Violeta Ilik

Galter Health Sciences Library, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Distributed Person Data: Using Semantic Web Compliant Data in Subject Name Headings

Abstract: Providing efficient access to information is a crucial library mission. Subject classification is one of the major pillars that guarantees accessibility of records in libraries. In this paper we discuss the need to associate person IDs and URIs with subjects when a person happen to be subject of the document. Person can be a subject of documents and retrieval, as in the case of biographies, schools of thought in philosophy, politics, art, and literary criticism. Using semantic web compliant data in subject name headings enhances the ability to collocate topics about a person. Also, in retrieval, books about a person would be easily linked to works by that same person.

In the context of semantic web, it is expected that, as the available information grows, one would be more effective in the tasks of information retrieval. Information about a person or, as in the case of this paper, about a researcher exist in various databases, which can be discipline specific or publishers' databases, and in such a case they have an assigned identifier. They also exist in institutional directory databases. We argue that these various databases can be leveraged to support improved discoverability and retrieval of research output for individual authors and institutions, as well as works about those authors.

1. Introduction

The library community currently does not leverage the rich information about persons that is available from various data sources. Most of academic researchers publish only academic articles; therefore they do not exist in the Library of Congress Name Authority File (LCNAF). They do however exist in other data sources such as their respective institutional directory databases, discipline specific repositories and publisher databases, to name a few. In this paper we propose leveraging the data available in non-traditional library systems to support improved discoverability and retrieval of information about persons. To utilize these data sources, the library specific bibliographic databases such as the Online Computer Library Center (OCLC) and SkyRiver need to support the integration of data coming from non-traditional library sources. Currently OCLC supports "Control Heading" from the LCNAF. We think the time is right to expand the available sources for controlled headings to other data sources.

In this paper we will give general background information about the use of distributed person data in subject name headings. We state the problem and the ideal solution; we discuss the current crowded landscape of person data; then, in the absence of ideal solutions, we discuss what currently can be done to enhance bibliographic records with non-traditional library data; we focus on

one specific source of person data – VIVO, a semantic web application; and we make the case why using URIs in bibliographic records is important.

1.1. Background

The Subject Authority Cooperative Program (SACO) document from 2013 "Name Headings as Subjects H 430" states "most RDA and AACR2 name headings are usable not only as main or added entries, but also as subject headings." The document points out that there are some situations where a valid name heading may not be assigned as a subject heading and it is catalogers responsibility to be aware of and be able to identify such situations. The SACO document "Subject Usage Information in Name Authority Records H 432" from 2014 states "Information applicable to the usage of a name heading in subject cataloging may appear in a 667 field in the name authority record. Because the 667 field is also used for descriptive cataloging information, there may be more than one such field in a given NAR." The following is not allowed as a subject heading:

110 1# $a United States. $b President (1801-1809 : Jefferson)
667 ## $a SUBJECT USAGE: This heading is not valid for use as a subject.
 Works about this person are entered under Jefferson, Thomas,
 1743-1826.

The note in 667 points the cataloger toward using the name authority record for Jefferson, Thomas, 1743-1826 which is the actual Name Authority Record (NAR) for President Jefferson. This NAR has a Unique Resource Identifier (URI) assigned from the LCNAF and that URI is: id.loc.gov/authorities/names/n79089957. The URIs represent the most important piece of the puzzle in the current environment. As Verborgh and De Rose (2015) note the Web has become the world's largest source of knowledge not just for people but for machines too. The Web has evolved and now it offers something more powerful: linked data, which represents each piece of data as a link between two objects. These links, or URIs, enable software reasoners to derive knowledge that is human-specific.

