# AUCTORITAS: A Semantic Web-based tool for Authority Control

Leandro Tabares Martín[1], Félix Oscar Fernández Peña[2], and Amed Abel Leiva Mederos[3]

[1] Universidad de las Ciencias Informáticas ltmartin@uci.cu
[2] Universidad Técnica de Ambato fo.fernandez@uta.edu.ec
[3] Universidad Central "Marta Abreu" de las Villas amed@uclv.edu.cu

**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 AUCTORITAS' web services, its Ontology-based Data Access model and how the semantic web languages make possible the semantic integration of heterogeneous data sources.

**Keywords:** Authority Control, Linked Open Data, Ontology-based Data Access, OBDA,Semantic Web

## 1 Introduction

Authority Control is the most expensive part of the cataloging process [28,8,29], it is a global problem, affecting not only libraries but organizations of all kinds [19]. 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. 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]. With the aim of adding semantics to describe the contents, ontologies are used in the Semantic Web. An ontology is an explicit specification of a conceptualization [11]. In Ontology-based Data Access (OBDA) an ontology can be used in order to create a semantic domain layer used in a software application [6]. Nowadays there are huge volumes of data stored in heterogeneous data sources that could be combined and reused by using the OBDA paradigm. Regarding to heterogeneity this paper refers to the different ways used to store and retrieve data.

The current work aims to present AUCTORITAS, a Semantic Web-based tool for Authority Control. This work is structured as follows: A section about the main related works covering Authority Control, the Semantic Web, Linked

Open Data, OpenLink Virtuoso, VIVO, AUCTORITAS and the Ontology-based Data Access paradigm. Next the ontological model supporting AUCTORITAS data access is described. After that the experimental evaluation of the proposal is explained, followed by the main conclusions and future work.

## 2 Related Work

### 2.1 Authority Control

Authority control is a matter that has demanded 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 [19].

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 [19]. In Asia, the Hong Kong Chinese Authority Name (HKCAN) is established. This meant recognition of the issue in just two organizations worldwide - far [19], however, from the syndetic goals set forth by Charles Cutter in the nineteenth century [7].
- Lubetzky [21] 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 [5] 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.
- ORCID organization [25] provides a persistent digital identifier that distinguishes researchers and organizations between them.
- Thomson Reuters created ResearcherID [27]. Each ResearcherID's member is assigned a unique identifier to enable researchers to avoid author misidentification.

The Online Computer Library Center (OCLC), IFLA and LOC have fueled initiatives for authority control by sharing the records of various cataloguing agencies [19]. Result 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 [4].

Authority control also includes the management of subject headings. In that way the LOC shares its subject headings [20], organizations like Food and Agriculture Organization (FAO) shares their thesauri with the aim that libraries can reuse them. Software tools like SKOSMOS [26] have been developed in order to make the thesauri available online, but SKOSMOS only provides subject headings-related authority control, so its scope does not cover the whole authority control spectrum.

### 2.2 Semantic Web

Since Resource Description Framework (RDF) made it possible to define the meaning of data in a machine readable form [23], 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 Logics [15]. OWL is a formal language for representing ontologies in the Semantic Web [15]. This language has been used in many specific scenarios for the construction of flexible data semantic models [12,16,17,10]. Several knowledge organization systems takes advantage of semantic web technologies [22,14,9], SKOS [22] is one of them. In this proposal we reuse SKOS structured information sources provided by institutions and reuse their data.

### 2.3 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 [18].

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 [13].

### 2.4 Virtuoso Open Source

Virtuoso Open Source[4] 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.

[4] http://virtuoso.openlinksw.com/

- 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[5] 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 also be managed by Virtuoso.

### 2.5 VIVO

VIVO[6] 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. VIVO provides integration with ORCID, so ORCID identifiers can be linked to authors and organizations profiles in VIVO.

