ISSN 1870-4069

Ontology-based Population and Enrichment of Researcher Profiles

Isabel Cruz-Ruiz, Maricela Bravo, José A. Reyes-Ortiz

Autonomous Metropolitan University, Systems Department,
Ciudad de México, Mexico

icruzr2014@hotmail.com, {mcbc,jaro}@azc.uam.mx

Abstract. The representation, management, and exploitation of researcher profiles is an important task that every research institution must achieve. In this paper, we investigate on the use of ontologies as the main solution approach to support the representation of researcher profiles in a given academic environment. We describe the ontology model design, the automatic ontology population processes, and the discovery and enrichment processes of interesting semantic relations between researcher profiles. The functional competency of the enriched ontology is evaluated utilizing a set of inference rules and queries.

Keywords. Ontology population, ontology enrichment, researcher profile.

1 Introduction

Currently, higher education institutions and research institutes have highly specialized human resources who count with high degrees of postgraduate studies. The capacity, expertise and talent accumulated by academic and research staff is one of the most important assets available to institutions. Representing, quantifying and knowing how to better manage these highly specialized human resources is a very important issue; however, it is not an easy task to perform, since it requires the acquisition, representation and intelligent treatment of large volumes of data. A good management of highly specialized human resources can be carried out through the administration of researcher profiles to enable: finding similar profiles to establish new collaborations, looking for specific profiles that allow to integrate a work team with specialists, discovering groups or classes of researchers that address similar topics, discovering groups of researchers that address different problems but that use similar approaches, among other possible applications.

A researcher profile consists of the relevant information regarding previous academic work experience in different research institutions, education and level of studies considering undergraduate, graduate, and specialization studies; an important aspect of a researcher profile is the scientific published articles, chapters, and books, as they represent the researcher topics of interest, and the researcher most active collaborations.

In this paper, we present an ontology solution approach for the acquisition, representation and management of researcher profiles. This ontology solution is evaluated using a set of competency questions through which the functional competency of the proposed solution is evaluated satisfactorily.

This paper reports a contribution in the area of ontology learning. Maedche and Staab [1] define *ontology learning* as the process of automatic or semi-automatic construction, enrichment and adaptation of ontologies. Accordingly, the main tasks involved in ontology learning are ontology enrichment, inconsistency resolution and ontology population.

a) *Ontology enrichment* is the task of extending an existing ontology with additional concepts and semantic relations in the ontology.
b) *Inconsistency resolution* is the task of resolving inconsistencies that appear in an ontology aiming at producing and maintaining a consistent ontology.
c) *Ontology population*, is the task of adding new instances of concepts to the ontology.

The methodological process followed for the construction and evaluation of the proposed solution consist of four phases: ontology design, ontology population, ontology enrichment, and ontology evaluation.

The rest of this paper is organized as follows. Section 2 presents the related work, which is briefly described to compare them with the approach presented in this paper. Section 3 describes the specification of ontology requirements. Section 4 describes the ontology design methods. Section 5 presents the automatic ontology population of each ontology. Section 6 describes the ontology research profile enrichment. Section 7 presents an evaluation based on competency questions. Finally, Section 8 shows the conclusions and future work.

2 Related Work

In this section, we first present the definition of ontology, and describe the related works that address the representation and management of researcher or academic profiles. We analyze their applications and concept coverage.

Over the last decades, different ontology definitions have been presented and discussed. According to Gruber [2] an ontology is an "explicit specification of a conceptualization", an ontology is used to formally define the important concepts of a terminology and the semantic relationships that may exist between concepts. It is frequent that the set of formally defined concepts belong to a specific area of knowledge, and the set of rules and axioms defined are congruent with the particular area of knowledge. In [3] Sowa stated that an ontology represents a catalog of categories to classify entity types that exist in a given domain. In [4] Cámara explained that an ontology can be conceived as an instrument for knowledge representation in a particular topic area, through which knowledge recovery and information retrieval can be executed. Ontologies were selected as the formal representational mechanism as they facilitate reusability, knowledge sharing, and

execution of formal reasoning tasks such as satisfiability of concepts, consistency checking, classification and inference.

