Indicators and Functionalities of Exploitation of Academic Staff CV using Semantic Web Technologies

Isaac Lera, Carlos Guerrero, and Carlos Juiz
Dept. Matemátiques i Informática
Universitat de les Illes Balears
Palma de Mallorca, Spain
{isaac.lera,carlos.guerrero,cjuiz}@uib.es

ABSTRACT

We have transformed five years of curriculum data of our academic staff from relational databases to a semantic model. Thanks to semantic queries, capabilities of NoSQL models, inference reasoners and data mining techniques we obtain knowledge that it improves the personal management of curriculum data, the quality and efficiency of exploitation tasks, and the transparency, dissemination and collaboration with citizens. The huge catalogue of CV data remains an underutilized resource. Private companies such as editorials have robust services based only on publications but academic institutions have the option of integrating other databases related with their staff to obtain more indicators. We analyse the transformation of data, highlighting the mapping process of authors, and we present two ways of exploitation using semantic queries and complex networks. Thus, institutions, researchers and citizens will have a quality data catalogue for diverse studies.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks; K.3.2 [**Computer and Information Science Education**]: Accreditation, Curriculum

Keywords

Curriculum data; Semantic Web; Complex Networks; Performance Indicators

1. INTRODUCTION

The management of CV data in research activities is crucial to offer a productivity window of the activity of a researcher, group or institution, or simply, to ask for research grants. Most of the institutions offer a service of CV management to their members that is able for standardising CV representations, storing researcher outcomes (publications, patents, projects,...) and exporting in diverse formats (Marie Curie Actions, FP calls, and so on). It makes more efficient the functions of management. Institutions can have benefits from CV data as well. The analysis of CV data generates indicators of research quality. These measures bring us the possibility to improve the quality of assignment, control and planning of economical and human resources. In our opinion, this huge catalogue of CV data remains an underutilized resource. Furthermore, it is a fact, the exploitation of academic outcomes has benefits. There are a number of social networks that offer CV services and they exploit publication data such as: Academia, Mendely, and ResearchGate, and we can not forget publishers such as Thomson Reuters, and Elsevier.

The current role of any curriculum application is the storing of outcomes of a researcher member: articles in journals or conferences, courses or seminars, patents, projects, tutorials, etc. together with exportation functions in several formats to apply for a grant, a project call or a vacancy job. In our institution, the application was developed by a third company, or better to say, by a group of institutions, in any case we do not have any right on the code. An extraordinary modification or consultation requires a considerable amount of time and some bureaucracy. This system comes up against the continuous presence of new descriptive attributes of items, other types of items and querying requirements of specific information. The system adversely affects transparency, management and governance of the institution. It is worth highlighting that designers developed extra database tables to request the number of persons or outcomes associated with a group, area, faculty and department. The values of these tables remain outdated and they are incoherent with the reality of the institution. It means the manager uses the data of previous year.

Traditional Systems based on relational databases (SQL based) do not offer: 1. Flexibility to integrate new items or attributes. All items of a CV have similar attributes: pages, year, volume, authors, etc. but some owners or institutions can add new attributes: identifiers of research networks, links to open journals,.... 2. Combination with other data sources. Institutions can define politics and planning strategies thanks to the definition of new indicators based on the combination of private information about academic staff and CV data. 3. Private or public data consultations without the participation of qualified personal.

Flexibility, integration, querying and content publication capabilities are achieved with Semantic Web databases. They are NoSQL as well. Non-relational databases present some advantages such as: elastic scaling, flexible data schemas, memory-footprint, cluster dispersion, less management, and

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579036.

big volumes of data. An item stored in a NoSQL database has the attributes that the owner wish to store. For example, a publication item in a journal may or may not have impact index. This feature does not produce data inconsistency but a free-flowing representation according with actual demands. But still, there is an obligatory number of attributes to define an item. The only disadvantage is the design of the visualization tool that it may not include all possibilities.