All the information managed by VIVO is structured as Linked Open Data, this structure improves information discovery [18] 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. In Cuban context, VIVO is intended to be used for creating a national researchers' directory, through which the Cuban scientific production can be exposed. VIVO is used as an external application which is queried by AUCTORITAS in order to retrieve the author's identifiers coming from their profiles.

### 2.6 Ontology-based Data Access paradigm

Ontology-based Data Access is a paradigm of accessing data through a conceptual layer [1]. Usually, the conceptual layer is expressed in the form of a RDF(S) or OWL ontology. Terms in the conceptual layer are mapped to values in the data layer. This is achieved by specifying each proper query that allows to retrieve actual data from data sources [1]. Formally, an OBDA system is a triple $\Omega =< \tau, \sigma, \mu >$ where:

- $\tau$ is the intensional level of an ontology. We consider ontologies formalized in description logics (DLs), hence $\tau$ is a DL TBox.

[5] http://aims.fao.org/es/agrovoc

[6] http://vivoweb.org/

- $\sigma$ is a data sources set.
- $\mu$ is a set of mapping assertions, each one of the form $\Phi(x) \leftarrow \Psi(x)$ where
    - $\Phi(x)$ is a query over $\sigma$, returning tuples of values for $x$.
    - $\Psi(x)$ is a query over $\tau$ whose free variables are from $x$.

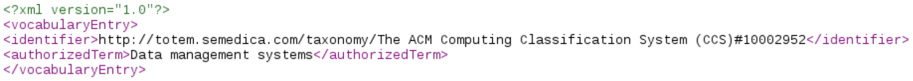
The OBDA paradigm has been used in software applications like the *Ontop* framework [1] for retrieving data stored in relational databases. More recently OBDA has been extended to NoSQL databases such as MongoDB [3], and in the current paper it is been used also for accessing RDF-based data sets and external applications that exposes their data through REST-based web services. The usage of the OBDA paradigm allows the applications to scale respecting to data sources. When there are changes in the data sources, the only component that needs to be modified is the assertional part of the ontology.

## 3 AUCTORITAS interface

AUCTORITAS interface is the main entry point for our applications ecosystem, it 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.

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 1 shows AUCTORITAS answer to an external system after searching for "database" term on the ACM Controlled Vocabulary.

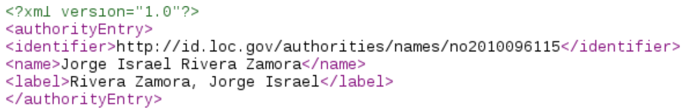


```
<?xml version="1.0"?>
<vocabularyEntry>
<identifier>http://totem.semedica.com/taxonomy/The ACM Computing Classification System (CCS)#10002952</identifier>
<authorizedTerm>Data management systems</authorizedTerm>
</vocabularyEntry>
```