Concerning researcher profile, Yao, Tang and Li [5] address the problem of researcher profiling by annotating a collection of researcher web pages, and defined a series of difficulties found using this approach. Authors identify tokens in the Web page heuristically, assign tags to each token (Position, Affiliation, Email, Address, Phone, Fax), using the tags, they perform the profiling extraction. In Liu et al [6] authors address the problem of finding experts with required expertise. They describe two ontologies: an *expert ontology*, which defines concepts such as: Person, Publication, Project, and Research Interests; and a *domain ontology* which stores the key concepts (research areas), the attributes of the concepts and the relations between concepts (for example, broader, narrower and part-of). In [7] authors address the problem of automatic extraction of topics of expertise of a person based on the documents accessed by the person through information extraction techniques. They define a user profile using a set of topics with weights determining his level of interest. In [8] authors present a multi-agent paradigm supported by a semantic web architecture to address the challenges of researcher profiling and association. Authors describe an ontological model to represent information such as researcher profiles, conference papers, research centers, etc.

In [9] authors describe ArnetMiner, to address the following questions:

a) How to automatically extract researcher profiles from the Web?
b) How to integrate the extracted information (e.g., researchers' profiles and publications) from different sources?
c) How to model different types of information in a unified approach?
d) How to provide powerful search services based on the constructed network?

In ArnetMiner the schema of a researcher profile was proposed consisting in two main entities: Researcher and Publication. Based on the work reported in [9] in this project we address the same questions and present a solution approach based on the use of ontologies and reasoning tasks.

In [10] authors describe a skill classification ontology model containing skills of research in the area of computer and information science. Their main contributions are:

a) A process to build the skill classification ontology.
b) A methodology to determine expertise of the researcher using the skill classification ontology.
c) A method to retrieve the relevant researchers who may have competency matched to the desired expertise.

Motivated by these related works, we propose an ontology-based solution approach for the acquisition, representation and management of researcher profiles.

3 Specification of Ontology Requirements

The main objective of the ontology model presented in this paper is to facilitate researcher profile processing and reasoning. Considering that every research

institution requires the efficient management and dissemination of information relative to the professors and research activities, etc. The ultimate goal of this project is the smart and provisioning of services to researcher communities in which researchers search for specialized publications (such as publications, coauthors, conferences) and are also interested in establishing collaborations with other researchers. Considering this motivation, the following requirements were defined in order to guide the ontology design, construction and evaluation.

3.1 Scope of the Ontology

In order to specify the scope of the ontology, we reviewed the concepts of research profile. Yao, Tang and Li [5] described profiling as the process of obtaining the values associated with the different properties that constitute the person model. Authors define the schema of a researcher profile containing: name, affiliation, position, phone, address, email, research interests, and postgraduate studies. From this definition we consider that the entities (or objects) that constitute a researcher profile are: ***Person*** (for example professors, students, staff, etc.), and ***Publication*** (to extract research interests).

From this initial analysis, we have defined the concept coverage requirements of the ontology and defined the main objective of the ontology, which is to facilitate researcher profile representation and processing in the academic environment.

3.2 Concept coverage

The ontology model should include the following concepts:

a) Data for the identification of persons and researchers such as name, economical number, staff card, etc.
b) Person profile information to represent the user data that is possible to gather from public networks, public Web pages, or public data bases available such as DBLP.
c) Data to represent publications such as: thesis, chapters, journals, etc.

3.3 Competency of the Ontology

Gruninger and Fox [11] proposed six characteristics to evaluate a Business Model. These characteristics were proposed to answer the question of "How can one determine which model is correct for a given task?" To give a guideline on the operation of these characteristics, the authors define the concept of competence of the model as follows: given an appropriately instantiated model and a demonstrator of theorems, the competence of a model is the set of questions that the model can answer. Based on this definition, we may state that

The competence of an ontology model is the set of questions that the ontology can answer.

Evaluation of the competency of an ontology system is crucial to verify that a representational model is complete with respect to a given set of competency questions. During the phase of requirements specification, a team of experts and programmers defined the following set of competency questions for this model:

1. To know how many scientific works does a given researcher has published
2. To find groups of authorship collaborations
3. To know the researcher´s publications from a specialized topic with high degree studies
4. To answer about statistical data of publications
5. To know the scientific productivity of a given department
6. To know the number of female researchers from a given department with published scientific works.

4 Ontology Design

Ontology design is the process of selecting and applying methods, techniques and principles with the objective of producing an ontology model. In this section, the design considerations that were taken into account are described.

