INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Tecnologías para el manejo de metadatos en artículos científicos

COMPUTER AND SYSTEMS ENGINEERING

Technologies for metadata management in scientific articles

Alexander Castro-Romero[*], Juan S. González-Sanabria[*], Javier A. Ballesteros-Ricaurte[*]

[*]Escuela de Ingeniería de Sistemas y Computación, Universidad Pedagógica y Tecnológica de Colombia. Tunja, Colombia.
alexander.castro01@uptc.edu.co,juansebastian.gonzalez@uptc.edu.co,javier.ballesteros@uptc.edu.co

(Recibido: Mayo 22 de 2015 – Aceptado: Junio 26 de 2015)

Resumen

El uso de tecnologías de la Web Semántica ha venido acrecentándose, por lo que es común usarlo en diferentes aspectos. Este trabajo evalúa como estas tecnologías pueden contribuir a mejorar la indexación de artículos en revistas científicas. Inicialmente, se hace una revisión conceptual de los metadatos, para posteriormente estudiar las tecnologías más importantes para el uso de metadatos en la Web y, de esta manera, escoger una para aplicarla en el caso de estudio de indexación de artículos científicos, determinando los metadatos con bases en los usados por las revistas de investigación de impacto y construir un modelo para la indexación de artículos científicos usando una tecnología de Web Semántica.

Palabras Clave: *Anotaciones semánticas, búsquedas semánticas, recuperación de la información, web semántica.*

Abstract

The use of Semantic Web technologies has been increasing, so it is common using them in different ways. This article evaluates how these technologies can contribute to improve the indexing in articles in scientific journals. Initially, there is a conceptual review about metadata. Later, studying the most important technologies for the use of metadata in Web and, this way, choosing one of them to apply it in the case of study of scientific articles indexing, in order to determine the metadata based in those used in impact research journals, and building a model for indexing scientific articles using Semantic Web technologies.

Keywords: *Retrieval information, semantic annotations, semantic search, semantic web.*

1. Introduction

The data has been, is and will a topic of constant importance for those human beings who, although being in a continual searching of how to manage it for ease its consulting and get useful elements for making decisions; are still facing numerous challenges about its management, so they have created a set of tools to contribute with its administration.

One of the earliest ways that appeared to manage the data was the use of short labels with relevant data about an element; for example, putting a piece of paper with the price of a product. The purpose of this practice was to have more data about the elements.

However, in the digital aspect, this technique was implement under the name of 'metadata', which appear as an alternative to register internally data from the resources. An example could be a MP3 song, since this, besides of the audio, stores data of the song as its name, the album cover and even its lyrics.

In the beginning, metadata management were devise for describing the resources, but along time, has been sought that machines understand these descriptions and implicit relations. Nowadays, the largest data network is Internet, where many resources are share between computing devices, and therefore, some technologies are necessary to manage and make easy the understanding of metadata in order to easily access to these resources.

One of these technologies is Semantic Web which is based on giving value to metadata by using them to improve the results of web searches (make them easier, faster and more accurate) and contribute with the arranging of the content. The above, starts from the fact that one way to achieve better quality of the search results is the standardization. Accordingly, for example, in a city is relatively easy to search for an address since there is a format adopted for its representation; in change, if every person used a different notation, although it would be possible to find an address, it would be also too complicated. By associating that example with the Web, it is evident the need to pose a standard to manage the metadata.

Although it is obvious that the Web has evolved in recent years, it is hard to find the data searched due to 'disorder' in the resources on the Internet. In 2001, the Semantic Web was present as a proposal to improve that situation but unfortunately, it has not had much success. In the words of one of its authors, this one "continues largely unfulfilled" (Shadbolt et al, 2006), and to make an actual analysis, the situation has not improved, the standards are formulated and in constant improvement but unimplemented; the technologies are not being used and the quality of the search results has improved but thanks to the efforts of big companies in the search industry such as Google, Facebook and Twitter, and not to the application of Semantic Web.

