# Surfacing Scientific and Financial Data with the Xcel2RDF Plug-In

Marcia Lucas Pesce, Karin Koogan Breitman, Marco Antonio Casanova
*Departamento de Informática, PUC-Rio*
*Fundação Getúlio Vargas*
*Rio de Janeiro/RJ – 22453-900 - Brasil*
*Marcia.pesce@fgv.br, {karin,Casanova}@inf.puc-rio.br*

***Abstract***

Given the astounding amount of data stored in spreadsheets and relational databases, a critical requirement for the evolution of the Semantic Web (SW) is the ability to convert data to SW compatible formats, such as RDF and OWL. The process by which data is transformed into RDF is known as triplification. This paper introduces Xcel2RDF, an MS Excel plug-in to support the triplification of spreadsheets, which minimizes the learning curve, as it is integrated into a widely used spreadsheet software tool. The plug-in is user-friendly, does not depend on the installation of additional software and does not require the user to leave his familiar environment, thereby avoiding problems reported as the major drawbacks of existing spreadsheet to RDF conversion tools. Finally, as a proof of concept, the paper illustrates how to use the tool to triplify statistical data.

***Categories and Subject Descriptors***
***D.2.2 [Software Engineering]: Design Tools and Techniques***

***General Terms***
***Design***

***Keywords***
***Spreadsheet, RDF, Excel, Linked Open Data***

## [1] INTRODUCTION

The term *Deep Web* refers to Web content that cannot be directly indexed by search engines. Studies showed that Deep Web content is particularly important. However, to obtain such content is challenging and has been acknowledged as a significant gap in the coverage of search engines [12].

Indeed, much of the information hidden in Web sites is dynamically generated from relational databases or spreadsheets, and search engines are not able to find them. Estimates suggest that the volume of data stored in data silos greatly exceeds that of the Surface Web – with nearly 92,000 terabytes of data on the Deep Web versus only 167 terabytes on the Surface Web, as of 2003 [10].

In particular, a large part of scientific and business data today is captured using Excel spreadsheets, and remains inaccessible to search engines [17]. Typically, one provides a wrapper that allows users to query datasets and their attributes, e.g., geographical location (states, regions), temporal data (month, quarters). The problem with this approach is that the data is only accessible to the few users that are aware of it, but hidden to users at large.

The late Jim Gray stressed the necessity to surface Deep Web data, i.e., to make it visible to search engines and, thus, largely findable [18]. The Linked Data approach, based on the Semantic Web stack of standards and technologies, provides a framework in which to publish, query and consume data in the Web [1]. The RDF standard, in particular, is very useful for data surfacing, for it provides a "*lingua franca*" in which heterogeneous datasets can interoperate.

Unfortunately, tools that support the transformation of data stored in the Deep Web to RDF formats are still in their infancy [21]. Most of the tools focuses on the transformation of relational data, as opposed to data stored in spreadsheets, and requires the installation of additional tools and add-ons [4, 19, 20].

The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. It is particularly intended for representing metadata about Web resources, but it can also be used to represent information about objects that can be identified on the Web, even when they cannot be directly retrieved from the Web. To some extent, RDF is a lightweight ontology language to support interoperability between applications that exchange machine-understandable information on the Web.

RDF has a very simple and flexible data model, based on the central concept of the RDF statement. We also consider the concept of vocabulary as part of the RDF data model, due to its relevance to ontology modeling. RDF offers three equivalent notations: RDF triples, RDF graphs, and RDF/XML.

The RDF triples notation translates RDF statements directly into character strings. More precisely, the RDF triple for an RDF statement (S, P, O) is a string of one of the two forms:

<S> <P> <O> . if O is an absolute or relative URIref
<S> <P> "O" . if O is a literal

The RDF triples notation for a set R of RDF statements is simply the concatenation of the RDF triples that represent each RDF statement in R, in any order.

In this paper, we propose an Excel plug-in that supports the transformation of data stored in spreadsheets to RDF. It promotes the reuse of standard RDF vocabularies, to secure interoperability with data published in the Linked Data format. It also promotes usability for Excel users, as they need not to leave their work environment.

The rest of the paper is divided as follows. Section 2 presents related work. Section 3 introduces the Xcel2RDF plug-in and describes the process by which the plug-in produces RDF triples from spreadsheets. Section 4 presents case studies that use financial series. Section 5 contains the conclusions and discusses future work.

## [2] RELATED WORK

There are several tools that claim automatic triplification. However, most of them focuses on the conversion of relational data, such as Triplify [4], Virtuoso RDF views [19] and D2RQ [20].