Harper & Tillett (2006) point out that the benefits of authority control apply to metadata about the creators of resources as well as to subject access. These benefits include search precision, powerful navigation, collocation and linking between various tools and resources. It is well known that libraries have been dealing with identification, disambiguation and collocation of names of content creators since the beginning of cataloguing and at the same time with collocation of works about those same authors. Harper & Tillett (2006) mention the Friend of a Friend (FOAF) project and the need to find ways to integrate similar initiatives with existing mechanisms for name authority control in libraries, which will in turn help bring library catalogues into the mix of tools available on the Web. They also point out that the availability of library authority data in a more Web friendly format has the potential to positively influence the organization of the broad spectrum of Web content already available. The

development of a Virtual International Authority File (VIAF) has been a key idea moving forward this initiative.

Rebecca Dean (2004) states that the rapid growth of the World Wide Web and the emergence of numerous metadata schemas have spurred a re-examination of the way subject data are provided for Web resources. She points out the need of having simple and interoperable subject schema for metadata to ensure usability by non-catalogers and to enable users to search across both discipline boundaries and across information retrieval and storage systems. OCLC adapted the Library of Congress Subject Headings (LCSH) to create FAST (Faceted Application of Subject Terminology) which represents a simplified syntax that retains the very rich vocabulary of LCSH and is easier to understand, apply, control and use. As Dean (2004) points out, FAST headings for personal and corporate names are very similar, and in most cases exact, to the established name heading in the LC authority file.

Tillett (2004) noted that the "Web environment opens up new uses for authority records and new objectives to augment the traditional objectives." She further argues that the cataloguing community is moving "from the stand-alone authority files of a single institution, or even from the shared online files towards a goal of sharing authority files among all communities" (ibid). Tillett further discusses how the existence of authority records for an entity opens up new possibilities for links to other resources, and she includes examples such as home page for the entity described, links to biographical dictionaries and other reference sources on the Web. She mentions the possibility that the records would serve to control the various forms of names for an entity rather than having any single heading be the only authorized form. The Getty Institute in its Union List of Artists Names (ULAN) provides a listing of the forms of name for an entity that have been found in various resources, brought together to use for searching and displays. The Getty Institute released its Union List of Artists Names as Linked Open Data on 30th of April 2015, which will provide for much better retrieval and will resolve the current issue of retrieving all variant forms of names that do not have a clear connection between them.

The use of standard number has been mentioned as a way to enable collocation when there are several variant forms of a name used by an entity. The use of the language-neutral number for the entity in the bibliographic record was also mentioned few decades ago. As Tillett points out in 1980 IFLA proposed using an International Standard Authority Data Number (ISADN). She reminds us that other options were also in place such as the use of an International Standard Authority Number (ISAN), as well as the ISO International Standard Text Code (ISTC to identify works and expressions).

1.2. The current landscape

The current landscape that represents various systems that capture data about researchers is crowded. The systems fall into ten categories or types, according to the report by the OCLC Research, Registering Researchers in Authority Files Task Group (2014):

1. Authority hubs that provide a centralized location of authority records for multiple institutions - LC/NACO Authority File; ResearcherID; Virtual International Authority File (VIAF)
2. Current Research Information System (CRIS), which stores and manages data about research conducted at an institution and integrates it with data from external sources: Symplectic
3. Identifier hubs, providing a centralized registry of identifiers (2): International Standard Name Identifier (ISNI); Open Researcher and Contributor ID (ORCID)
4. National research portal, providing access to all research data stored in a nation's network of repositories.
5. Online encyclopedia: Wikipedia
6. Reference management, a system that helps scholars organize their research, collaborate with others, and discover the latest research: Mendeley
7. Research and collaboration hub, a centralized portal where scholars in a particular discipline can work together: nanoHUB
8. Researcher profile systems, networks that facilitate professional networking among scholars: VIVO, Google Scholar; LinkedIn
9. Subject author identifier system, a registration service to link scholars with the records about the works they have written: AuthorClaim
10. Subject repository such is arXiv hosted by Cornell Library.

The challenge of compiling scholarly output, measuring it, finding collaborators, disseminating research, tracking funded research is substantial, considering that each system creates identifiers for each author and co-author and those identifiers are valid and usable only in that particular system. Additional challenges include disambiguating researchers across institutions, preserving the institutional authority over researcher scholarly contributions, handling multiple names/transliteration, change of names, pseudonyms, multiple profiles, identifiers that are not linked and, most importantly, the various data standards that are not interoperable. One data source cannot solve all these problems and this is why one has to "think globally and act locally".