Nevertheless, it should be mentioned that one of the causes for the application of Semantic Web is that the largest companies don't build the internet, if not the final users for whom it is complex to incorporate a standard when sharing its resources, and that's the reason why it is necessary bring the users towards these technologies in a friendly way and where possible, in a "transparent" way.

Therefore, this article studies the theory of the metadata, its integration with Semantic Web through standards, and how to integrate these technologies, initially, with the field of scientific publications.

2. Metadata

Metadata are a technique to manage data from an element. It is said that these are "data about data"; a more formal definition is that "these are data and documentation that make the data to be comprehensible and sharable for users through the time" (ISO/IEC 11179-1, 2004). On the Web, metadata play an important role since these are a key part of the infrastructure needed for helping to create order in the chaos of the Web, by

injecting description, classification and arranging for creating data stores more useful (Duval et al, 2004).

Its working is easy: to use words to describe resources in terms of qualities, for example, if is wanted to describe a photo, metadata can be used to store the resolution in an established unit, the date when it was taken, the place, the people appearing and even something as complex as the feeling of someone when watching it. These data are not part of the photo itself and even the photo can exist without this information, but if these data are registered and well managed, can allow more information about the photo that can be related with another relevant data for those who consult the photo. Following with the example, if the date and the person appearing in the photo are registered, this can be used to determine that this person was in that place or even it could be determined the weather and up to make predictions about it. Namely, the more data are register about an element, the more likely to obtain information and that this one can be useful to make decisions.

However, despite of its importance, metadata have not had the expected relevance since is complex to make that Web users use them correctly and the big searchers are doing this task. If a parallel with a library were made, the Web would prefer to contract many employees for organize and deliver to people the book that they are looking for, than to order properly those books for being sought by the people.

It is necessary to highlight the following four aspects about metadata:

2.1 Classification

Although there are different classifications for metadata, one of the most accepted makes a division according to its function (Baca, 2008):

Administrative: These ones are used for managing the resources, for example, the location of a resource.

Descriptive: These ones are used to identify and describe the resources, for example, the version of a software.

Preservation: These ones refer to the status of resources, for example, the current condition (if it is working or not).

Technical: These have to do with the relation of the resource with the system, for example, the format of an image.

Use: These represents data related with the level and type of the resource use, for example, the records of a Web server.

2.2 Characteristics

Taking into account that metadata are importable in the field of data management, these ones must keep the following characteristics (Audit Commission Publishing Team, 2007):

Accurate: These must be detailed enough.

Valid: These must be consistent with what purport to represent.

Reliable: These must be stable and credible.

Timely: These must be updated.

Integral: These must contain the promised information.

2.3 Functions

Metadata are used for (National Information Standards Organization, 2004):

Discover resources: Allowing searches among resources group them and skew them.

Organize resources: Grouping resources under the same metadata, without being a rigid scheme, easing its management and giving a sensation of natural order.

Interoperability: When metadata are used to describe resources, both human and machines

can understand it; besides allows resources migrate to another platform without suffering data losses.

Identification: Allow assigning a unique identifier to a resource, either a code or an URL (Uniform Resource Locator); moreover, with metadata, a resource must be distinguished from another.

Archiving and preservation: Metadata are key for assuring that a resource is going to be useful in the future and its information is not going to be altered.

2.4 Issues

This topic has received little discussion, however, it could be mentioned that metadata do not work because of (Doctorow, 2001):

People lie

People are lazy

People are partial to their contents

The schemes are not neutral

The metrics influence results

There is more than one way to describe something

These issues make metadata not as reliable as they should be, but thanks to the standards, the technologies developing and the commitment of the Web users, these ones can be overcome.

Therefore, metadata are a powerful choice to organize the Web, by describing resources through labels to be understood by machines and humans. However, there is still much work for a correct implementation, therefore a first step to build a solution in this sense, consists of studying current technologies for metadata management and the standards ruling these ones.