Tools that triplify spreadsheets are less popular. Table 1 shows an evaluation of such tools with respect to their usability (in a scale of 1-3, 1 being the lowest), license (open source or paid) and RDF vocabularies supported. As can be observed there is no tool that, at the same time, supports multi-dimensional data conversion, is open source and promotes vocabulary reuse and usability.

## [3] THE EXCEL2RDF PLUGIN

*a)* Overview

The Xcel2RDF plug-in provides support to the spreadsheet triplification. A good metaphor for the process is the translation from the spreadsheet model to the RDF model.

Most relational triplification tools maps tables to RDF classes and attributes to RDF properties, with no concern to identifying possible matches with existing standard vocabularies, thereby most often creating new vocabularies unnecessarily. The following table depicts a comparison between existing and the proposed tool. The following criteria is used: usability (tools with poor usability are ranked 1, adequate usability 2, and good usability 3), range of input data formats, cost, and RDF vocabulary support.

SPREADSHEET TRIPLIFICATION TOOLS

| Tools | Usability | Data | Lic. | Vocabularies |
|---|---|---|---|---|
| Stats2RDF | 2 | Multidimensional | No | DataCube Vocabulary |
| Cambridge Semantics' Anzo for Excel | 3 | Bidimensional | Yes | |
| XLWrap | 1 | Multidimensional | No | SCOVO Vocabulary |
| Excel2RDF Plugin (proposed Tool) | 3 | Multidimensional | No | Foaf DataCube Vocabulary Others |

Our background in databases, particularly past experiences with the construction of mediators, schema and ontology matching, convinced us that the use of standards in schema design is the only viable way to guarantee future interoperability [11, 22]. The Xcel2RDF plug-in is anchored on this principle and strives to promote the reuse of standards by implementing a guided, four-step process. The first step consists in choosing the vocabulary to be used, the second step is selecting the data to be converted, the third is assigning semantics to the data and the fourth and final step consists in assigning provenance to the data generated. These four steps are described in the following subsections.

Xcel2RDF was developed using C#, in conformance with Microsoft's add-in guidelines [2]. Figure 1 shows the graphical user interface of the proposed plug-in.

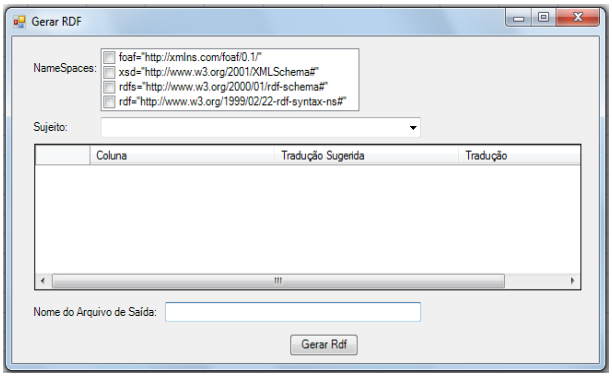


**Figure 1** – Xcel2RDF plug-in graphical user interface

*b)* Step 1: Choice of Vocabularies

We believe that the only way to secure interoperability is by a priori design, i.e., by selecting appropriate standards, if one exists, to guide the design of the data source [11]. The same philosophy is applicable to Linked Data, as stated by Bizer, Cyganiak and Heath "In order to make it as easy as possible for client applications to process your data, you should reuse terms from well-known vocabularies wherever

possible. You should only define new terms yourself if you cannot find required terms in existing vocabularies" [23]. Unfortunately, that is not what happens in practice. Most users prefer creating new vocabularies (as do the vast majority of triplification tools) to spending the required time and effort to search for adequate matches [21]. There are notwithstanding numerous standards that designers cannot ignore when specifying triple sets and publishing their content.

The current version of the plug-in has two standard, built-in RDF vocabularies. The first is FOAF (Friend of a Friend) [8], which is widely used to describe people, their activities and their relationships with other people and objects. The second is the Data Cube Vocabulary [9], often used to describe multidimensional statistical data and metadata. Xcel2RDF, however, is extensible, and allows the adoption of new vocabularies as they become necessary. This happens when there is no existing RDF vocabulary adequate to describe the data in the spreadsheets in question. In this case the user has the opportunity to create a new vocabulary, with the terminology he or she needs to describe the data.

*c)* Step 2: Selection of Data

Spreadsheets contain both data and metadata. Typically, the top rows and leftmost columns contain metadata, i.e., concepts (what the spreadsheet is about) and the center contains the data (instance values). During the publication process, it is very important to separate concepts from values, for concepts need to be represented using RDF vocabularies. The mapping of spreadsheet concepts to RDF vocabularies anchors the semantics of the triple set to be created.