Semantic Web (SW) paradigm promotes common data formats on the Web. This framework allows to share, publish and reuse data across heterogeneous applications. A semantic schema provide data coherency, disambiguation and implicit statements. In fact, following with our case of extra data tables, researcher outcomes are directly associated to the outcomes of a group, department, or similar entity where the researcher belongs. This is asserted by subsumption properties: subclasses and subproperties (and respective logic axioms) under an inference engine.

Roughly speaking, we define the new schema using classes, properties and links among another classes and properties [3]. In WS architecture, our outcomes are defined by URIs that it provides web addressability. We can link our data with other data to provide context. All these features and semantic languages RDF and OWL together shape the Linked Data [1]. This benefits us both the management and exploitation.

At exploitation level, the definition of a new indicator is non trivial. The manager has to know all possible data related with a goal, it defines the data wished and thereafter, a qualified personal with access to the system may do the request. Often, the request goes through a refinement and filtering process, occasionally with secondary tools. Unfortunately, either the lack of data or difficulty in accessing them the institution does not obtain a valuable knowledge. Our university publishes an annual report with the outcome of their academic staff.[1]. It is a web form with search functionalities such as: author, department, or type of outcome filters. These data are plain without any added structural value. In order to obtain knowledge from this source, the process should be a web crawler with a set of specific text patterns. If the web changes is not a reusable process. Such strategy is not available for anyone, it is not useful in terms of public analysis. Moreover, the creation of tools or the improvement of web to visualize these data is, in our opinion, a waste of time and resources. The institution already has this data and it has only to facilitate the dissemination in analytical terms and free use. It means in terms of Linked Data umbrella.

In this article, we apply semantic web paradigm and technologies to manage CV data obtaining useful knowledge in governance tasks, transparency and curricular services better than current methods. We present some details regarding ethical issues of public publication of CV, and a set of quality indicators not based on publication indexes. We have used a CV catalogue of four years thanks to the support our university, with which we can show real cases.

For hence, we propose a model of representation and publication based on Semantic Web for the correct exploitation of personal and institutional CVs and greater transparency of the institution and members to society. We have used a CV catalogue of five years with which we can show real cases. And, we present two ways of exploitation: querying and latent information from structural organization, using complex networks.

Definitely, we have grouped in three points of view the advantages of this approach. These points are justified throughout the document.

- **Personal**
  - Discovering new items. When an author defines an item and writes down the rest of authors or entities, the new item should be in their CVs as well.
  - Complementation of the definition of an item. Other entities or authors may complement the definition of an item. All related items should be updated automatically.
  - Establishment of collaborative networks. All authors are tagged in areas or departments, but their topics may have on several branches.
  - Exportation into any format. The adaptation of items well-tagged is simple using a list of queries or Document Type Definition (DTD) parsers.
- **Institutional**
  - Profit impact and performance indexes
  - Support to strategies and best practices
  - Reduction of administrative burdens
  - Monitoring mechanisms: successful degree of politics and changes to the scope of the work
- **Public**
  - Transparency
  - Social and technology trends
  - Establishment of collaborative networks among other public or private institutions.
  - Feedback: the management in terms of knowledge extraction can be used by citizens in different ways, for example, developing smartphone apps to localize resources or academic staff, etc.

2. RELATED WORK

Data manipulation and transformation into knowledge to improve the operations of an institution or company, the representation and publication of data on Semantic Web paradigm, and establishing metrics of scientific productivity of researchers have been analysed and developed in several jobs. This work is a combination of all these issues. It is a list of solutions addressed on the right exploitation of data of public institution. This work covers multidisciplinary areas: governance, semantic data modelling, data mining and complex networks.