3. Resource description standards

In the data management field, as had been mentioned before, is necessary the use of standards for metadata management on Web, highlighting the following:

3.1 Microformats

These are "a simple way to add bookmarks to readable data elements such as events, contact information or places, in web sites, so the information within them can be extracted, indexed, searched, saved, by crossing references or combining them" (Microformats community, 2014). In simple terms, these are a way to add a structure to data and marking the metadata taking into account some labels, all on HTML (HyperText Markup Language) code. Such sets of labels are defined for common elements, to illustrate, if it is wanted to describe an event (h-event); some labels as name (p-name), start and end date (dt-start - dt-end) and location (p-location) are used.

Going into details, the model works with three basic elements: the classes or vocabularies that reference elements from the real world (for example a contact or a review), the prefixes that indicate the type of data with which it works (for example "dt-* for data related with time) and those characteristics of the classes (for example "dt-start" that refers to the start date of the event class). Further, it must be taken into account the use of links, because if the description of the relationship is indexed when a link is placed, the Web would make more sense (Khare & Çelik, 2006).

This technology brings several advantages: promotes standardization, helps to enable the construction of software services distributed as aggregation and indexation, and allows the interoperability between Web apps and desktop apps; as Bill Gates said "We need Microformats and to get people to agree on them. It is going to bootstrap exchanging data on the Web" (Allsopp, 2007).

On the other hand, it is said that Microformats are usable, unlike Semantic Web robust technologies (Khare, 2006); these may not be such complex, but

right there is its strength, because of its simplicity, these ones are easy to use for final users, who must index its resources on Web. Shortly, these may be the starting point to "evangelize" people on Semantic Web, making these a useful step to start ordering the Web.

In summary, Microformats are a simple technology to endow the metadata of a resource with a basic semantic structure. These provide basic templates to describe objects of the real world, its disadvantage lie in that these fall short when connecting resources, "publishers need to understand that Microformats are not a widespread universal solution to add semantic to all possible ontologies within HTML" (Suda, 2006).

3.2 Microdata

It is another way to include metadata on HTML code. It is composed by vocabularies, which offer a way to describe elements using a key value scheme; "MicroData allow search engines and another automated processes making sense to data on a Web site, such as the identification of the title, the author and the identification number of a book, in all the content of a site" (Scott, 2013).

This model consists of three parts: the itemscope, which indicates to the browser that a content is going to be described; the itemtype, which indicates what type of content is going to be described and the itemprop, which refers to the attribute of the entity that is going to be described. For example, there could be an Organization itemtype such as "street-address" along with the address where the meeting is going to take place; there are other elements such as itemref, which serves to manage item lists by grouping them, and itemid, which is used to save an identifier own to the entity.

Emphasizing that this standard is used along with a library called Schema.org, which allows standardizing templates to describe resources, counts on schemes to describe many elements as books, events, movies, people; the idea is that the fields of every resource are defined and can be interpreted correctly by different machines (Ronallo, 2012).

An advantage about integrating this system with HTML is that when using a library of standard vocabularies such as Schema.org, it can be possible indicate to the browsers and searchers how to show relevant content in Web searches. For example, if it is wanted to search for a restaurant and metadata such as the number of stars or the price range were modelled, this information can be useful to the user, because it wouldn't even have to enter in the restaurant Web site to know if that is what is looking for (Pabitha et al, 2011).

Nevertheless, this model has had some problems when adopting it (Tomberg & Laanpere, 2009), because it is supported by WHATWG (Web Hypertext Application Technology Working Group) (WHATWG Editors, 2014) which deviates from the World Wide Web Consortium (W3C), a leader in the Semantic Web Project.

3.3 RDF (Resource Description Framework)

This is a standard to describe Web resources by using a triplet model created by the W3C, "is a language to represent data about Web resources, designed for situations in which is necessary that the apps process the information, rather than just be shown to people; RDF provides a common framework to express this information and be exchanged between apps without loss of meaning" (Manola et al, 2004).