All of us working in all these various systems dealing with person data at the moment seem to have an open door to experiment and do something about solving the problem of uniquely identifying researchers within our systems, but also to think beyond individual systems in terms of global approaches that would connect and enable the exchange of information between all the systems. Producing and publishing data that is interoperable and/or reusable is the most important aspect of the problem.

1.3 Uniquely identifying a person

Everyone wants to uniquely identify a researcher: the researchers themselves, institutions, publishers, libraries, funding organizations and identity management systems. However, we also want to identify works that are about that individual researcher, works in which he/she is the actual subject. If there were a unique identifier for each researcher, all these tasks would be easier, but no such identifiers exist at the moment, and this is a source of considerable difficulties.

We need to address these issues and aid researchers in compiling and disseminating their scholarly output and to identify works about those

researchers. We need to be able to pull all the information about researchers from various databases and obtain the best possible aggregated data. Interoperability and reuse of data can be and should be addressed, since we live in an age that offers unique opportunities – we already have ways of publishing semantic web compliant data that can be interlinked and useful across databases.

Many of us are asking the same question: how to best utilize the various types of researcher identifiers to be able to accurately identify researchers and their scholarly outputs for better research outcomes and for better retrieval?

Identifying published works about those same researchers, works where researchers are the subject is equally important, since data containing that information would be extremely valuable not only to researchers themselves, but also to other stakeholders. In order to be able to represent that information in structured form, such as RDF, we need to have appropriate classes and properties defined in existing ontologies.

2. **The problem and the ideal solution**

Given that the library community currently does not leverage the rich information about persons that is available from various data sources, we propose to move forward with such an initiative. As stated above, most of academic researchers do not exist in the LCNAF, since they primarily publish academic articles. Since information about those researchers exist in other data sources such as discipline specific repositories (arXiv), publishers' databases (Scopus) and their respective institutional directory databases, to name a few, we propose that the library community opens up to the use of that data and moves towards implementation of technologies that would make this possible. OCLC, the biggest provider of bibliographic records to libraries around the world, is an essential partner in accomplishing this ideal solution, since they would have to open their doors to working with data from the non-traditional library world and enable the use and distribution of that data to the whole community.

In the ideal scenario, when a cataloger comes across a person name that does not have a NAR, it would make sense to attempt to control that name heading with possible established form of the name not only in the LCNAF but also in various other databases and systems, either library specific (other national libraries), discipline specific, or any other global data source. What this means is that a cataloger should be able to control the name heading of a person even if it comes from ISNI, ORCiD, the Getty ULAN Catalog, the Dutch Digital Author ID, the Brazilian Lattes Platform, just to name a few. We need to see this happen, so that when we come across authors that have no authority record created in the traditional library authority hubs we should be able to use non-traditional headings when importing authority records in our bibliographic data. This is nothing new since similar proposals have been known to exist. One can see from the many articles by Barbara Tillett that the need for using data from other data sources has already been proposed. What is new is that now we actually have the means to do it because we live in an age of semantic web

where data is published and structured in Resource Description Framework (RDF) format.

However, the bibliographic utilities (OCLC and SkyRiver) and the vendor community which includes Integrated Library Systems (ILS), automated authority control vendors, contract cataloging vendors and publishers, are not ready yet to offer a technical environment that enables us to control the name headings that come from linked data products and services.

2.1. What can we do now: proposed solution

The focus of this paper is the researcher profile system VIVO and the semantic web compliant data that can be leveraged in different systems. VIVO is a discovery tool that connects researchers across disciplines, institutions, geography and time. With this in mind, knowing that the ideal solution is not in the immediate future, we propose the use of alternative and non-traditional library data in NARs when used as subject headings – more specifically we propose enhancement of Name Authority Records by adding trusted VIVO person URIs.