The Spanish Foundation for Science and Technology (FECYT) develops the project called Normalize Curriculum Vitae. The goal is to simplify to national researchers the manual introduction of data into the different calls. They use the Extensible Metadata Platform (XMP), XML-based format, to define the schema of the items and to include them in PDF documents. In 2012, XMP has been standardized as ISO 16684-1:2012. It is a powerful tool for researchers since

[1]https://webgrec.uib.es/cgi-bin/Memoria/crgen.cgi?IDI=ANG

it generates incremental PDF with the information introduced by them, it imports manually information from Scopus, WOK, Cab Direct, etc. and it supports co-official Spanish languages. This XML information is represented using Web Semantic XML-annotation which facilitates the parsing to a real semantic schema (to RDFs and OWL formats). This representation does not offer the functionalities of a SW model. The data are isolated in a PDF document that the researcher has and the institution not. To the present day, the FECYT does not publish the common schema and It does not provide mechanism to extract XML information from PDF.

There is a publication in a journal of medical research that mentioned the use of Semantic Web languages to implement collaborative curriculum mappings for a complete medical model curriculum. This work by Spreckelsen et al. [9] focused on data modelling and on the web tool to use the CV. They introduce a semantic MediaWiki (SMW)-based Web application for the elicitation and revision process of the Aachen Catalogue of Learning Objectives (ACLO). The semantic wiki uses a domain model of the curricular context and offers structured (form-based) data entry, multiple views, structured querying, semantic indexing, and commenting for learning objectives (âĂIJLOsâĂİ). Thus, students can prepare for examination and teachers, a multidisciplinary course or trying to find out about the prior knowledge of the students enrolled. Their opinion was that SMW-based approach enabled an agile implementation of computer-supported knowledge management. In our opinion, they obtained a great performance of SMW tool but it was only applied to a specific goal of exploitation. One disadvantage was that students and teachers had to introduce their curriculum again. It does not comply with the suggestions of Linked Data.

There are parses that transform SQL to semantic models but we have avoided them since they use raw data from the SQL system and they only relates primary and external keys.

In [2] there is an interested study about the development of electronic CV national information systems in countries as: Norway, Spain, Portugal, UK, France, Israel, etc. As future recommendations they mentioned some points to develop that our solution include. The first one, since 2007 there is an application to the Swiss National Science Foundation to manage CV information but it is uploaded as PDFs where the extraction of data is more expensive. In SW , the representation and identification of data is easy and the querying, extraction and incorporation of more data is relatively simple. They presented another issue that it is the well-known reason of manual modification when researchers ask for grants or calls. In our case, we can extract detailed information from a set of queries for each specific purpose.

Thomson Reuters, one of the most important scientific editorials, has published documents about the evaluation of research performance using citation data [6]. These indicators have been used for institutions, and researchers for evaluation for purposes of accreditation, faculty review, etc. and planing of economic resources. Thomson Reuters have been analysing and developing tools for 50 years in natural and social sciences and humanities. It provides information about journals, entities or authors. This work does not attempt to supplant all previous services but the institution have more information about their academic staff than publication data. In our model, data integration from other sources is feasible and it provides new performance indicators.

Regarding with the potential of published data there are several universities that provide Open Data Services under Linked Data paradigm (e.g. University of Southampton, University of Leeds, The Open University, University of Oxford,...). The data catalogues are numerous: events, catering, phonebook, places, points of services, vacancies, equipment inventory, etc. Obviously, we open an ethical gap trying to have a curricular dataset of the academic staff. As a perfect example, we cite The University of Southampton: "There's data we have which isn't in any way confidential which is of use to our members, visitors, and the public. If we make the data available in a structured way with a license which allows reuse then our members, or anyone else, can build tools on top of it without needless bureaucracy ".

3. DATA MODELLING

As we have mentioned, Semantic Web databases models are based on RDF and OWL languages. Each language provides a repertoire of constructors which give a logical interpretation to the data. We have transformed each table of the SQL model in a list of classes. This is not a systematic process since there are attributes which also are classes.

