Software Applications Ecosystem for Authority Control

Leandro Tabares Martín[1], Félix Oscar Fernández Peña[3], Amed Abel Leiva Mederos[2], Marc Goovaerts[4], Dailién Calzadilla Reyes[1], and Wilbert Alberto Ruano Álvarez[1]

[1] Universidad de las Ciencias Informáticas ltmartin@uci.cu
dcreyes@estudiantes.uci.cu waruano@estudiantes.uci.cu
[2] Universidad Central "Marta Abreu" de las Villas amed@uclv.edu.cu
[3] Instituto Superior Politécnico "José Antonio Echeverría"
felix@ceis.cujae.edu.cu
[4] Universiteit Hasselt marc.goovaerts@uhasselt.be

Abstract. Authority control is recognized as an expensive task in the cataloging process. This is actually an active research field in libraries and related research institutions even when several approaches have been proposed in this research area. In this paper, we propose AUCTORITAS, a tool for exposing high value services on the web for the authority control in a generic institution environment. This paper describes the application ecosystem behind AUCTORITAS and how the semantic web languages make possible the semantic integration of heterogeneous applications. Likewise we evaluate the applicability of the proposal for academic libraries.

Keywords: Authority Control, Linked Open Data, Semantic Web

1 Introduction

Authority Control is the most expensive part of the cataloging process [20,7,21], it is a global problem, affecting not only libraries but organizations of all kinds [16]. Authority Control is necessary for meeting the catalog's objectives of enabling users to find the works of an author and to collocate all works of a person or corporate body. Authority control virtues have been debated and restated for decades. Catalogers for at least a century and a half have documented their decisions on how the single, authorized form of name for each entity should be represented in their catalog [20]. Several efforts has been made by library institutions in order to share their authority records [20,10] but, the publication of authority data on the Web in an heterogeneous or arbitrary way produces inefficiency in information retrieval and creates complications when attributing authority to a given work. The need to improve the interoperability within the world Wide Web gave rise to the development of the Semantic Web [2].

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [2]. The current work aims to create an applications ecosystem enabling authority control capacities for external applications, by reusing semantically structured data shared by different institutions. This work is structured as follows: a section exposing related work where authority control state of the art and specifically AUTHORIS, semantic web, linked open data, Openlink Virtuoso and VIVO are addressed. After that the applications ecosystem is explained in detail, evaluated and we conclude with future steps to follow in order to improve our proposal.

2 Related Work

2.1 Authority Control

Authority control is a matter that has exacted the efforts of generations of librarians and catalogers. The need to uniformly record information on each author included in a catalog is addressed in work and research stemming from several international organizations. Libraries and organizations of international prestige such as the United States Library of Congress (LOC), the Bibliothèque Nationale de France and International Federation of Library Associations (IFLA) acknowledge the fact that the information exchange protocols on the Web are insufficient means of controlling authority in the catalogs and systems of library management [16].

A brief outline of authority control would include the following landmarks:

- The need for authority control is made explicit, and the Name Authority Cooperative (NACO) comes to light with the US Library of Congress [16]. In Asia, the Hong Kong Chinese Authority Name (HKCAN) is established. This meant recognition of the issue in just two organizations worldwide - far [16], however, from the syndetic goals set forth by Charles Cutter in the nineteenth century [6].
- Lubetzky [17] improves the search and retrieval of authored works in bibliographic records, eliminating the deficiencies that interfered with the retrieval and location of authors in a catalog.
- Bregzis [4] creates the ISADN (International Standard Authority Data Number) to overcome difficulties when retrieving bibliographic records with works relative to a given author and with works recorded under a uniform title.

The Online Computer Library Center (OCLC), IFLA and LOC have fueled initiatives for authority control by sharing the records of various cataloguing agencies [16]. Fruit of this work is the Virtual International Authority File (VIAF), which has meant advances in the construction and generation of authority entries, though it has no reached all the major information institutions at the international level [3].

2.2 AUTHORIS

The need of creating high quality authority records has led the creation of tools like AUTHORIS [16]. AUTHORIS aspires to facilitate the processing of authority data in a standardized fashion, following the principles of Linked Data [1]. AUTHORIS allows the automatic generation of authority records by using learning rules [16], however AUTHORIS does not takes advantage the high quality authority records shared by library institutions. Further AUTHORIS does not implement an interface for providing data services for other software applications.