A good quality ontology design depends mainly on the selection and incorporation of design principles. Uschold and Grüninger [12] presented their initial ideas and detailed a set of ontology design criteria. For the design and construction of the researcher profile ontology the following design principles were taken in consideration:

i. *Clarity* principle states that an ontology should effectively communicate the intended distinctions. Ambiguity should be minimized, distinctions should be motivated, and examples should be given to understand definitions that lack necessary and sufficient conditions.

ii. *Coherence* design principle specifies that an ontology should be internally consistent. Coherence should also apply to the parts of the definitions that are not axiomatic.

iii. *Extensibility* principle states that an ontology should be designed anticipating possible uses of the shared vocabulary.

Additionally, an initial set of competency questions were used for term elicitation and for final competency evaluation.

4.1 Person Ontology

Person ontology was designed to represent all possible academics, which hold a permanent or temporal position as professor or researcher at the university, such as: academic visitor, full time professor, external sabbaticals, etc. This ontology also represents postgraduate students, and research oriented undergraduate students, among others. Figure 1 shows the main class hierarchy of the ***Person*** ontology. An

important characteristic of this ontology is that it uses a unique identifier for every type of person.

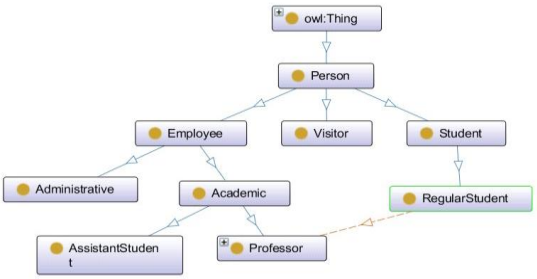


Fig. 1. Person ontology class hierarchy.

The concept ***Person*** is defined as an equivalence through the *hasName* and *hasGender* data properties, indicating that every person individual is obligated to have name and gender to be classified as type of ***Person*** class. The concept ***Employee*** is defined as a sub class of a ***Person*** that *hasEconomicNumber* data property. Whereas the concept ***Student*** is defined as a sub class of ***Person*** that *hasStudentId*. An important concept is a ***Professor*** which is an ***Academic***, is an ***Employee*** and is a ***Person*** that *hasCategory*, *hasDepartment*, and *hasEmail*; and inherits the data property of an ***Academic*** *hasProject*. The class hierarchy of the ***Person*** ontology shows the sub-classification of the class ***Student*** into ***RegularStudent*** and ***AssistantStudent***. This classification addresses the particular need to represent the two types of students that exist in the university where an individual of the ***AssistantStudent*** class is considered to be an ***Academic***, an ***Employee*** and a ***Student***.

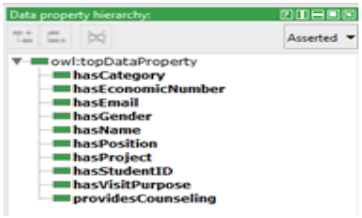


Fig. 2. Data type properties defined for the Person ontology.

The full list of data type properties defined for the ***Person*** ontology are shown in Figure 2.

4.2 Publication Ontology

Scientific published articles, chapters, and books are the most important sources of information in order to integrate a researcher profile. Scientific publications contain the author's topics of interest, conferences and journals of preference, the years of publications and periodicity; also the researcher most active collaborations. In order

to build a researcher profile, the design and construction of a publication ontology considered as input the information extracted from the DBLP computer science bibliography on-line reference, extracting the most relevant bibliographic information on major computer science publications.

The ***Publication*** ontology defines the same attributes utilized in DBLP. Figure 3, shows those data type attributes.

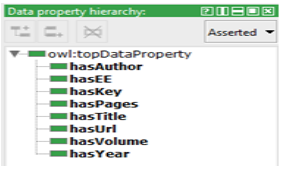


Fig. 3. Data properties of the Publication ontology.

4.3 Researcher Profile Ontology

Researcher Profile ontology was designed to incorporate conceptualizations from ***Person*** and ***Publication*** ontologies. From ***Person*** ontology imports ***Professor*** personal data, such as full name, and economical number; from the ***Publication*** ontology imports publications organized by year, type of publication, among others. Additionally, incorporates ***Department*** and ***Academic Title*** concepts. All these conceptualizations are used to complete the definition of a ***Researcher Profile***, considering the associated publications, the affiliated department, the academic title obtained, and the rest of personal data. Figure 4 shows the main concepts that integrate the ***Researcher Profile*** Ontology.