In the Xcel2RDF plug-in, the separation of concepts from the data is guided by the user. Using the mouse-over Xcel GUI, he or she selects the range of data values from the spreadsheet that are to be converted. Internally, the plug-in assumes all rows above and to the left of the selection to be headers. The headers are then fed to the mapping wizard, which assists the user in the next step (finding adequate mappings of row/column concepts to RDF vocabularies).

It is important to note, however, that spreadsheets are often used to represent multi-dimensional data, i.e., as bi-dimensional representations of aggregated data. Figure 2 illustrates this point:

- Lines 2 and 3 define a temporal series (months: row 2; months aggregated by trimester: row 3)
- Columns A and B list Brazilian states, aggregated by region (Sudeste - Southeast region - aggregates the states of Rio de Janeiro, São Paulo, Espírito Santo and Minas Gerais; the Sul - South region - aggregates of the states of Paraná, Santa Catarina and Rio Grande do Sul)

Xcel2RD uses the DataCube vocabulary, whose terms represent *dimensions, attributes and measures,* typical of multi-dimensional data [9].

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | | 1º Trimestre | | | 2º Trimestre | | |
| 3 | | | Janeiro | Fevereiro | Março | Abril | Maio | Junho |
| 4 | Sudeste | RJ | 10 | 11 | 30 | 30 | 18 | 7 |
| 5 | | SP | 20 | 2 | 7 | 12 | 12 | 5 |
| 6 | | MG | 30 | 20 | 12 | 18 | 10 | 15 |
| 7 | | ES | 5 | 10 | 10 | 7 | 5 | 18 |
| 8 | Sul | RS | 10 | 12 | 5 | 15 | 18 | 7 |
| 9 | | SC | 15 | 15 | 10 | 30 | 12 | 30 |
| 10 | | PR | 10 | 7 | 15 | 2 | 10 | 2 |

**Figure 2 -** Data selection step

In the above example the user must indicate that the data he or she is interested in is located in between cells C,4 and H,10. The rest of the information, i.e., located in columns A and B, and rows 1-3, will be used in the identification of headers, possible merges, and groupings. The groupings represent the multiple dimensions of the data.

*d)* Step 3: Assigning Semantics to the Data

Once the rows and columns to be published are decided, the user is assisted in the process of assigning adequate semantics to represent data in terms of existing RDF vocabularies. This process consists in mapping concepts in the spreadsheets, e.g., *Full Name*, to concepts in standard RDF vocabularies, e.g. *foaf:Person*. Hence, one secures that the RDF triples possess the semantics of the data contained in the spreadsheet.

It is important to note that this step is manual. It is sometimes the case that the names used to tag the rows and columns are devoid of semantics, e.g., mnemonics and jargon. Only the owner of the spreadsheet is able to provide the correct semantic mappings for them.

To reduce the effort, however, the choices made in this step will be stored in a separate file, for future use. The idea is that in subsequent conversion processes, the plug-in provides suggestions of possible mappings from previous choices. This feature is part of the functionality package previewed for the next release of the plug-in.

*e)* Step 4: Provenance Allocation

After producing the RDF triples, we include additional metadata and provenance information to facilitate retrieval and future interoperability of the published triple set.

Provenance metadata is central to guarantee information reliability about the people, data sources, collection, and transformation processes the data went through. Provenance metadata is also considered fundamental in securing data quality in several domains such as databases [13], e-Science [14], and workflows [15] where concrete solutions take into consideration the specificities of the domains in question. The plug-in currently supports the inclusion of the date of the triplification, file name used, the data source file, name of the user who performed the triplification, among others.

In addition, Xcel2RDF also captures information on the vocabulary choices made during the triplification process. It

is often the case that there is more than one choice of RDF vocabulary to be used in the process of annotating data. Registering the user choices is a very important step because it helps identify mappings to other vocabularies and schemas. The mappings are useful for data integration, and in the construction of mediators, that will consume the triple set in question.

## IV - THE EXCEL2RDF PLUG-IN

In what follows, we illustrate two case studies, with data from the Getúlio Vargas Foundation (FGV), a Brazilian institution that publishes economical and financial indicators. These are typical examples of data originally produced in spreadsheets. Our choice of case study serves us twofold: to enforce the need for tools that support spreadsheet to RDF conversion, as opposed to relational to RDF, and to demonstrate the usability of the proposed plug-in.

The Xcel2RDF plug-in works within the MS Excel environment, therefore it capitalizes on its general functionality, e.g. cut & paste, thus reducing the learning curve for plug-in use. Additionally, and more importantly, it neither requires the installation of additional software nor the use of unfamiliar software tools.