This data model is based on three elements of the common grammar: the subject, person or entity that will be described; the predicate, which refers to the characteristic of that subject; and the object that is the value of that characteristic. For example, in the triplet "Diego has brown eyes", the subject is Diego, the predicate is the eye color and the object is the brown color; as can be appreciated both the subject and the object are resources while the predicate is the existing relation between these ones (Decker et al, 2000).

Other important elements in RDF are:

The Internationalized Resource Identifier

(IRI), used to identify a resource, the URLs are types of IRIs because these indicate where is the resource on the Web and give a way to find something that differs from the others,

Literals are values that do not worth storing in an IRI. Data to be taken into account but without enough characteristics to be a resource,

White nodes, which are resources without IRI, represent something but it is no interesting to know its characteristics.

Thus, RDF provides a scheme to describe vocabularies through classes, properties, types, domains and ranges. Indeed, the purpose of this technology is to construct a semantic graph of knowledge on Web, "the main and differentiator value of the capacity of the Semantic Web is the ability to connect stuff" (González, 2014). The advantage of RDF versus other technologies lies in this point.

In summary, these three technologies are relevant to add a semantic structure to resources represented in HTML (Pastore, 2012). This will allow both machines and humans understand that a Website not only shows text if not also a resource that has characteristics, that can be related with other resources, as in the real world. In this point, it would be irrelevant to suggest a specific technology, taking into account that, although these pursue the same thing, have their own architectures, advantages and disadvantages; but it can be recommended avoiding to fall into technical discussions and use the one which fits better to the needs of the organization.

Finally, it is important that these technologies serve to layout content; however, it becomes necessary to speak the same language that is, use the same terms, so that is recommended to use a standard vocabulary as Schema.org does (García, 2013).

4. Metadata in scientific journals

One of the areas where metadata have a huge importance is in scientific journals, since its correct use in published articles, helps to spread the knowledge in a more accurate way.

It is for this that metadata are a key factor in the record and distribution of the scientific information, since with its good use, scientists can publish and share data, allowing that results from experiments and studies can be examined, searched and cited, thus encouraging the reuse of data among scientific disciplines (Matthews et al, 2010).

Nonetheless, the upgrade of the scientific journals has been slow. The articles still being shared printed or through PDF files (Portable Document Format) and its transition to HTML Websites has not been completed yet, which impedes the correct indexation of the articles by the web searchers since no use is made of metadata technologies and neither exists a standard about this.

However, the elements in metadata standards for scientific data transmit the essential information about the creator, the contexts, temporal and geospatial parameters, and the details of the process quality (Qin & Li, 2013). The purpose is that the entities form an interrelated network of nodes, which allows an optimal management of the resources. However, the representation of the characteristics of all these entities and its relations as metadata turns out to be a daunting task due to the huge amount of elements in metadata standards on science domains and the complex linguistic and syntactic forms used.

An example of the aforementioned challenges is this article that was presented in two different websites (Figure 1 and 2).