**Fig. 1.** AUCTORITAS answer to a query over 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.

Also external applications may query AUCTORITAS services for personal author entries. Figure 2 shows AUCTORITAS answer to a query about Jorge Israel Rivera Zamora over LOC's graph processed information.



```
<?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. 2.** AUCTORITAS answer to a query about Jorge Israel Rivera Zamora

## 4 Ontology Model for Accessing Data

In order to provide a conceptual layer to AUCTORITAS for OBDA an ontology was designed. Figure 3 depicts the classes, object properties and data properties used in the designed ontology.

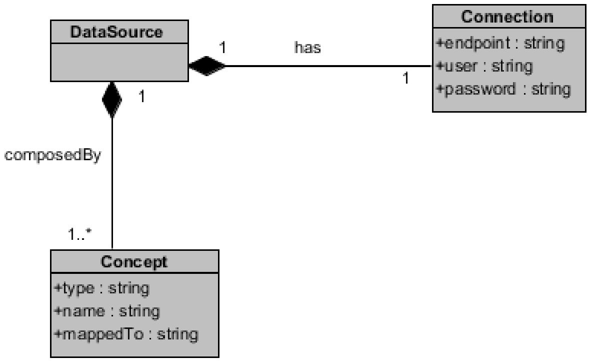


**Fig. 3.** Class diagram containing the classes, object and data properties of the designed ontology

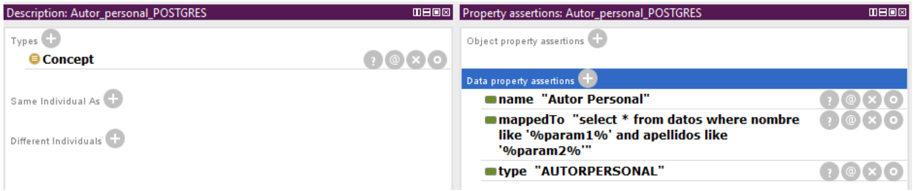
Each *DataSource* class instance is an identifier (URI) representing a data source consumed by AUCTORITAS. The only requirement that a data source must meet in order to use it in AUCTORITAS is that its data can be retrieved by a syntactical query. The *DataSource* instance is related with a *Connection*'s class instance by the object property *"has"*. The *Connection*'s class instance is integrated by the following data properties:

- *endpoint*: A string representing the path of the data source where it is listening for queries.
- *user*: A string representing a user needed for authentication purposes when running the query. It is optional.
- *password*: A string representing a password needed for authentication purposes when running the query. It is optional.

Each *DataSource* class instance is related with the *Concept*'s class instances by the object property *"composedBy"*. *Concept*'s class instances are abstract representations of the data stored in the data source. Each *Concept*'s instance is integrated by the following data properties:

- *type*: A string discriminating what the concept is about. The value can be only one of the following strings "AUTORPERSONAL", "AUTORCORPORATIVO", "CONTROLEDTERMS".
- *name*: A string representing the name of the concept in natural language.
- *mappedTo*: A string representing a syntactical query expressed in a query language (e.g. SQL, SPARQL).

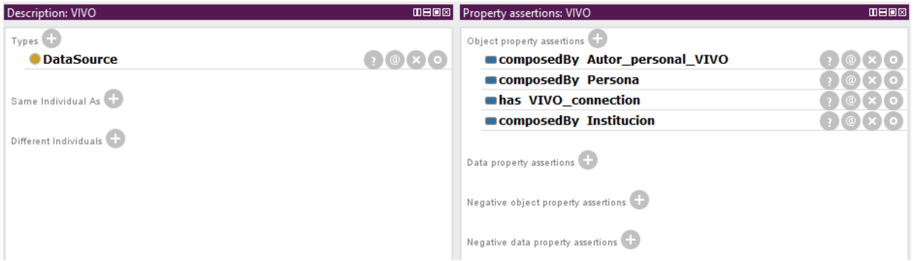
Figures 4, 5 and 6 depict the assertional part of the designed ontology in Protégé. A convention was used for naming parameters: the string "param" followed by a number. Those parameters are replaced inside AUCTORITAS by the values passed by external applications.


Description: Autor_personal_POSTGRES
Types
Same Individual As
Different Individuals
Property assertions: Autor_personal_POSTGRES
Object property assertions
Data property assertions

Concept

name "Autor Personal"
mappedTo "select * from datos where nombre like '%param1%' and apellidos like '%param2%'"
type "AUTORPERSONAL"

**Fig. 4.** Ontological description of the "Personal Author" concept


Description: VIVO
Types
Same Individual As
Different Individuals
Property assertions: VIVO
Object property assertions
Data property assertions