2.3 Semantic Web

Since Resource Description Framework (RDF) made it possible to define the meaning of data in a machine readable form [19], it seems that the semantic web technologies could be helpful in the integration of data managed between heterogeneous software applications. The evolution of RDF into Web Ontology Language (OWL) allows a richer semantic description based on Description Logic [12]. OWL is a formal language for representing ontologies in the Semantic Web [12]. This language has been used in many specific scenarios for the construction of flexible data semantic models [9,13,14]. Several knowledge organization systems takes advantage of semantic web technologies [18,11,8], SKOS [18] is one of them. In this proposal we reuse SKOS structured information sources provided by institutions and reuse their data.

2.4 Linked Open Data

The concept of Linked Open Data (LOD) is based on the idea of linking publicly available data "silos" on the internet. By linking data, all of the data objects become related to each other. By determining a number of rules about these relationships, such inter-linked data can be "understood" by machines and algorithms, which enables global data mining approaches and the discovery of truly new associations, patterns and knowledge. LOD is based on the Resource Description Framework (RDF) data model, which formulates syntax and rules about data and resources as well as their location on the internet [15].

Implementation of LOD approaches requires adherence to the four basic components as formulated by Berners-Lee [1]:

- Use Uniform Resource Identifiers to uniquely identify data.
- Use the Hypertext Transfer Protocol (HTTP) so that people, web agents and data mining tools can access and refer to data.
- A URI has to refer to usable information that can be provided with the RDF and queried with the Simple Protocol and RDF Query Language (SPARQL).
- Links to other RDF resources should be established in support of growing a world wide network of publically available and allowing for truly interdisciplinary data mining.

There is a tremendous potential for the library community to play a significant role in realizing Berners-Lee's vision, the idea of moving thesauri, controlled vocabularies, and related services into formats that are better able to work with other Web Services and software applications is particularly significant. Converting these tools and vocabularies to Semantic Web standards will provide limitless potential for putting them in a myriad new ways [10].

2.5 OpenLink Virtuoso

OpenLink Virtuoso[5] is an innovative enterprise grade multi-model data server for agile enterprises and individuals. The hybrid server architecture of Virtuoso enables it to offer traditionally distinct server functionality within a single product that covers the following areas:

- SQL Relational Tables Data Management.
- RDF Relational Property Graphs Data Management.
- Content Management.
- Web and other Document File Services.
- Linked Open Data Deployment.
- Web Application Server.

Virtuoso capabilities managing Linked Open Data allow us to expose vocabularies such as AGROVOC[6] through its SPARQL endpoint and make them query available for other applications such as AUCTORITAS. AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization of the United Nations with over 32000 concepts. CCS vocabulary for Computer Sciences and MESH for Medicine and Life Sciences can be managed by Virtuoso also.

2.6 VIVO

VIVO[7] is an open source semantic web application originally implemented at Cornell University that enables the discovery of research and scholarship across disciplines, it supports browsing and search function which returns faceted results for rapid retrieval of desired information. VIVO allows also to manage authors and institution profiles and generates a Uniform Resource Identifier for each one of them.

All the information managed by VIVO is structured as Linked Open Data, this structure improves information discovery [15] and also facilitates the generation of authorship relations graphs. Information inside VIVO is SPARQL queriable and new ontologies can be added in order to expand VIVO's capabilities of semantically manage data. VIVO and AUCTORITAS integration is intended to use the information coming from institutions with intellectual production, so integrated library systems and digital repositories can use VIVO's data for uniquely identifying its authors.

[5] http://virtuoso.openlinksw.com/
[6] http://aims.fao.org/es/agrovoc
[7] http://vivoweb.org/

3 Applications Ecosystem

3.1 Preprocessing Tool

The Library of Congress of United States of USA has shared their authority graph[8] to the international community with the aim their data can be reused. In that graph information like author names and authoritative labels can be found in several different languages. AUCTORITAS is intended to initially use data expressed as Latin characters, so with the goal of extracting relevant information for AUCTORITAS coming from the LOC's graph, we created a preprocessing tool[9] that populates a relational database. The database structure is represented on figure 1.