The most illustrative example of the whole process is using a publication outcome and its SQL entry. To develop an automatic tool that extract the information from database we had to understand the schema under the model without documentation. We have performed reverse engineering helping us with specific known cases: our outcomes. We have only 30 tables. The table of journal outcomes has 33 attributes: Code, External Key to unknown table, Authors -string-, Title, External Key to Publisher table, Additional Information -string-, volume, number, initial page, final page, year, data of publication, DOI, physical deposit -string-, Noval2 -string-, Noval3, Noval4. Observations -string-, Reste -string-, Keywords -external keys-, owner, acceptable, number of authors, character -external key to unknown table-, Link to the online document, Noval1 -string-, year of this entry, checked, abstract, Noval -string-, data of publication-, include in the annual report -boolean-. We realize there two types of attributes: descriptors of the outcome (title, abstract, keywords, etc.) and attributes to manage the item ($Noval_i$, owner, acceptable, number of authors, year of this entry, checked, etc.). The first ones are useful the rest are to support the management. There are 12 of 33 attributes really useful to describe the item, the rest are for internal management tasks. Some of them are unknown for us such as: Noval, data of publication (x2) and Reste. In fact, most of these attributes are not used for the users, tools or even the administrators. In 2012 year, there is 978 journals publications, lines in the file, and also, 2004 participations in conferences. For example, we show an item, a publication in a journal, where more or less you can identify each field separated by a token symbol.

```
539317#%#AA#%#Juiz, C.; Gomez, M.; Barcelo, M.I.#%
#Implementing Business/IT Projects Alignment through the
Project Portfolio Approval Process#%#203418#%##%#180#
%##%#1#%#8#%#2012#%#20120000#%
#10.1007/978-94-007-5082-1_1#%##%##%##%##%#
#%##%#CURRICUL#%#20121023#%#N#%##%#IN#%##%##%
##%##%##%##%#
```

```
<owl:NamedIndividual rdf:about="http://...#
    ImplementingBusinessitProjectsAlignment
ThroughTheProjectPortfolioApprovalProcess">
 <rdf:type rdf:resource="...#JournalOutcome"/>
 <volume>1</volume>
 <initialPage>1</initialPage>
 <author rdf:resource="http://...#JuizGarciaCarlos"/>
 <author rdf:resource="...#GomezMercedes"/>
 <author rdf:resource="...#BarceloMariaIsabel"/>
...
</owl:NamedIndividual>
```

Figure 1: OWL Transformation of a journal outcome

```
<owl:Class rdf:about="..#WellFormedOutcome">
  <owl:equivalentClass rdf:resource="..#Outcome"/>
  <owl:equivalentClass>
    <owl:Restriction>
      <owl:onProperty rdf:resource="...#author"/>
      <owl:onClass rdf:resource="...#Researcher"/>
      <owl:minQualifiedCardinality rdf:datatype="&xsd
          ;nonNegativeInteger">1</owl:
          minQualifiedCardinality>
    </owl:Restriction>
  </owl:equivalentClass>
 ...
</owl:Class>
```

Figure 2: OWL Transformation of a journal outcome

This data does not have sense out of the context. Even, a computer interprets strings and numbers. When we use semantic constructors we are creating knowledge. We are able to provide solutions to an imaginable list of questions that it will define performance indicators. For example, what is the number of authors? Has this journal an impact factor?, Has Juiz more publications regarding IT Business?, Has Juiz common projects with the institution where Gomez or Barcelo are working?, etc.

To do that, we have to transform plain text to tagged text or tagged URIs where these tags have logical implications. Firstly, URIs have to be friendly because it simplifies human interpretation. In the figure 1, there is a part of the transformation of previous example into OWL language. The *NamedIndividual* constructor defines an item. This item can be addressed using the *about* URI which is friendly. This item has some properties among other items, which are called *ObjectProperties*. *author* is an *ObjectProperties* that it has a *resource* attribute with a friendly-URI value as well. This author is unique. Instead, the text "Juiz, C." can give possibilities to another interpretations of authors. When we use friendly-URIs, we do not need numerical keys (i.e. 53917).

Secondly, there are no semantic fields. The semantics is useful when you can relate terms among them but in some fields are not. In some cases, you cannot identify the usefulness until there is an exploitation goal. In our example, the attributes are volume, initial page, observations, and so on. These attributes are tagged with *DatatypeProperties*.