Negative object property assertions
Negative data property assertions

DataSource

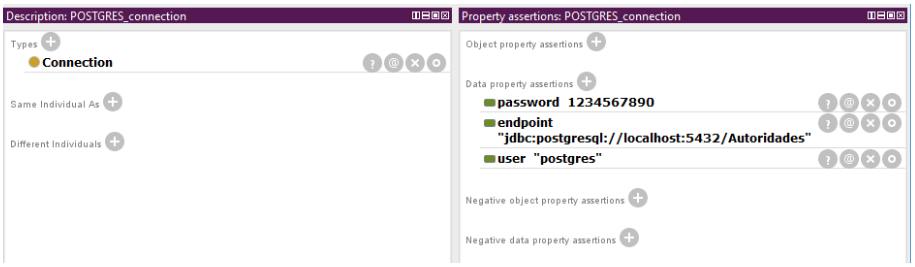
composedBy Autor_personal_VIVO
composedBy Persona
has VIVO_connection
composedBy Institucion

**Fig. 5.** Ontological description of VIVO data source

When an external application requests AUCTORITAS' web services, AUCTORITAS uses its OBDA mechanism to fulfill the request as depicted in figure 7.

## 5 Evaluation

In order to evaluate the proposal, an experimental environment was set. The experiment was designed as follows.



Description: POSTGRES_connection

Types

Connection

Same Individual As

Different Individuals

Property assertions: POSTGRES_connection

Object property assertions

Data property assertions

password 1234567890

endpoint
"jdbc:postgresql://localhost:5432/Autoridades"

user "postgres"

Negative object property assertions

Negative data property assertions

**Fig. 6.** Ontological description of the connection to PostgreSQL server

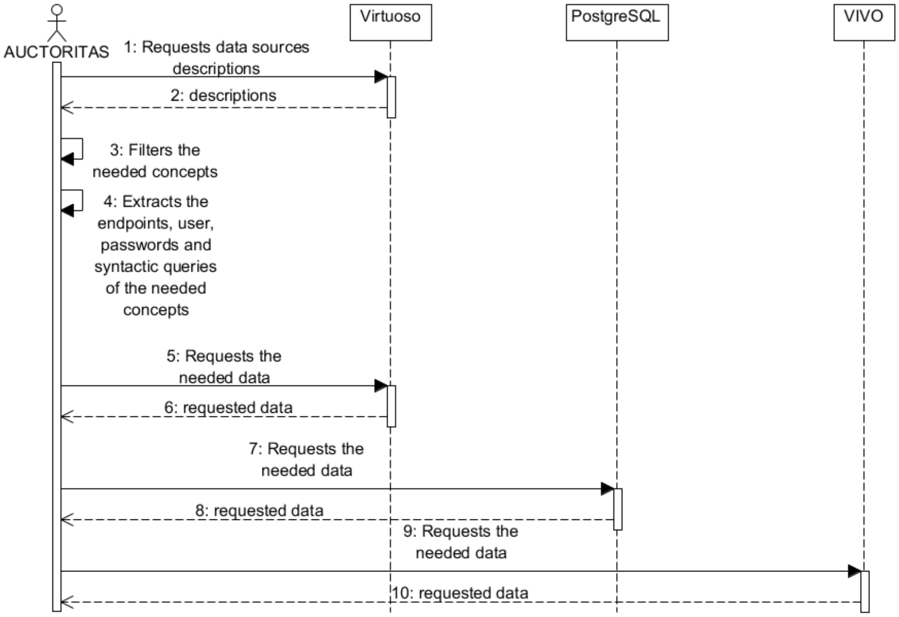

AUCTORITAS
Virtuoso
PostgreSQL
VIVO
1: Requests data sources descriptions
2: descriptions
3: Filters the needed concepts
4: Extracts the endpoints, user, passwords and syntactic queries of the needed concepts
5: Requests the needed data
6: requested data
7: Requests the needed data
8: requested data
9: Requests the needed data
10: requested data

**Fig. 7.** AUCTORITAS OBDA mechanism workflow.

**Goal:** To evaluate AUCTORITAS web services working over an OBDA mechanism according to integrity and performance as recommended by Mustafa [24].

**Participants:** Regarding integrity sixty rounds of experimentation were conducted in order to verify the accurate connection between AUCTORITAS and the described data sources, twenty per data source. Respecting to performance ten rounds of experimentation over each web service were conducted per each increase in the requests amount, being one hundred and sixty performance evaluation attempts.

**Research question:** Is AUCTORITAS OBDA mechanism able to perform in the concurrent expected scenarios?

**Experiment materials:** A computer with an Intel Core-i5 2450 processor at 2.5 GHz, 8 Gb of RAM and hard disk drives at 5400 RPM was used to serve the application and the data sources. The operating system of that computer was OpenSuse 42.1, the relational database server was PostgreSQL version 9.4.1, the RDF data storage was Virtuoso Open Source version 7.2.1, the web server was Oracle Glassfish version 4.1.1 and VIVO version 1.8 as data source. In the client side a computer with an Intel Core-i5 2410 processor at 2.3 GHz, 8 Gb of RAM and hard disk drives at 5400 RPM was used. The operating system of that computer was Microsoft Windows 10 x64 and to simulate the concurrency conditions SoapUI version 4.6.1 and LoadUI version 2.6.5 were used. As web browser Mozilla Firefox version 46.0 was utilized. The connection between AUCTORITAS, its data sources and the client computer is depicted in figure 8.