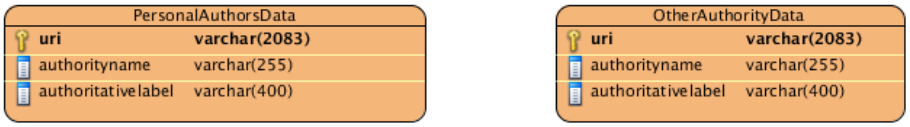

Fig. 1. Physical model of the preprocessing tool relational database

For improving the graph processing we have divided the original LOC's authority graph into one hundred and ninety-five RDF files containing around four million triples each one. The preprocessing tool uses parallel processing to optimize the processing of the graph by splitting the load in multiple threads. A regular expression is used for matching patterns contained in the graph, determining which information is about a personal author. After a preprocessing with testing purposes for six files of the LOC's we get 18592 personal authorities records stored in our database, ready to be exposed through AUCTORITAS services. Other 13215 records were identified as non-personal authority records or non-latin characters personal authority records, so they were stored into another table for further processing. This preprocessing phase allows us to reduce the significant-data table size in a 41.5% by eliminating non-relevant information for the tool.

3.2 AUCTORITAS interface

AUCTORITAS interface is the main entry point for our applications ecosystem, it can be seen as a three dimensional vector A(v,p,w) where:

- v is a linked data datasource stored at Virtuoso.
- p is the relational database stored at PostgreSQL.
- w is the data managed by VIVO.

[8] http://id.loc.gov/static/data/authoritiesnames.rdfxml.madsrdf.gz
[9] https://drive.google.com/file/d/0B-Pkaic4zIO8T2FnQVIxdWR5WFU/view?usp=sharing

AUCTORITAS interface has four main functionalities exposed as REST web services:

- Search for personal authors information.
- Search for corporate authors information.
- Retrieve registered controlled vocabularies list.
- Search for an authorized term on a specified controlled vocabulary.

All these functionalities are explained in figures from 2 to 5.

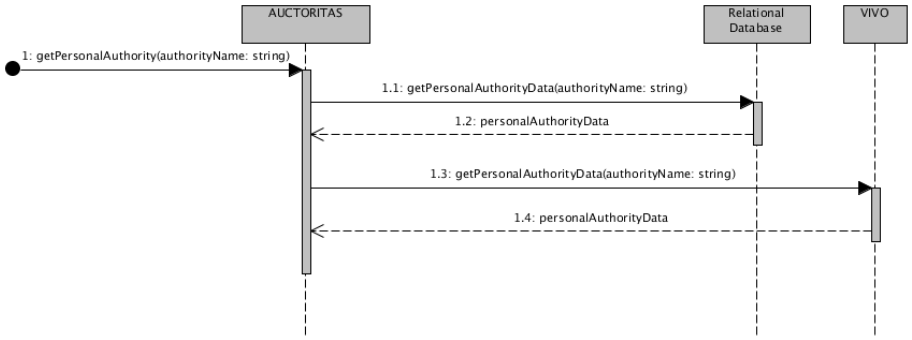

Fig. 2. Sequence diagram for a request about personal authors

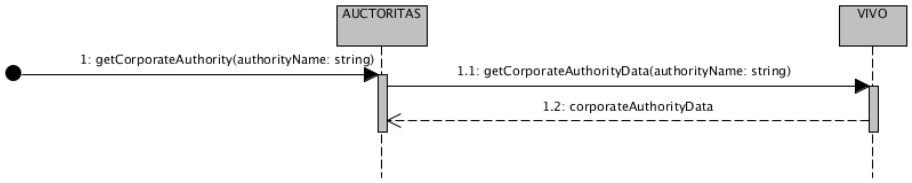

Fig. 3. Sequence diagram for a request about corporate authors

External applications like integrated library systems (ILS) and digital repositories send requests to AUCTORITAS with the objective of uniquely identify their authority entries, then AUCTORITAS queries its available information sources and retrieves the requested information structured as a XML. Figure 6 shows AUCTORITAS answer to an external system after searching for "database" term on the ACM Controlled Vocabulary.