Thirdly, all specific cases, the individuals (*NamedIndividual*) always are of a type, at least. This type points to a class. In this case, the class is a *JournalOutcome*. It means this individual has the same logical implications of that class. In our case, these individuals is a JournalOutcome, but it is also an Outcome, and it is also a BookPublication. We are creating knowledge: How many researchers of our institution do book chapters have ? The definition of a classes is dynamic. Either you can introduce the definition in the model or you can import the model with your definition inside and automatically the inference engine filters that information. The individuals can have implicitly or explicitly that type tag. For example, we can define a "well formed of an outcome" class, see figure 2. Basically, this class is defined by all researchers outcomes that they do have at least one *author* property. Thus, the previous individual is also a well-formed definition. This list of items can be used to manage those outcomes that are poorly defined.

Fourthly, the degree of exploitation is also determined by the hierarchical relationships among ObjectProperties, properties among classes. Often, we use the comparative of Semantic Modelling with Object Oriented Programming to try explain the definition of classes. In the case of properties, it has nothing to do. ObjectProperties can be subsumed by other ObjectProperties and they have mathematical characteristics. An ObjectProperty can be functional, inverse functional, transitive, symmetric, asymmetric, reflexive and irreflexive. It can have inverse, disjoint and equivalent properties. For example, the *author* of a patent is also an *inventor* and *ownership*. This fact simplifies the obtaining of results since it is equal to say: who is the author of this patent? or who is the ownership of this patent?

In this first approach, we have keeping the actual schema removing unnecessary fields, and establishing relationships between items that they are useful to manage decisions. We list our classes: area, professional category, conference, qualification, editorial, entity, centre, superintendent office, department, external department, building, school, faculty, institute, laboratory, other, service, group, researcher, outcome, patent, project, publication, journal, work, work of degree, master, or thesis. We do not list ObjectProperties and DatatypeProperties.

We have incorporated the catalogues of the 2008, 2009, 2010, 2011 and 2012 in a semantic Web database that it incorporates an inference engine for providing suitable responses according with the group of semantic constructors that it manages.

3.1 Author identification problem

There is a critical point applying linguistic parsing techniques to transform plain data in some known data. In our case, It happens when we have to identify the different authors from a text that researchers write down freely. For the institution, It is a serious matter to base policies on unrealistic information. Most of performance indicators are based on the outcomes of the academic staff, the authors.

In the current model, we have support tables to identify the owner of each outcome and we can relate it with one of the authors. Thus, Carlos Juiz introduced the previous item in the system. But perhaps, Juiz is not the author of the publication and he had other reasons to store the item: he collaborated with these authors, he related the publication with the results of a project, etc.

Our algorithm identifies the full name (name, middle name, first and second surname), if any, of each member and it tries to establish a mapping between the full name and the split text of author field. We have detected extreme cases from cases where the full name is set by the only one available surname to cases where the middle name has more length than both surnames together, doing hard its identification. We have found several cases to separate the authors in the

corresponding field. Thus, we have the classical separation using semicolon : "Lera, I.; Juiz, C.", or other cases such as: "Lera I. Juiz C", Isaac Lera and Carlos Juiz", "Lera, Isaac . Carlos, Juiz", "Sr. Lera", "Dr. Lera and Dr. Juiz" and a long list of possible combinations. And this is exacerbated when there is a middle name and a second surname, plus in Spanish official variants the 'and' is 'y' and 'i', We have solved this using multiples text patterns, owner item of database and a final mapping based on sub-entity where authors work together. It means, for all possible combination the algorithm chooses the mapping that it has a working relationship. For example, the text in the author field is:"Lera, I. and Guerrero, C.". There is two possible maps for the first well-split author: Isaac Lera who works in ACSIC group and Isabel Lera who works in IUNICS institute . The second author is mapped to: "Carlos Guerrero" who works at ACSIC group. Thus, our approach, it is to map the first author with the first choice.

The evaluation of this mapping algorithm should be carried with an active collaboration of the academic staff. Each member should be checked each item. It is problematic the explanation of this task when there is a system that it is right working and giving the functionalities that they are necessary from his or her point of view. The dimension of the problem is around of 7752 mappings for year.