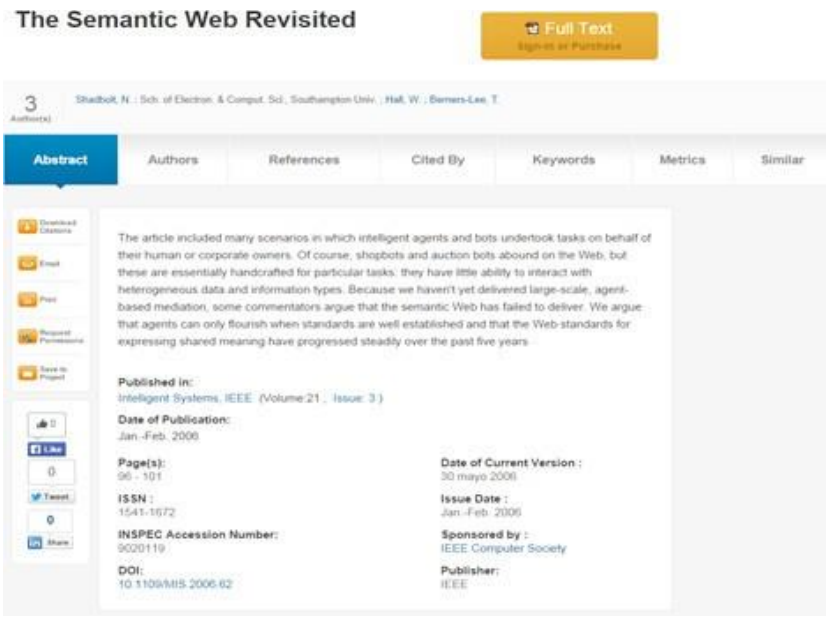

Figure 1. *Article "The Semantic Web Revisited" on IEEE website.*

Figure 2. *Article "The Semantic Web Revisited" on ACM website.*

As can be appreciated, with the same resource there are important changes in the manner in which the information is presented, even in those metadata in which are taken into account, so it becomes necessary unify the basic information of a scientific article in order to design a model supported in some of the aforementioned technologies.

For this, "Scimagojr" was consulted, one of the most important journal indices which "classifies the specialized journals on the basis of citation weighting systems" (González-Pereira et al, 2010), and the top six worldwide journals and the first four journals in Colombia were taken as a sample to make comparison of its metadata, obtaining as a result the Table 1.

Table 1. *Top journals and metadata.*

Journal\Metadata	Title	Author	Publication Date	Journal dimension	Edition	First and last page	DOI	URL	ISBN/ISSN	Reception date	Number of pages	Publisher	Key words	Language	Copyright
Ca-A Cancer Journal for Clinicians	x	x	x	x	x	x	x	x	x						
Reviews of Modern Physics	x	x	x	x	x	x	x	x			x	x			
Annual Review of Immunology	x	x	x	x	x	x	x	x	x			x			
Cell	x	x	x	x	x	x	x	x				x			
Annual Review of Biochemistry	x	x	x	x	x	x	x	x	x	x		x			
Quarterly Journal of Economics	x	x	x	x	x	x	x		x						
Colombia Médica	x	x	x	x	x		x	x				x	x	x	x
Revista Colombiana de Estadística	x	x	x	x	x	x									
Universitas Psychologica	x	x				x									
Livestock Research for Rural Development	x	x	x	x	x									x	

It must be clarified that it was verified that metadata were accessible for both people and machines, that is, that had a clear query format. Three outstanding aspects were found in metadata management:

RIS technologies are used (George, 2006), which is a format for bibliographic data on web, composed by two uppercase letters that act as key and a space and a dash followed by a data that acts as value. For example, "JO- UPTC Engineering Journal", where "JO" refers to the journal name metadata and is followed by the value of the journal example. Even though this format is very useful for bibliographic managers and it is understandable for humans, it is not so transparent and it has several shortcomings.

Importantly as the most of the search sites of scientific articles allow exporting its metadata in a CVS (Concurrent Versions System) file, as mentioned in Antwerp & Madey, 2010. Sorrowfully, the mentioned options do not contribute to Semantic Web.

The only effort for a semantic indexation is the use of the Dublin Core vocabulary (DC initiative, 2012), composed by fifteen labels used for describing resources.

Nevertheless, there are worrying things as the fact that the key words, the abstract, the type of file, the area, the acceptance or shipping date, the email of the author, among other useful fields when consulting, are not included.

In summary, in such an important area it is necessary a better management of the metadata, as possible with a standard format, thus in the next chapter it is proposed an approach to a format for describing a scientific article by using Semantic Web technologies.

5. Using semantic web technologies to describe a scientific article

Although there are efforts for scientific data modelling, these are so complex that few people use them, for example, the Core Scientific MetaData Model (CSMD) (Yang et al, 2010). For this reason, it is proposed avoiding reinventing the wheel and, in change, using Semantic Web technologies mentioned with a vocabulary defined to represent a scientific article.