Two main elements are sent as answer in this case, the identifier of the term in the requested vocabulary and the authorized term by itself. The identifier of the term is computer oriented for uniquely identify it by using an URI and the authorized term is what the person using the system sees.

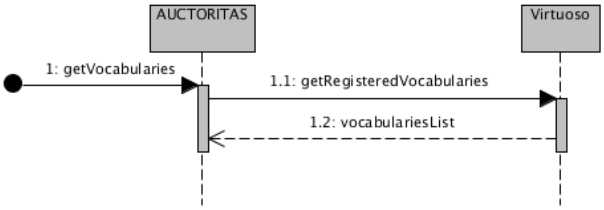

Fig. 4. Sequence diagram for a request about the registered vocabularies list

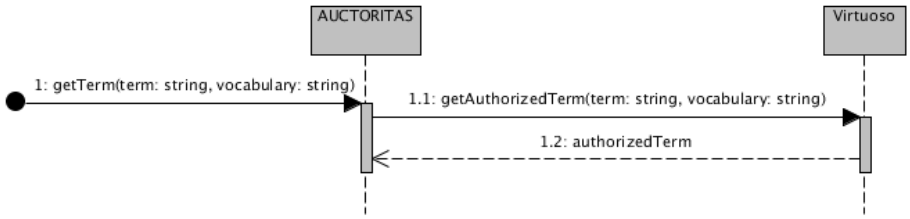

Fig. 5. Sequence diagram for a request about a term from a registered vocabulary

Also external applications may query AUCTORITAS services for personal author entries. Figure 7 shows AUCTORITAS answer to a query about Jorge Israel Rivera Zamora over LOC's graph processed information.

In our proposal four applications are integrated to conform what will call the Applications Ecosystem[10] as shown in the figure 8.

The ILS is represented by ABCD, which is a system that allows librarians to manage their library data in a digital way. For managing Digital Repositories a customization of DSpace was developed by the University of Computer Sciences of Cuba and that customization was integrated with AUCTORITAS. By exposing AUCTORITAS features as Web Services other software developers are allowed to consume AUCTORITAS services. Also AUCTORITAS provides a mechanism to reuse SKOS-structured controlled vocabularies, so it is not limited only to use the presented vocabularies. This mechanism is to add the string "vocab" before the last section of the URI that identifies the vocabulary to register it in Virtuoso, for example: *http* : *//ccs.vocab.cu*. AUCTORITAS uses a regular expression to identify this URI structure and use it as a controlled vocabulary.

[10] This proposal has been developed thanks to the Flemish Project VLIR-UOS.

```
<?xml version="1.0"?>
<vocabularyEntry>
<identifier>http://totem.semedica.com/taxonomy/The ACM Computing Classification System (CCS)#10002952</identifier>
<authorizedTerm>Data management systems</authorizedTerm>
</vocabularyEntry>
```

Fig. 6. AUCTORITAS answer to a query over ACM controlled vocabulary

```
<?xml version="1.0"?>
<authorityEntry>
<identifier>http://id.loc.gov/authorities/names/no2010096115</identifier>
<name>Jorge Israel Rivera Zamora</name>
<label>Rivera Zamora, Jorge Israel</label>
</authorityEntry>
```