The non-traditional library data in this case is the URI assigned to a person/researcher in an existing VIVO site. The non-traditional data can also include the identifier (ID) or a URI of a person in question that comes from ORCiD, ISNI, arXiv and various other databases and/or repositories. Similar project was conducted by the Fundacion Ignacio Larramendi and is described in their paper from 2012. The process of authority control is much broader than the one that traditionally takes place in large libraries or bibliographic agencies since it covers not only the task of identification of the entity "person," but also the tasks of contextualization.

2.2. What is VIVO?

VIVO was developed by Cornell University in 2003 and, with the help of an NIH grant in 2009, grew to become an open source project. VIVO is an open, shared platform for connecting scholars, research communities, campuses and countries using Linked Open Data. VIVO links data from institutional and public sources to create web profiles populated with researcher interests, activities and accomplishments.

The data sources used in VIVO are authoritative data sources that include university, agencies, repositories and aggregators. Private and sensitive information is never used or imported in VIVO. This rich information in VIVO profiles can be repurposed and shared with other institutional web pages and consumers. Data is housed and maintained at the local institution, which provides for regular updates. VIVO provides search results in faceted way. Researcher profiles are created via automated data feeds, but can also be customized to better fit the needs of the individual. Information is open source and it is available for use and re-use in other applications.

VIVO works with semantic standards defined by the World Wide Web Consortium (W3C), Resource Description Framework (RDF) standard for data

encoding and the Web Ontology Language (OWL) because of their wide adoption and because of the increasingly rich and diverse set of associated tools available (Börner, B.; Conlon, M.; Corson-Rikert, J.; Ding, Y., 2012 pg. 16). VIVO can be searched by keyword, as well as browsed by name, organization, or research area.

When discussing searching it is worth noting that providing efficient access to information can be approached in different ways, and almost always implies the use of an indexing language such as a taxonomy or an ontology. VIVO semantic web application uses the VIVO-ISF ontology. As noted on the Dura Space (2014) VIVO wiki page "The Integrated Semantic Framework ontology modules for VIVO (the VIVO-ISF ontology) provide a set of types (classes) and relationships (properties) to represent researchers and the full context in which they work." The ISF ontology leverages a number of ontologies in a unified, semantic structure. As stated on the DuraSpace (2013) VIVO wiki page "many of these ontologies are part of the Open Biological and Biomedical Ontologies organization, known as the OBO Foundry." These source ontologies may be imported in their entirety or included selectively through the MIREOT approach, which stands for minimum information to reference an external ontology term.

2.3. Why URIs and the current solution

Fran (2012) discussed the Semantic Web and Linked Data communities and their attempt to tackle the problem of the semantics of metadata by giving each concept a URI, so that the URI can be linked with other URIs by an equivalence or related by a relationship. Brickley (2011) states that "the strength of a URI-based approach is that the same URI can be given different expressions, allowing semantically different forms to be mapped and rich connections to be made between content." The problem that comes from such an approach is that governance is required to ensure that the stated relationships are valid and useful in the contexts that they are practically being used. As Fran points out:

> "in the context of knowing what an author has written, it is valid to assert that a book is the same as an e-book, whereas in the context of delivering a purchase to a customer, treating the two as the same is not valid." (Fran, 2012: 90)

In the context of the VIVO semantic web application the possibility to express the relationship of a person with the book that he/she is the subject of is not clearly represented in the VIVO-ISF ontology. Currently the VIVO-ISF supports only the featuredIn property and it is not clear if this property is sufficient to express the fact that the person is the subject of the specific work. For example featuredIn is defined in VIVO-ISF ontology as "This relates a person to an information resource that contains a featured article on that person" (vivoweb.org/files/vivo-isf-public-1.6.owl). Extending the definition of what featuredIn means will be sufficient, but as currently defined this property leads us to believe that we can only use it in cases when a person is featured in an article. Experimentation we performed proved that this property can be

successfully used to express the relationship a person has with any type of information resource (publication type) and not just an article.

