted them with the idea of an International Standard Authority Data Number (ISADN) and an authority data cluster. The ISADN could work independently of particular languages and systems, although numbers could change if there were duplicate or disambiguated entities, or if the concept of entity itself changed. She deemed the data cluster "probably the most practical approach" (p. 358) because it could record all variants, languages, and scripts without international administration.

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). Despite the time it takes, she encourages catalogers to create authority headings at the same time they create a bibliographic record.

Several possible solutions to the problems surrounding name authorities exist, including: arXive, Research Papers in Economics (rePEc), Elsevier's Scopus Author Identifier, Thomson Reuters' Researcher ID, ProQuest Scholar Universe, NISO's International Standard Name Identifier (ISNI), the JISC Names Project, and the Virtual International Authority File (VIAF) (Carpenter, 2009).

When in Rome...

Differences in language and script only compound the challenges in disambiguating names. 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.

After their survey of librarians and end users of several non-Roman alphabets, El-Sherbini and Chen (2011) argue for increased research and use of non-Roman subject access. They found that end users have more difficulty than librarians with English search terms and

controlled vocabularies and they prefer using their own language and script. End users reported inconsistent Romanization and one librarian specifically mentioned "good subject access to Slavic materials may be lacking because language expertise in cataloging is lacking" (p. 470).

Kudo (2010) studied 950 Romanized Japanese titles. Although the amount of inconsistencies was small (2.63%), she points out that 30% of the inconsistencies were from Japanese vendors, who would perhaps be considered a more reliable source, and were not corrected in OCLC.

Seikel (2009) points to Resource Description and Access (RDA) as an improvement over AACR2 when it comes to non-Roman languages and scripts. Because RDA focuses on transcription, a cataloger can enter a name or title as it appears on the resource and link to other authority records.

Harnessing the Semantic Web

The Polymath Virtual Library (PVL) is a powerful example of the necessity and ability to link various sources and offer multiple access points. Agenjo, Hernandez, and Viedma (2012) describe the PVL as covering "Spanish, Hispano-American, Brazilian, and Portuguese polymaths from all times" (p. 803). The PVL creates MARC21 authority records that they consider "digital aggregates." Names are used both to identify and contextualize resources. All variants of a name are included to ease discovery, consistency, and queries. They rely heavily on the Virtual International Authority File, although taking into consideration its weakness in not identifying the language of its headings. Continuing the European tradition, libraries, archives, and museums

are considered equally important partners. To that end, the MARC21/RDA records can be easily downloaded as EAC-CPF.

In general, it seems the PVL is an outlier. When Diekema (2012) conducted a review of multi-lingual digital libraries, she determined that the subject was “understudied” (p. 175).

If access is more important than preservation, as Kiebuzinski (2012) argues, then archives must be more pro-active in providing multi-language, multi-script access points. Linking finding aids to the name authorities available in the Semantic Web, such as the Virtual International Authority File, would improve access.

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.

Bachman, R.D. (2005). The Comintern archives database: Bringing the archives to scholars. *Slavic & East European Information Resources, 6*(2/3), 23–36.

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.

Carpenter, T. (April, 2009). Working to solve the problems of name authority: The International Standard Name Identifier and other projects. Against the Grain. Retrieved from http://www.against-the-grain.com/TOCFiles/Standards_v21-2.pdf

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

Doorn-Misseenko, T. (2005). The Comintern archives online. *Slavic & East European Information Resources, 6*(2/3), 37–44.

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.

Harper, C.A. & Tillett, B.B. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging & Classification Quarterly, 43*(3–4), 47–68.

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.

Kudo, Y. (2010). A study of Romanization practice for Japanese language titles in OCLC WorldCat records. *Cataloging & Classification Quarterly, 48*(4), 279–302.

Seikel, M. (2009). No more Romanizing: The attempt to be less Anglocentric in RDA. *Cataloging & Classification Quarterly, 47*(8), 741–748.

Snyman, M.M.M. & van Rensburg, M.J. (1999). Reengineering name authority control. *The Electronic Library, 17*(5), 313–322.

Szary, R.V. (2005). Encoded archival content (EAC) and archival description: Rationale and background. *Journal of Archival Organization, 3*(2/3), 217–227.

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.