Fig. 7. AUCTORITAS answer to a query about Jorge Israel Rivera Zamora

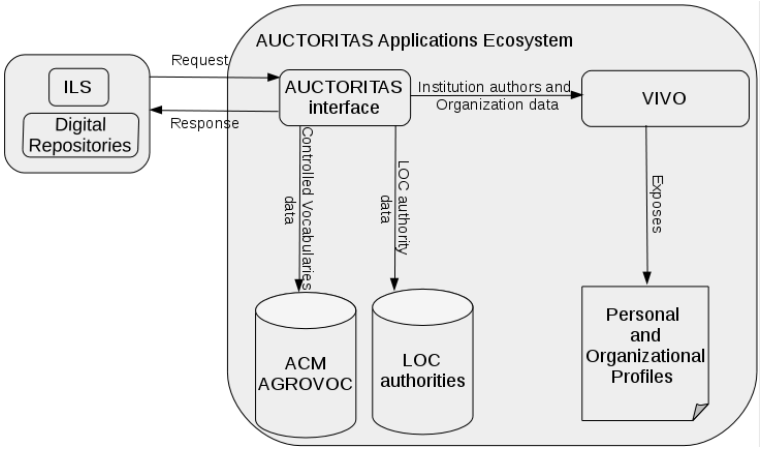

Fig. 8. Applications ecosystem overview

3.3 Querying VIVO

AUCTORITAS queries to VIVO are done through VIVO's SPARQL endpoint, which is deployed in the URL [*vivoAddress*]/*api*/*sparqlQuery*, for example: http://localhost:8080/vivo/api/sparqlQuery. Queries to this endpoint must contain the parameters specified in Table 1:

Table 1. VIVO's SPARQL endpoint parameters

Parameter name	Parameter value
email	The email address of a VIVO administrative account
password	The password of the VIVO administrative account
query	The SPARQL query

VIVO 1.7 was used in order to manage personal and organizational data as LOD. Besides the main authority control that we achieve with the integration of this tools, VIVO also allows our institutions to make scientometric studies like the generation of science maps and the creation of graphics illustrating the authorship relations. At the same time the information managed by VIVO can be browsed, so the user can discover new related information and access to the full institutional scientific production.

4 Evaluation

The Applications Ecosystem has been evaluated in the University of Computer Sciences of Cuba according to the criteria proposed by RDA [5]. The elements taken into account in this evaluation were:

- Be usable primarily within the library community, but able to be used by other communities.
- Enable users to find, identify, select, and obtain resources appropriate to their information needs.
- Be compatible with the descriptions and access points in existing catalogs and databases.
- Be readily adaptable to new emerging database structures.
- Be optimized for use as an online tool.
- Be easy and efficient to use, both as a working tool and for training purposes.

A total 14 of users participated, divided into the following categories:

- Twelve Library and Information Science specialists with more than ten years of experience in cataloging.
- Two Computer Software Engineering specialists with more than five years of experience in programming.

All of these users interacted with the ILS and the Digital Repository System introducing new records into them. A total of one hundred new records were created in both systems. Figure 9 shows the amount of errors detected during the evaluation.

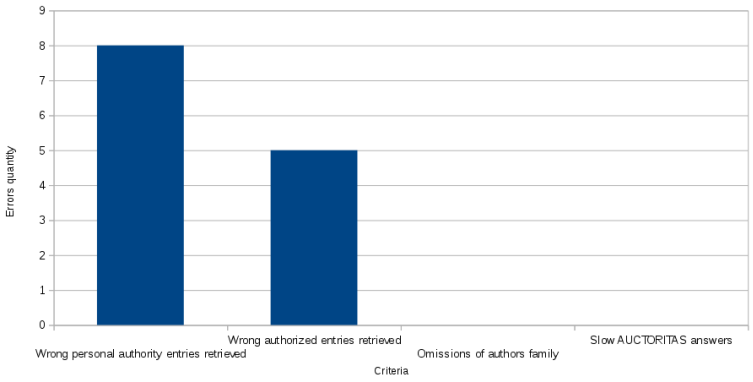

Fig. 9. Errors detected during the evaluation

The evaluation concluded that the Applications Ecosystem is usable by library institutions and extensible to other institutions that needs it. The retrieval

of information appropriate to users needs is partially met because there were problems about the precision in retrieving personal authority entries. The solution is compatible with existing catalogs and databases structured as SKOS and is easily adaptable to new databases structures. The Applications Ecosystem is designed to be used as an online tool and is easy and efficient to use in production or in training environments.

5 Conclusions and Future Work

The development of authority control faces new challenges in the Semantic Web. The need to facilitate interoperability and connection among non-bibliographic and bibliographic entities is one promising area to be implemented by the designers and developers of future cataloguing and authority control systems.

The tools presented in this paper are one step further in the development of new authority control systems. Still there is a lot of work to do in order to fully reuse available authority data shared by institutions. In new versions of AUCTORITAS similarity measures will be incorporated in order to create a better information retrieval mechanism. Also the incorporation of corporate authors coming from available authorities data sources has to be added to the preprocessing tool.