For example if we have a person that is represented in our VIVO instance with the following URI: http://vivo.northwestern.edu/individual/nbf813378 and we have an article featuring him in our VIVO instance the relationship is represented as in Figure 1.

Figure 1: VIVO featuredIn property

One can also represent the featuredIn relationship between the person and a book, a biography, or any other publication type (see Figure 2).
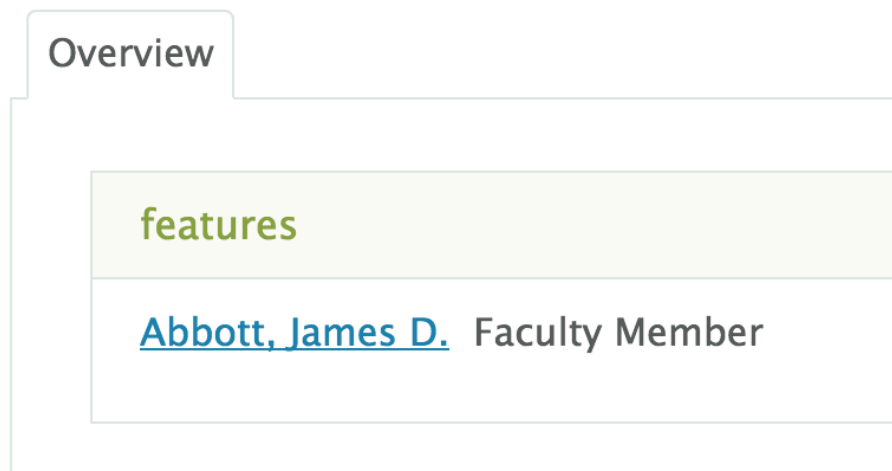

Figure 2: Person as subject of a book

As mentioned above few potential confusing points may be that it is not clear from the definition of the featuredIn property that one can represent the relationship between a person as a subject with any publication types, not just articles, and another one is that this property is not representing the aboutness in a sense as understood by traditional library world. As a reminder, according to Žumer, Zeng, & Salaba (2012, p. 5)

> "aboutness is a statement of what a work (an information resource) is about, what topic it covers. It usually is expressed with the relationship "work is about/covers/has subject x, y, and z." Information professionals when analysing the content of the resource through the process of identifying the topics discussed or otherwise represented in the resource, are trying to assess the user information needs that can be fulfilled by the resource at a future time. The user, on the other hand, when trying to fulfil and information need, wants to find the appropriate resources or verify weather the resource(s) obtained contain relevant information."

One way to solve this problem is to possibly extend the VIVO-ISF ontology to represent the aboutness aspect in more clear way since it is very important to know when a person is a subject of a work. The property will need to have *agent* in the range and needs to be defined in the domain of *information content entity (ICE)*.

As of now the only way to capture non-traditional data in the form of IDs or URIs is to enhance the authority records with the VIVO URIs, or any other URIs, by adding the values in the field 024, first indicator 7 with the value of the URI in subfield *a* as shown below:

024 7 # $a http://vivo.northwestern.edu/individual/nbf813378 $2 uri

In case the cataloguer is not at a NACO institution and cannot create the NAR for the person there should be a possibility to add the URI in 600 - Subject Added Entry-Personal Name, in subfield $0 - Authority record control number or standard number, as shown below:

600   10$a   Abbott,   James   D. $0(uri) http://vivo.northwestern.edu/individual/nbf813378

The authority record control numbers, such as the Library of Congress Control Numbers, can also be used as unique persistent identifiers. Lately we are seeing the trend of using the ORCiD in authority records, more specifically in the 024, Other Standard Identifier field, with the value of ORCiD entered in subfield 2. The same is true for the VIAF, ISNI, or the LCCN (see Figure 3).