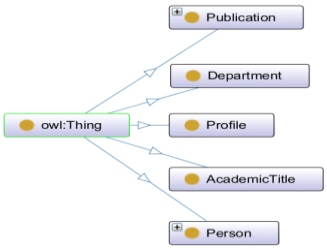


Fig. 4. Researcher Profile ontology class hierarchy.

5 Ontology Population

Ontology population is the process of adding (instantiating) new individuals in the ontology concepts (classes). Automated ontology population is desirable due to the large amount of data that must be extracted and instantiated in the ***Person*** and ***Publication*** ontologies.

5.1 Person Ontology Population

For ***Person*** ontology population, the data source comes from a set of excel files that the management staff of the university uses for different purposes. These excel files contain the information of all academic staff who are affiliated with the university, such as: professor's full name, gender, department, email, economical number, academic projects, and alias.

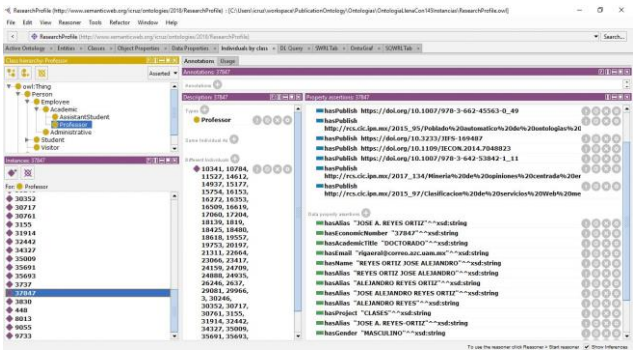
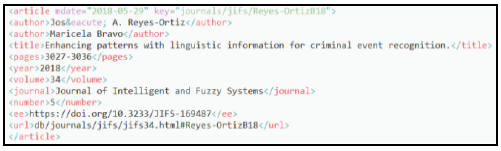


Fig. 5. Person ontology population.

For the automated population of the ***Person*** ontology, two Java modules were developed: a module to *parse and extract* the information from the source files; and another module to *interact* with the ontologies using the Java OWL Application Programming Interface (OWL API) to load and manipulate ontologies, creating new individuals, instantiating object properties and data properties with individuals, and register them in the ontology. Figure 5 shows the values of data type attributes registered for professor "REYES ORTIZ JOSE ALEJANDRO", and the recognized alias names. Alias names are important in order to facilitate the semantic association of the researcher individual with all his publications.

5.2 Publication Ontology Population

For ***Publication*** ontology population, the data was extracted from the DBLP (Digital Bibliography & Library Project) [13], a compressed XML file, which contains more than a million of Computer Science publications. The XML file from the DBLP contains publication title, author names, publication year, volume, EE (a unique publication identifier), URL, and pages (see Figure 6). However, it does not provide the abstract and keywords of publications.



```
<article mdate="2018-05-29" key="journals/jifs/Reyes-OrtizB18">
<author>Jos&eacute; A. Reyes-Ortiz</author>
<author>Maricela Bravo</author>
<title>Enhancing patterns with linguistic information for criminal event recognition.</title>
<pages>3027-3036</pages>
<year>2018</year>
<volume>34</volume>
<journal>Journal of Intelligent and Fuzzy Systems</journal>
<number>5</number>
<ee>https://doi.org/10.3233/JIFS-169487</ee>
<url>db/journals/jifs/jifs34.html#Reyes-OrtizB18</url>
</article>
```

Fig. 6. DBLP XML file extract.

A Java module was built to *interact* with the ontologies using the OWL API to load and manipulate ontologies, creating new individuals, instantiating object properties and data properties. Figure 7 shows the instantiation of new ***Publication*** individuals correlated with the identification of the author that published.