**Tasks:** For the integrity evaluation the users made sixty requests to AUCTORITAS web services through a web browser. By obtaining a successful answer with the requested data was stated that the connection between AUCTORITAS and the corresponding data sources was successful. For the performance evaluation ten rounds of experimentation were conducted per each increase in the requests amount. The requests amounts were five, ten, fifteen and twenty per second. More than those amounts are not probable in the context where the application will be deployed, this is due to the application is not a critical mission one and only is requested in some specific parts of the cataloging process. The duration time of each round was ten seconds.

**Hypothesis:** AUCTORITAS with an OBDA mechanism will be able to successfully perform in the expected concurrency scenarios.

**Variable:** Number of successfully completed requests.

### 5.1 Experiment results

In each iteration of the integrity experiment the user made a request through the web browser and checked the answer. After sixty successfully requests, it was concluded that the connection between AUCTORITAS and its data sources was performing well.

For the performance experiment ten rounds for each increase in the requests concurrency were carried out per web service. The averages of successfully completed requests and standard deviation were tabulated in table 1.

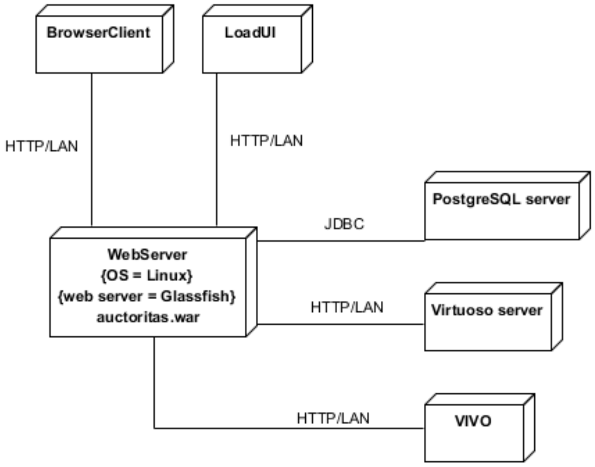


**Fig. 8.** Deployment diagram for the integrity evaluation

| Quantity of requests per second | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | | 10 | | 15 | | 20 | |
| Average | $\sigma$ | Average | $\sigma$ | Average | $\sigma$ | Average | $\sigma$ |
| Personal authorities web service | | | | | | | |
| 49.1 | 0.3 | 99.1 | 0.3 | 149.1 | 0.3 | 199 | 0 |
| Corporate authorities web service | | | | | | | |
| 49.2 | 0.4 | 99.3 | 0.9 | 149.1 | 0.3 | 199.4 | 0.49 |
| Controlled vocabularies web service | | | | | | | |
| 49.5 | 0.5 | 99.4 | 0.49 | 149.1 | 0.3 | 199.5 | 0.5 |
| Controlled term web service | | | | | | | |
| 49.2 | 0.4 | 99.4 | 0.49 | 149.2 | 0.4 | 199.1 | 0.3 |

**Table 1.** Measurement of the completed requests during the experiment

## 6 Conclusions and Future Work

The development of authority control faces new challenges in the Semantic Web. The need for increasing interoperability capabilities between software applications and information stored in heterogeneous structures is a promising area. Designers and developers of future cataloging and authority control systems should use the benefits of semantic web technologies to improve interoperability.

The usage of ontology-based data access mechanisms provides better scalability to applications in order to plug in new data sources for consuming data. In the context where AUCTORITAS will be deployed the designed OBDA mechanism acceptably performs to provide external applications authority control features.

## References

1. Bagosi, T., Calvanese, D., Hardi, J., Komla-Ebri, S., Lanti, D., Rezk, M., Rodríguez-Muro, M., Slusnys, M., Xiao, G.: The ontop framework for ontology based data access. Communications in Computer and Information Science 480, 67–77 (2014)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American magazine 284(5), 34–43 (2001)