Although it would be easy to model a scientific article with any of the aforementioned

technologies and vocabularies, RDFa (a simple version of RDF) and Schema.org were selected, since these are the ones with more future (Mika & Potter, 2012) and are the more robust ones. It emphasizes that in the process of data integration to the Web in a semantic way, the technology is important but the model used to describe the resource and the popularity of its vocabulary is much more relevant.

Below it should be mentioned that content indexation has several phases, but first it is recommended the analysis phase where it is scanned which metadata are taken into account for modelling a resource and its format. For this, characteristics in Chapter 2 must be taken into account. For the case study, the analysis phase results are on Table 2.

Table 2. *Metadata for a scientific article*

Name	Format	Description
Title	Uppercase text	Title
Author	Surnames, Name	Author's full name
Key words	Separated by a coma	Words describing the topics covered
Publication date	yyyy/mm/dd	Date when the article was published
Abstract	Text	Abstract
Volume	Number	Publication volume
Number	Number	Journal number
First and last page	pp-pp	Location on the journal
URL	Text	Url or the page of the abstract
DOI	Text	Digital object identifier
Language	Text	Language in which it is written
Affiliation	Text	The place where the author works or is attached to
ISBN	Number	ISBN of the journal
Journal/conference	Text	Name of the publication where the article is registered
Citations	Comma- separated list	List of the articles mentioned in this article
References	Comma- separated list	List of the mentioned articles
Copyrights	Text	License
Format	Text	HTML, PDF, Other
City	Text	The city where it was produced
Contact	Text	Web, email or social networks of the author
Publisher	Text	Publisher entity

After establishing a format that can even be filled in paper, a parallel with a vocabulary or ontology is made, Schema.org is used to have a better standardization. This process results on Table 3.

Table 3. *Vocabulary for a scientific article*

Name	**SCHEMA PROPERTY**
Title	NAME
Author	AUTHOR
Keywords	KEYWORDS
Publication date	DATEPUBLISHED
Abstract	DESCRIPTION
Volume	VOLUMENUMBER
Number	ISSUENUMBER
First and last page	PAGESTART PAGEEND
URL	URL
DOI	SAMEAS
Language	INLANGUAGE
Affiliation	SOURCEORGANIZATION
(e)ISSN	ISSN
Journal/conference	NAME RELEASEDEVENT

Name	SCHEMA PROPERTY
Citations	CITATION
References	CITATION
Copyrights	LICENSE
Format	?
City	WORKLOCATION
Contact	EMAIL
Publisher	PUBLISHER

The process was simple because Schema.org provides a hierarchy in which resources to describe are found, in this case, "CreativeWork, Article, ScholarlyArticle" were the basis to describe the scientific article, although "Book" and "Organization" were also used, the complexity lay in the description of the format. Furthermore, it can be appreciated that this vocabulary has many more characteristics that describe an article, but from experience it is known that many fields in a format tend to be disregarded, therefore the work is going to be only with the 20 mentioned. Finally, the syntax of the selected semantic technology is integrated; in this case, the RDF triplets are assembled. In Figure 3, the basic data of a scientific article are shown, especially those from the author.

```
<div vocab="http://schema.org/" typeof="ScholarlyArticle" resource="#article">
  <strong>Título:</strong> <span property="name">Utilidad y funcionamiento de las
  bases de datos NoSQL</span><br/>
  <strong>Autor:</strong>
  <div  property="author" typeof="Person">
    <span property="name">Castro Romero, Alexander</span>
    <span property="workLocation">CO</span>
    <span property="email">alexander.castro01@uptc.edu.co</span>
  </div>
  <span property="inLanguage" content="es">
  <strong>Palabras clave:</strong> <span property="keywords">NoSQL, Bases de datos,
  Arquitectura en bases de datos.</span><br/>
  <strong>Resumen:</strong>
  <span property="description">Reflexiona sobre las bases de datos NoSQL,
  <strong>Afiliación:</strong> <span property="sourceOrganization">UPTC</span><br/>
  pp.<span property="pageStart">21</span>-<span property="pageEnd">32</span>
</div>
```