4. DATA EXPLOITATION

We open two ways to analyse the data knowing what we are looking for and discovering latent information on the structure of data.

The first one is using SPARQL query language [7]. SPARQL can be used to request data across several catalogues and it has capabilities for optional graphs -sources- and diverse function logics. Thus, the manager knows the information that it needs for the governance goal and It only has to define the query. For example, it can establish a new performance index between the relationship between a newbie researcher -using the employee category- and their related outcomes. We can use the query of the fig. 3 to obtain a list of the researchers, outcomes and its type (publication, degree project, etc.). It has a representation RDF-based statements: subject, predicate and object. The variables are represented with '?'-symbol. Thus, we can literally to read that a researcher has outcomes, the outcomes are of a type, and the researcher has a category. Obviously, the resulting data requires a suitable processing to obtain more useful knowledge but this task has lower cost than the current process. Moreover, the complexity of a query depends on the expressiveness and richness of the schema. The second query (fig. 4) is a little more complex than the first. It obtains the number of times and countries where Isaac has travelled in function of places where he has published. Perhaps, this information provides a future governance task: a specific call of collaborations with a determined countries.

The second way is using other techniques such as complex networks to highlight latent facts. Semantic Web catalogues can be transformed into graph models. Classes are the nodes and the ObjectProperties are the edges. Moreover, we can include attributes to describe them visually: the type of outcome, the label of the journal, the country, and so on. Thus, we export our databases in RDF/OWL format and, at the same time, in GraphML format. It is supported by

```
//Prefix section
SELECT  ?investigator ?employeeCategory ?outcomes ?type
WHERE {
?investigator model:has ?outcomes
?outcomes rdf:type ?type
?investigator model:has ?employeeCategory
}
```

Figure 3: A SPARQL query of researcher's outcomes

```
//Prefix section
SELECT ?country (COUNT(?country) as ?total)
WHERE {
?outcome base:autor base:LeraCastroIsaac .
?outcome base:country ?country
} GROUP BY ?country
```

Figure 4: A query example to provide the countries visited by a researcher

multiple analysing and representing network tools such: Cytoscape [8], Gephi [4] or NetworkX [5].

For the next analysis, we have used the catalogue of the 2012 year, the most current network. It is the last available catalogue that we have in DDBB format, although it is published at the web of the institution the 2013 dataset. We obtain a quick approach of the data organization using Gephi. We have 28072 nodes and 26309 edges.

The first observation is compared to the current mismanagement of the information about publishers stored in the each catalogue. That number is 62.79% and these values are replicated each year. A node is a row, an item, in the database. In any case, the number of researchers are 20.75%, publications in conferences or workshops are 5.02%; journals are 3.42% and thesis are 0.3%. In the figure 5, there are represented the network removing the nodes of publishers. Elements without edge that are outside of the networks most of them are institution services, where their management is unnecessary. But also, there are elements from the CV of researchers that they do not have well-defined. For example: projects, areas or groups without assigned members. It means, each year the institution gathers information unnecessary or without any utility and the researchers do not receive any feedback about the degree of CV completeness.

As a viable case of exploitation with networks, we have analysed the degree of collaboration among academic personal of our institution and others in terms of scientific publications (SP). We have calculated the eccentricity distribution (see figure 6). The eccentricity is the maximum graph distance between a vertex and any other vertex. Thus, there are around 4000 sub-graphs that they have one node. The rest presents lower length paths. In terms of publications, it means there are separations among the personal: lower values better degree of collaboration. But also, values from 14-24 are indications of people who work in several groups.