010		‡a no2005043559
024	7	‡a 0000000043605828 ‡2 isni
024	7	‡a 0000-0002-7442-635X ‡2 orcid
024	7	‡a 36648407 ‡2 viaf

Figure 3: Use of person IDs in a NAR

Instead of using the value of ORCiD, ISNI, or VIAF as a number string why not use the machine readable URI as already described by Agenjo, X.; Hernández, F; & Viedma, A (2012)? Using URIs in the 024 field in authority records would make possible the retrieval of relevant data about the person in real time. For example in the case of an individual by the name Michael Conlon whose ORCiD is: orcid.org/0000-0002-1304-8447 and the VIVO URI that represents him is: vivo.ufl.edu/individual/n25562 we can add those values in 024. The problem with this specific individual is that he does not have his name established in the LCNAF, since he mostly publishes academic articles. For a large number of scholars and researchers this is the case. In our earlier papers and presentations we have argued for the need to leverage the existing data from databases either sematic web or relational (Ilik, 2015). Not every person needs to exist in the LCNAF, but the information about that person that exists

in other systems can still be leveraged by library specific databases like OCLC and SkyRiver. This brings us back to the problem presented in this paper and to its ideal solution, as indicated above.

3. VIVO and the ideal solution

We have discussed the possibility of moving towards creating persistent VIVO URIs that resolve at the central VIVO Registry, which represents a hub aggregation of all of the local deployments. Only at that point we would be able to introduce the VIVO Central Registry as a partner to OCLC, SkyRiver and the Program for Cooperative Cataloging (PCC), to name the widely known library specific bodies dealing with name authorities. This partnership means that catalogers would be able to control the person name from not only LCNAF but also from other trusted data sources, VIVO being one example (see Figure 4).
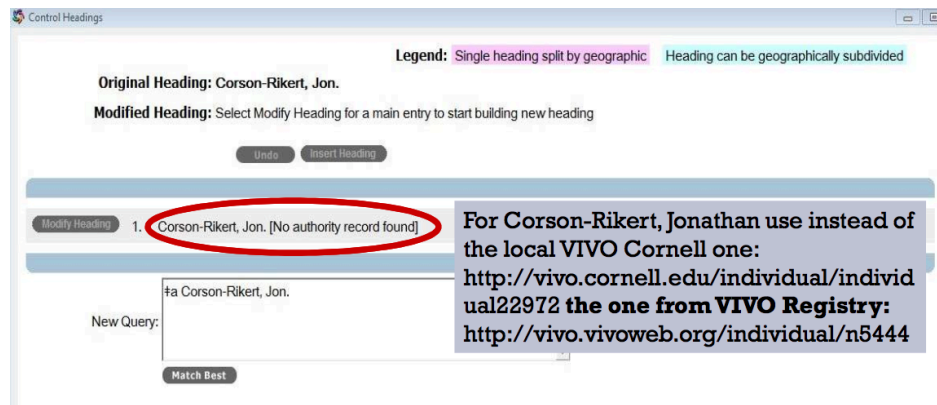

Figure 4: OCLC, VIVO and possible partnership

The goal is to work on asserting rdfs:seeAlso between local URIs and the central URIs. These URIs are RDF representation of a story or set of statements about that person that each institution makes. To assert relationships between external data sources and VIVO, like for example ORCiD or ISNI, we need to assert the sameAs relationship. In order to support maximal data integration across sites and platforms, VIVO needs to support persistent VIVO person URIs that resolve to a central VIVO Registry, management of seeAlso assertions for person instances across various VIVO platforms, management of sameAs assertions for person instances across different data sources, which would lead to interoperability.