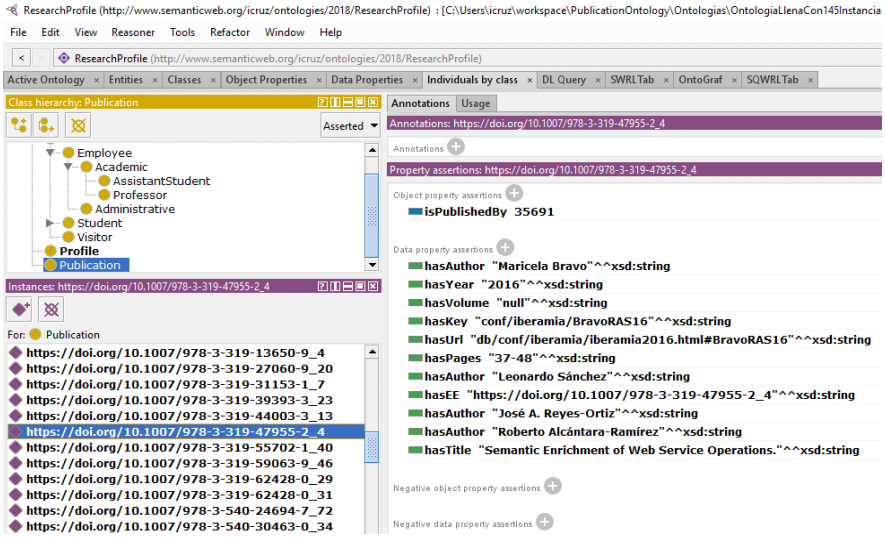


Fig. 7. Publication ontology population.

6 Ontology Enrichment

Ontology enrichment is the automatic process of analyzing the population data values and discovering new interesting semantic relations between individuals. Of particular interest in this enrichment process is the automatic discovery of collaboration relations between authors of publications. For this, the following object properties and inference rules were defined.

collaborateWith is an object property with domain ***Person*** and range ***Person***. This object property was defined to establish semantic relationships between authors of publications.

Figure 8 shows the SWRL rule that was executed to find collaborations between authors of publications is:

```
Publication(?pub) ^ isPublishedBy(?pub, ?prof) ^
isPublishedBy(?pub, ?prof2) ^ differentFrom(?prof, ?prof2) ^
-> collaborateWith(?prof, ?prof2)
```

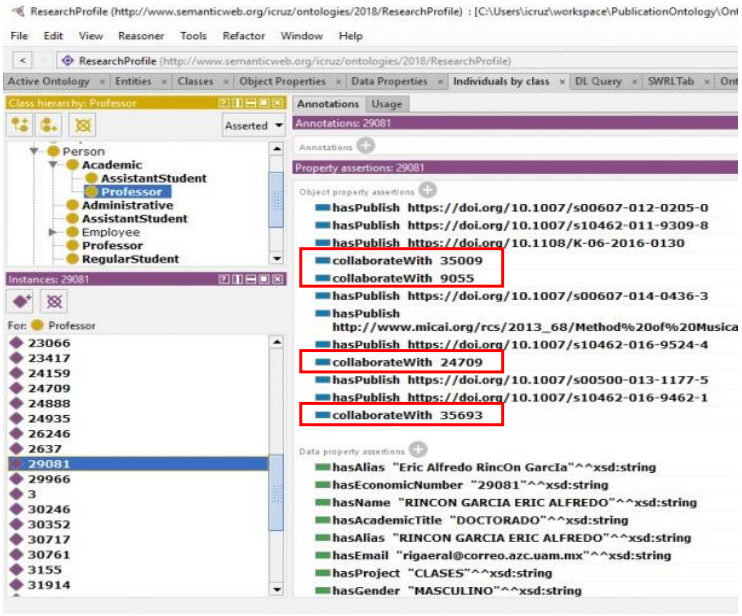


Fig. 8. Discovery of authorship collaborations.

7 Ontology Evaluation

Ontology Evaluation [14] concerns the correct building of the ontology, ensuring that its definitions correctly implement the ontology requirements and competency questions. For evaluation the ***competence of the ontology*** was considered, that is, if it is able to respond to a set of competency questions; and the verification of ***requirements compliance***. The following competency questions were coded in SWRL language and their results were correct.

7.1 Researcher Publications

To know how many scientific works does a given researcher has published, the following rule was defined and executed. Figure 9 shows the result of this competency question.

```
Professor(?prof1) ^ hasPublish(?prof1, ?pub) ^
hasEconomicNumber(?prof1, ?e) ^ swrlb:equal(?e, "14233") ->
sqwrl:count(?pub)
```

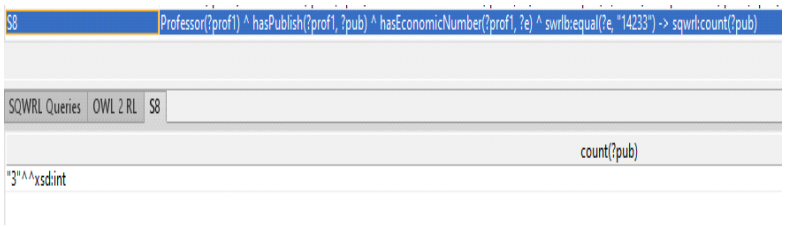


Fig. 9. Discovery of authorship collaborations.