5. CONCLUSIONS AND FUTURE WORK

The transformation of CV from row data of SQL databases to an accessible, addressable and querying semantic model under Semantic Web and NoSQL paradigms provides knowledge that it improves the personal management of curriculum data, an increase in the quality and efficiency of exploitation tasks, and the transparency, dissemination and collaboration with citizens . Furthermore, the semantic model facilitates normal queries (e.g. list of IP, list of projects or

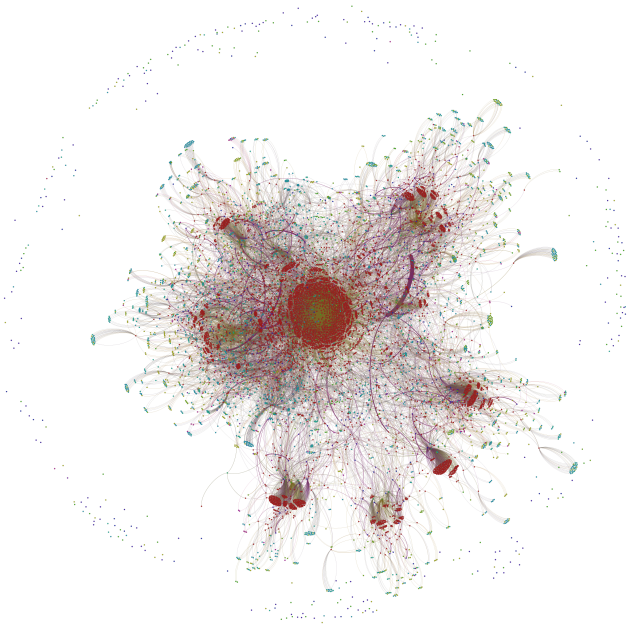

Figure 5: Relationships between researchers and outcomes in 2012

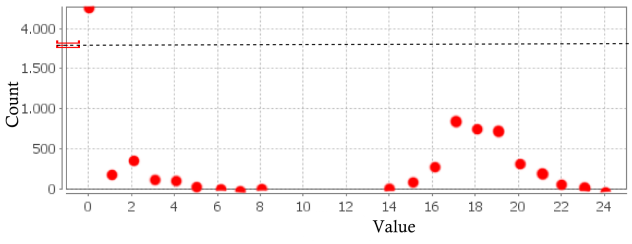

Figure 6: Eccentricity Distribution values

groups, etc.) alleviating the impact of the management staff. And, It facilitates the integration with other data sources. Along the document, we have analysed the main advantages of a semantic representation.

We have opened an ethical gap. Curricular data is sometimes confidential. In any case, most of the research institutions publishes quality reports. We have unified the representation of CV data in a common format to extract useful knowledge and we have published these catalogues that may be analysed by anyone. This requires previous processes of supervision to avoid erroneous or malicious interpretations. The exploitation of data with the creation of new performance indicators can be based on a specific set of querying or latent indicators applying complex networks or other data mining techniques. We have proposed two SPARQL queries and an interpretation of an eccentricity distribution of networks.

As future work, we are developing the web management tool for browsing and querying data from the whole database and we are analysing new performance indicators in function of evolution of the outcome along years.

6. ACKNOWLEDGEMENTS

Thanks to Vice-chancellor for Innovation and Dissemination and to *Oficina de Suport a la Recerca* by providing data catalogues and exploitation guidelines. This work is partly financed by the Spanish Ministry of Education and Science through the TIN11-23889 project.

7. REFERENCES

[1] Design Issues about Linked Data http://www.w3.org/designissues/linkeddata.html.

[2] Building new indicators for researchers careers and mobility based on electronic curriculum vitae. http://hdl.handle.net/10400.9/401, 2009.

[3] OWL 2 Web Ontology Language (Second Edition) http://www.w3.org/tr/owl2-primer/, 2012.

[4] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[5] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug. 2008.

[6] D. A. PenDlebury. White paper: using bibliometrics in evaluating research. thompson reuters. 2009.

[7] E. Prud'hommeaux and A. Seaborne. Sparql query language for rdf. Latest version available as http://www.w3.org/TR/rdf-sparql-query/, January 2008.

[8] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Osgi alliance cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, 2003.

[9] C. Spreckelsen, S. Finsterer, J. Cremer, and H. Schenkat. Can social semantic web techniques foster collaborative curriculum mapping in medicine? *J Med Internet Res*, 15(8):e169, Aug 2013.