4. Conclusion

In this paper we discussed the need to associate person IDs and URIs with subjects when a person happen to be subject of the document. Person can be a subject of documents and retrieval, as in the case of biographies, schools of

thought in philosophy, politics, art and literary criticism. Authoritative information about researchers exist in various databases, which can be discipline specific, publisher's, or institutional. These various databases can be leveraged to support improved discoverability and retrieval of research output for individual authors and institutions, as well as works about those authors. We propose the use of alternative and non-traditional library data in NARs when used as subject headings – more specifically we propose enhancement of Name Authority Records by adding trusted VIVO person URIs. URIs for VIVOs that commit to maintaining the historical data should be registered for on-going discoverability. VIVO has the capacity of recording the ORCiD ID or the Scopus ID to aid in sameAs linking. Creating extensions to the core VIVO-ISF ontology would make possible the recording of the ISNI, VIAF and many other IDs that come from various databases and systems. We need to see this happen so that whenever we come across authors that have no authority record created in the traditional library authority hubs we would be able to use non-traditional headings in order to import authority records in our bibliographic data. Such headings would be utilized in the process of subject name authority control.

Acknowledgment

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).

Börner, B.; Conlon, M.; Corson-Rikert, J.; Ding, Y. (2012). *VIVO: A Semantic Approach to Scholarly Networking and Discovery; Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool Publishers, pp. 178 pages, doi:10.2200/S00428ED1V01Y201207WBE002

Brickley D (2011) Video Linking: Archives and Encyclopedias(blog post). Available at: http://danbri.org/words/2011/02/01/658

Dean, R. (2004). FAST: Development of Simplified Headings for Metadata, *Cataloging & Classification Quarterly*, 39 (1-2), pp. 331-352.

DuraSpace, (2012). VIVO Glossary. Available at: https://wiki.duraspace.org/display/VIVO/VIVO+Glossary

DuraSpace, (2014). VIVO-ISF Ontology. Available at: https://wiki.duraspace.org/display/VIVO/VIVO-ISF+Ontology

DuraSpace, (2013). Source ontologies for VIVO. Available at: https://wiki.duraspace.org/display/VIVO/Source+ontologies+for+VIVO

Fran, A. (2012). Building bridges: Linking diverse classification schemes as part of a technology change project. *Business Information Review*, 29 (2), pp. 87-94. doi: 10.1177/0266382112446284

Harper, C.; Tillett, B. (2007). Library of Congress controlled vocabularies and their application to the semantic web. *Cataloging & Classification Quarterly,* 43 (3-4). Available at: https://scholarsbank.uoregon.edu/dspace/handle/1794/3269

Ilik, V. (2015), Cataloger Makeover: Creating Non-MARC Name Authorities. *Cataloging & Classification Quarterly*, 53 (3/4), pp. 382-398.

Name Headings as Subjects H 430 (2013). The Subject Authority Cooperative Program (SACO). Available at: http://www.loc.gov/aba/publications/FreeSHM/H0430.pdf

Smith-Yoshimura, K.; Altman, M.; Conlon, M.; Cristán, A.; Dawson, L.; Dunham, J.; Hickey, T.; Hook, D.; Horstmann, W.; MacEwan, A.; Schreur, P.; Smart, L.; Wacker, M.; & Woutersen, S. (2014). Registering Researchers in Authority Files. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearchregistering-researchers-2014-a4.pdf

Subject Usage Information in Name Authority Records H 432 (2014). The Subject Authority Cooperative Program (SACO). Available at: http://www.loc.gov/aba/publications/FreeSHM/H0432.pdf

Tillett, B. (2004). Authority control : state of the art and new perspectives. *Cataloging & Classification Quarterly*, 38 (3/4), pp. 23-41. Available at: http://www.sba.unifi.it/ac/relazioni/tillett_eng.pdf

Verborgh, R.; De Roo, J. (2015), Drawing Conclusions from Linked Data on the Web, IEEE Software, 32 (5), pp. 23-17. Available at: http://online.qmags.com/ISW0515?cid=3244717&eid=19361#pg25

Verify and Control Headings in Bibliographic Records (2013). OCLC. Available at: http://www.oclc.org/content/dam/support/connexion/documentation/browser/authoriti es/apply_ac_bib_records/apply_ac_bib_records_pdf.pdf

Žumer, M.; Salaba, A.; Zeng, M. L. (2012). *FRSAD : conceptual modeling of aboutness.* Santa Barbara, CA: ABC-CLIO, LLC.