Figure 3. *Author's Data.*

Figure 4 shows the journal data ranked.

```
<strong>Revista:</strong>
<div property="isPartOf" typeof="PublicationIssue" resource="#issue">
    <span typeof="Periodical" resource="#periodical">
      <span property="name">Revista Ingenieria UPTC</span>,
    </span>
    <span property="datePublished">2012</span>,
    Vol.<span property="isPartOf" typeof="PublicationVolume"><link
                property="isPartOf" resource="#periodical" /><span
                property="volumeNumber">21</span></span>(<span
            property="issueNumber">33</span>),
</div>
<span resource="#periodical">
    <strong>Publisher:</strong> <span property="publisher">UPTC</span><br />
    <strong>ISSN:</strong> <span property="issn">0121-1129</span> ;<br />
    <strong>E-ISSN:</strong> <span property="issn">2357-5328</span> ;<br />
</span>
```

Figure 4. *Journal data.*

In Figure 5, there is a reference to another article.

```
<strong>Referencias:</strong>
  <div property="citation" typeof="ScholarlyArticle">
    <span property="author">N. Leavitt</span>
    "<span property="name">Will NoSQL Databases Live Up to Their Promise?,</span>"
    <div property="isPartOf" typeof="Periodical">
    <em><span property="name">Computer</span></em>
  </div>
  <span property="isPartOf" typeof="PublicationVolume">
    vol. <span property ="volumeNumber">43</span>
  </span>,
  <span property="isPartOf" typeof="PublicationIssue">
    no. <span property="issueNumber">2</span>
    (<time datetime="2010-02" property="datePublished">Febrero 2010</time>):
  </span>
  <span property="pageStart">12</span>-<span property="pageEnd">14</span>
</div>
```

Figure 5. *Reference data.*

The indexation process is very simple once the model is ready. It is important that metadata have the mentioned characteristics and that the user can have a graphic interface or an automated media for registering these data. For example, Twitter is a user that stands out the metadata and the relations inadvertently, and if some codes as ISSN (International Standard Serial Number) and DOI (Digital Object Identifier) were used, the process would be automated to include the references.

6. Conclusions

The union of Semantic Web technologies along with the knowledge of the users can contribute to improve the quality in internet searches, but it is necessary using the existing standards for not to fall into the construction of a "Tower of Babel" of indexation of technologies.

Metadata are a way to manage the information, but it is necessary it integration to the Web in an organized way. In order to do that, there are several technologies and it is urging to use the ones that have more backrest and try to unify efforts to improve the experience on Web.

Finally, it is important to mention that there is still a long path to find a universal way for describing contents, but there are many initiatives in this area like W3C and some particulars like Schema.org. It is time to contribute with the efforts of the research community in order to promote these technologies to reach the final user in a transparent way.

7. Bibliographical references

Allsopp, J. (2007). *Microformats: empowering your markup for Web 2.0.* Berkeley: Apress.

Antwerp, M. V., & Madey, G., (2010). Warehousing and studying open source versioning metadata. In: Ågerfalk, P.J., Boldyreff, C., González-Barahona, Madey, G.R., & Noll, J. (eds.), *Open Source Software: New Horizons: 6th International IFIP WG 2.13 Conference on Open Source Systems* 319, 413-418.

Audit Commission Publishing Team (2007). *Improving information to support decision making: standards for better quality data,* Millbank. London: Wales Audit Office.

Baca, M., (2008). *Introduction to metadata: second edition.* Los Angeles: Getty Research Institute.