Multilinguality in non-latin characters is one aspect that has to be incorporated, for the purpose of allowing to other countries the usage of AUCTORITAS benefits. AUCTORITAS still has some limitations to be solved, but it provides a flexible mechanism to be extended in order to support the different authority control scenarios needed by Cuban institutions.

References

1. Berners-Lee, T.: Linked Data (2006), http://www.w3.org/DesignIssues/LinkedData.html
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American magazine 284(5), 34–43 (2001)
3. Bourdon, F., Zillhardt, S.: Author: Vers une base européenne de notices d'autorité auteurs. International cataloguing and bibliographic control 26(2), 34–37 (1997)
4. Bregzis, R.: The syndetic structure of the catalog. Authority control: the key to tomorrow's catalog. Proceedings of the 1979 Library and Information Technology Association Institute, Mary W. Ghikas ed. Phoenix: AZ (1982)
5. Committee, U.R.T.C., et al.: Report and recommendations of the us rda test coordinating committee. Executive Summary. Online unter: http://www. nlm. nih. gov/tsd/cataloging/RDA_report_executive_summary. pdf (letzter Zugriff: 10.09. 2011) (2011)
6. Cutter, C.A.: Rules for a printed dictionary catalogue. US Government Printing Office (1889)
7. Diaz-Valenzuela, I., Martin-Bautista, M.J., Vila, M.A., Campaña, J.R.: An automatic system for identifying authorities in digital libraries. Expert Systems with Applications 40(10), 3994–4002 (2013)
8. Dunsire, G., Willer, M.: Standard library metadata models and structures for the semantic web. Library hi tech news 28(3), 1–12 (2011)

9. H. Agus-Santoso, S.C.H., Abdul-Mehdi, Z.: Ontology extraction from relational database: Concept hierarchy as background knowledge. Knowledge-based Systems 24(3), 457–464 (2011)
10. Harper, C.A., Tillett, B.B.: Library of Congress controlled vocabularies and their application to the Semantic Web. Cataloging & Classification Quarterly 43(3-4), 47–68 (2007)
11. Hodge, G.: Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. ERIC (2000)
12. Ian Horrocks, P.F.P.S., van Harmelen, F.: From SHIQ and RDF to OWL: The Making of a Web Ontology Language. Web Semantics 1(1), 7–26 (2003)
13. K. Čerāns, G.B.: RDB2OWL: a RDB-to-RDF/OWL Mapping Specification Language. Proceeding of the 2011 Conference on Databases and Information Systems, Amsterdam, The Netherlands. IOS Press pp. 139–152 (2010)
14. K. Munir, M.O., McClatchey, R.: Ontology-driven relational query formulation using the semantic and assertional capabilities of OWL-DL. Knowledge-based Systems 35, 144–159 (2012)
15. Lausch, A., Schmidt, A., Tischendorf, L.: Data mining and linked open data – New perspectives for data analysis in environmental research. Ecological Modelling 295, 5–17 (2015), http://dx.doi.org/10.1016/j.ecolmodel.2014.09.018
16. Leiva-Mederos, A., Senso, J.a., Domínguez-Velasco, S., Hípola, P.: AUTHORIS: a tool for authority control in the Semantic Web. Library Hi Tech 31(3), 536 – 553 (2013), http://softwaredocumental.org/repositorio/Texto-completo/2013 - Leiva-Mederos et al. - AUTHORIS a tool for authority control in the Semantic Web.pdf
17. Lubetzky, S., Hayes, R.M.: The Principles of Cataloging: Report. Institute of Library Research, University of California (1969)
18. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference. W3C recommendation 18, W3C (2009)
19. Motik, B., Horrocks, I., Sattler, U.: Bridging the gap between owl and relational databases. Web Semantics: Science, Services and Agents on the World Wide Web 7(2), 74–89 (2009)
20. Tillet, B.B.: Authority Control: State of the Art and New Perspectives. Cataloging & Classification Quarterly Volume 38(3-4), 23–41 (2004), http://www.tandfonline.com/doi/abs/10.1300/J104v38n03_04
21. West, W.L., Miller, H.S., Wilson, K.: Electronic journals: Cataloging and management practices in academic libraries. Serials Review 37(4), 267–274 (2011)