7.2 Collaboration between Researchers

To find groups of authorship collaborations, the following semantic Web rule is defined. The result is shown in Figure 10.

```
Professor(?p1) ^ collaborateWith(?p1, ?p2) ^
hasEconomicNumber(?p1, ?e) ^ swrlb:equal(?e, "14233") ->
sqwrl:select(?p1, ?p2)
```

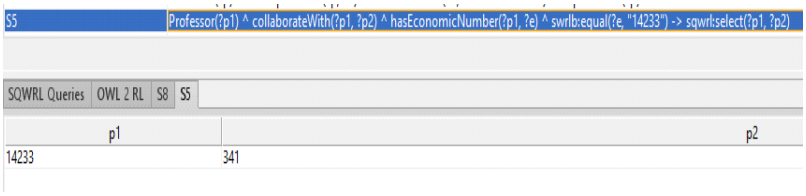


Fig. 10. Discovery of authorship collaborations.

7.3 Qualified and Specialized Researchers

In order to know the researcher´s publications from a specialized topic with high degree studies, the following semantic Web rule was utilized. Figure 11 shows the result of the execution.

```
Professor(?prof1) ^ hasDepartment(?prof1, ?dep) ^ swrlb:contains(?dep,
  "ELECTRONICA") ^ hasAcademicTitle(?prof1, ?at) ^ swrlb:equal(?at,
    "DOCTORADO") ^ hasPublish(?prof1, ?pub) -> sqwrl:count(?pub)
```

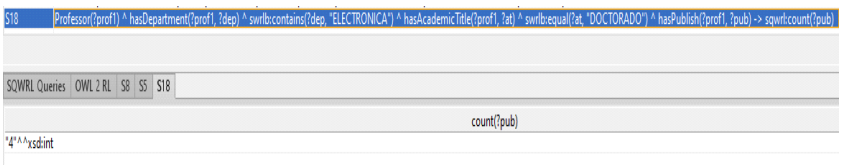


Fig. 11. Highly specialized researchers.

7.4 Publications by Year

The ontology is capable of answering statistical data of publications, for instance: How many publications were there in the year 2017? Figure 12 shows the result of this rule:

```
Publication(?pub) ^
hasYear(?pub, ?y) ^
swrlb:equal(?y, "2017") -> sqwrl:count(?pub)
```

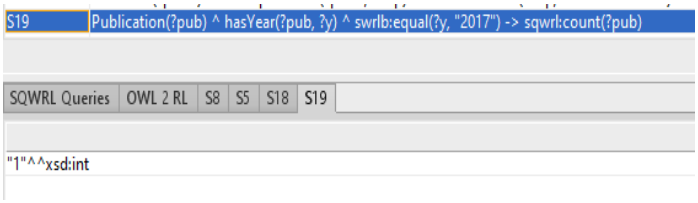


Fig. 12. Publications by year.

7.5 Publications by Department

In order to know the scientific productivity of a given department, the following semantic Web rule is used. Figure 13 shows the result.

```
Professor(?prof1) ^ hasDepartment(?prof1, ?dep) ^ swrlb:contains(?dep,
    "SISTEMAS") ^ hasPublish(?prof1, ?pub) -> sqwrl:count(?pub)
```

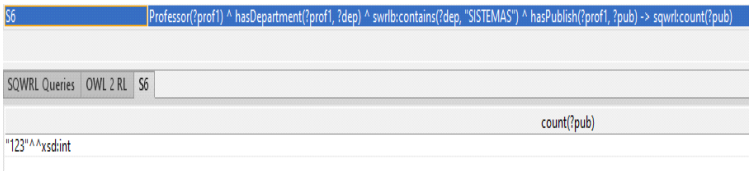


Fig. 13. Publications by department.