3. Botoeva, E., Calvanese, D., Cogrel, B., Rezk, M., Xiao, G.: OBDA Beyond Relational DBs : A Study for MongoDB. In: Lenzerini, M., Peñaloza, R. (eds.) International Workshop on Description Logics. CEUR Workshop Proceedings, Cape Town,South Africa (2016), http://ceur-ws.org/Vol-1577/paper_40.pdf
4. Bourdon, F., Zillhardt, S.: Author: Vers une base européenne de notices d'autorité auteurs. International cataloguing and bibliographic control 26(2), 34–37 (1997)
5. 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)
6. Calvanese, D., Liuzzo, P., Mosca, A., Remesal, J., Rezk, M., Rull, G.: Ontology-based data integration in EPNet: Production and distribution of food during the Roman Empire. Engineering Applications of Artificial Intelligence pp. 1–18 (2016), http://linkinghub.elsevier.com/retrieve/pii/S0952197616000099
7. Cutter, C.A.: Rules for a printed dictionary catalogue. US Government Printing Office (1889)
8. 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)
9. Dunsire, G., Willer, M.: Standard library metadata models and structures for the semantic web. Library hi tech news 28(3), 1–12 (2011)
10. Fernández-Peña, F., Urrutia-Urrutia, P., Cañete, R., Acosta-Sánchez, R., Yañez-Márquez, C., Nummenmaa, J.: A conceptual data model for the automatic generation of data views. Applied Mathematics & Information Sciences (2016)
11. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.7493

12. 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)
13. 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)
14. Hodge, G.: Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. ERIC (2000)
15. 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)
16. K. C̆erāns, 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)
17. 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)
18. 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
19. 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
20. Library Of Congress: Library of Congress Subject Headings (2016), http://id.loc.gov/download/
21. Lubetzky, S., Hayes, R.M.: The Principles of Cataloging: Report. Institute of Library Research, University of California (1969)
22. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference. W3C recommendation 18, W3C (2009)
23. 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)
24. Mustafa, A.S., Kumaraswamy, Y.S.: Performance Evaluation of Web-Services Classification. Indian Journal of Science and Technology 7(October), 1674–1681 (2014)
25. ORCID: ORCID Connecting Research and Researchers (2016), http://orcid.org/
26. Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., Retterath, A.: Publishing SKOS vocabularies with Skosmos. Tech. rep. (2015), http://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf
27. Thomson Reuters: ResearcherID (2016), http://www.researcherid.com
28. 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
29. West, W.L., Miller, H.S., Wilson, K.: Electronic journals: Cataloging and management practices in academic libraries. Serials Review 37(4), 267–274 (2011)