D. C. Initiative (2012). *Dublin Core Metadata Element Set. Version 1.1*. http://dublincore.org/ documents/dces/

Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann M., & Horrocks, I. (2000). The Semantic Web: the roles of XML and RDF. *IEEE Internet Computing* 4 (5), 63–73.

Doctorow, C., (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia. http://www.well.com/~doctorow/metacrap.htm. Duval, E., & Hodgins W. (2002). Metadata principles and practicalities. *D-Lib Journal* 8 (4), 2002.

García, F., (2013). Schema.org: la catalogación revisitada. *Anuario ThinkEPI* 7 (1), 169-172.

George, R. P., (2006). *Scaling the technology opportunity analysis text data mining methodology: data extraction, cleaning, online analytical processing analysis, and reporting of large multi-source datasets.* Ph.D. Dissertation. Capella University, Minneapolis, Estados Unidos.

González, R., (2014). *RDF 101, Cambridge Semantics.* http://www.cambridgesemantics.com/ semantic-university/rdf-101.

González, B., Guerrero, V. P., & Moya, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics* 4 (3), 379–391.

ISO/IEC 11179-1 (2004). *Information technology metadata registries (MDR).* http://metadata- standards.org/11179/

Khare, R., & Çelik, T., (2006). *Microformats: A pragmatic path to the semantic Web.* Proceedings of the 15th International Conference on World Wide Web. New York, NY, USA, p. 865–866.

Khare, R., (2006). Microformats: the next (small) thing on the semantic Web? *IEEE Internet Computing* 10 (1), 68–75.

Manola, F., & Miller, E. (2004). RDF Primer. *W3C recommendation* 10 (1), 1-107.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., & Kleese, K. (2010). Using a core scientific metadata model in large-scale facilities. *International Journal of Digital Curation* 5 (1), 106–118.

Microformats community (2014). *Getting started with microformats 2*. http://microformats. or g/2014/03/05/getting-started-with-microformats2.

Mika, P., & Potter, T. (2012). *Metadata statistics for a large web corpus.* In C. Bizer, T. Heath, T. Berners-Lee & M. Hausenblas (eds.), LDOW 2012: Linked Data on the Web: CEUR-WS.org.

National Information Standards Organization (2004). *Understanding metadata.* http://www. niso.org/publications/press/UnderstandingMeta-data.pdf.

Pabitha, P., Vignesh, N. K., Pandurangan, N., Vijayakumar, R., & Rajaram, M. (2011). Semantic annotation of wiki using wiki markup for Html5 Microdata. *IJCSI International Journal of Computer Science and Technology* 8

(1), 388-394.

Pastore, S. (2012). Website development and web standards in the ubiquitous world: Where are we going? *WSEAS Transactions on Computers* 11 (4), 309-318.

Qin, J., & Li, K., (2013). *How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure.* International conference on Dublin core and metadata applications. Lisboa, Portugal, p. 25–34.

Ronallo, J., (2012). HTML5 Microdata and Schema.org. *The Code4Lib Journal* (16).

Scott, D., (2013). *Microdata: making metadata matter for machine.* https://zone.biblio.laurentian. ca/dspace/bitstream/10219/1993/3/microdata_ matters.pdf.

Shadbolt, N., Berners, T., & Hall, W., (2006). The semantic Web revisited. *IEEE Intelligent Systems* 21 (3), 96–101.

Suda, B. (2006). *Using Microformats.* O'Reilly Media, Inc.

Tomberg, V. & Laanpere, M. (2009). *RDF a versus Microformats: exploring the potential for semantic interoperability of mash-up personal learning environments.* International workshop on Mashup Personal Learning Environments (MUPPLE) 506, 102-109.

Whatwg Editors (2014). *HTML living standard.* https://html.spec.whatwg.org/.

Yang, E., Matthews, B., & Wilson, M., (2010). *Enhancing the core scientific metadata model to incorporate derived data.* IEEE Sixth International Conference on e-Science (e-Science), Brisbane: Australia, p. 145–152.