7.6 Publications by Gender

To know the number of female researchers from a given department that have published scientific works, the following semantic Web rule was used. Figure 14 shows the results.

```
Professor(?prof) ^
hasGender(?prof, ?gen) ^
swrlb:equal(?gen, "FEMENINO") ^ hasDepartment(?prof, ?dep) ^
swrlb:equal(?dep, "SISTEMAS") ^ hasPublish(?prof, ?pub)
-> sqwrl:count(?prof)
```

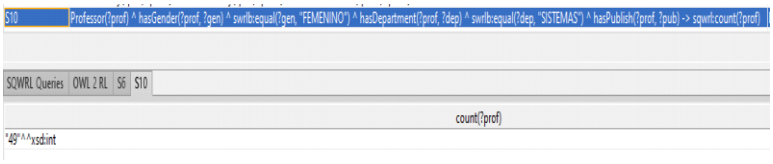


Fig. 14. Publications by gender.

8 Conclusions

In the work reported in this paper, an automated ontology population was used to build researcher profiles. For the population of the ***Person*** ontology, a collection of 373 professors was used, two departments were instantiated in the ***Department*** class, 50 professor individuals are from the Systems ***Department*** and 60 from Electronics. 100% of them were correctly inserted in the ***Person*** ontology, specifically in the ***Professor*** class.

For experimentation purposes two ontologies were generated: one was used for professors of the electronics department and another for professors of the systems department. The systems ontology had a total of 50 researchers in total. Once the universe of professors was divided, the publishing ontology was populated using as a data source the DBLP file, which contains approximately one million articles and more than 56 million lines. The result of the population of the ***Publication*** ontology with professors of the systems department resulted in a total of 135 publications that coincided between the aliases of the professors and the authors indicated within the *<author>* label of the DBLP. At the same time 116 collaboration relationships were found among the professors of the systems department.

The ontology of professors of the electronics department, with a total of 58 professors, was subjected to the same test as the ontology of professors of the systems department and 22 publications were found from the same sample of the DBLP, that is, all the aliases were compared of the researchers of the electronics department against the authors of the million articles of the DBLP. In this ontology 4 collaborative relationships were found, that is, in two publications two or more professors from the same department participated.

As future work, other sources can be considered to continue enriching the ontologies with more semantic relationships, such as ArnetMiner [9], which contains abstracts and keywords of publications in order to enable the semantic relationship between publications, researchers and topics of interest. In this way, the ontology would comply with the characteristic of being scalable and make the profile of each researcher a more complete.

References

1. Maedche, A., Staab, S.: Ontology Learning. In: Handbook on Ontologies (2004)
2. Gruber, Tom. What is an Ontology?, 1993 ksl.stanford.edu/kst/what-is-an-ontology.html

3. Sowa, J. F.: Guided Tour of Ontology, 2001 http://www.jfsowa.com/ontology/ guided.htm.
4. Cámara, J. C.: Learning Metadata Standards: Ontologias Barcelona: UPF 2002 http://www.iua.upf.es/ jblat/material/doctorat/ students/jccbis/Ontologias.htm
5. Yao, L., Tang, J., Li, J.: A unified approach to researcher profiling. In: Proceedings of the IEEE International Conference on Web Intelligence, pp. 359–366 (2007)
6. Liu, P., Liu, K., Liu, J.: Ontology-based expertise matching system within academia. In: Proceedings of the IEEE International Conference on Wireless Communications, Networking and Mobile Computing, WiCom, pp. 5431–5434 (2007)
7. Thiagarajan, R., Manjunath, G., Stumptner, M.: Finding experts by semantic matching of user profiles (Doctoral dissertation, CEUR-WS) (2008)
8. Adnan, S., Tahir, A., Basharat, A., de Cesare, S.: Semantic agent oriented architecture for researcher profiling and association (semora). In: Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, vol. 3, pp. 559–562 (2009)
9. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 990–998 (2008)
10. Punnarut, R., Sriharee, G.: A researcher expertise search system using ontology-based data mining. In: Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling, vol. 110, pp. 71–78 (2010)
11. Grüninger, M., Fox, M. S.: Methodology for the design and evaluation of ontologies (1995)
12. Uschold, M., King, M.: Towards a methodology for building ontologies Edinburgh. In: Artificial Intelligence Applications Institute, University of Edinburgh, pp. 15–30 (1995)
13. Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In Proceedings of the 9th International Symposium on String Processing and Information Retrieval, pp. 1–10 (2002)
14. Gómez-Pérez, A.: Towards a framework to verify knowledge sharing technology. Expert Systems with Applications, vol. 11, no. 4, pp. 519–529 (1996)