The first example is a worksheet that contains data on social indicators that inform the proportion of Brazilian household income in four strata: richest 1%, richest 10%, poorest 20%, and poorest 50% (individual distribution according to household income per capita). This dataset is published in a yearly basis. Figure 3 shows data ranging from 1976 to 2009.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Data | Household Income - Participation of the richest 1% - (%) | Household Income - Participation of the richest 10% - (%) | Household Income - Participation of the poorest 20% - (%) | Household Income - Participation of the poorest 50% - (%) |
| 2 | 1976 | 17,08 | 51,04 | 2,43 | 11,58 |
| 3 | 1977 | 18,47 | 51,64 | 2,42 | 11,68 |
| 4 | 1978 | 13,64 | 47,71 | 2,05 | 11,96 |
| 5 | 1979 | 13,61 | 47,45 | 2,67 | 12,77 |
| 6 | 1980 | - | - | - | - |
| 7 | 1981 | 12,67 | 46,4 | 2,66 | 13,14 |
| 8 | 1982 | 13,02 | 46,91 | 2,53 | 12,7 |
| 9 | 1983 | 13,47 | 47,38 | 2,55 | 12,51 |
| 10 | 1984 | 13,19 | 47,27 | 2,76 | 12,99 |
| 11 | 1985 | 13,61 | 47,75 | 2,54 | 12,46 |
| 12 | 1986 | 13,77 | 46,95 | 2,67 | 13,02 |
| 13 | 1987 | 14,11 | 47,75 | 2,36 | 12,22 |
| 14 | 1988 | 14,41 | 49,47 | 2,17 | 11,46 |
| 15 | 1989 | 16,48 | 51,5 | 2,01 | 10,62 |
| 16 | 1990 | 14,2 | 48,78 | 2,14 | 11,45 |
| 17 | 1991 | - | - | - | - |
| 18 | 1992 | 13,23 | 45,78 | 2,36 | 13,11 |
| 19 | 1993 | 15,09 | 48,64 | 2,26 | 12,31 |
| 20 | 1994 | - | - | - | - |
| 21 | 1995 | 13,81 | 47,85 | 2,31 | 12,35 |
| 22 | 1996 | 13,53 | 47,52 | 2,16 | 12,09 |

**Figure 3** – Household income distribution in Brazil – 1976 to 2009 source: Fundação Getúlio Vargas

The triplification of this dataset, using the Xcel2RDF plug-in, took 2 seconds. Figure 4 illustrates some of the output triples.

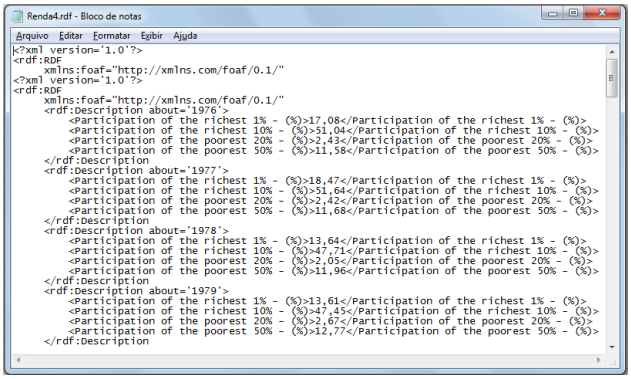


**Figure 4** – Example RDF triples from the conversion of the Household income distribution in Brazil spreadsheet

The second spreadsheet contains the US dollar daily exchange rate, determined by the Brazilian Central Bank (BCB). The period covered by this dataset includes significant changes to the exchange rate policy, including periods where the rate was fixed by the BCB and others where the rates freely fluctuated. The dataset is updated every business day. The dataset is this case study covers a period a little over fifteen years, ranging from July 1994 to Feb. 12, 2012, as illustrated by Figure 5.

| | A | B |
|---|---|---|
| 1 | Data | Commercial Dollar (PTAX) - Daily Exchange Rate - Values in BRL |
| 2 | 01/07/1994 | 1 |
| 3 | 04/07/1994 | 0,94 |
| 4 | 05/07/1994 | 0,932 |
| 5 | 06/07/1994 | 0,915 |
| 6 | 07/07/1994 | 0,91 |
| 7 | 08/07/1994 | 0,92 |
| 8 | 11/07/1994 | 0,925 |
| 9 | 12/07/1994 | 0,92 |
| 10 | 13/07/1994 | 0,92 |
| 11 | 14/07/1994 | 0,925 |
| 12 | 15/07/1994 | 0,935 |
| 13 | 18/07/1994 | 0,935 |
| 14 | 19/07/1994 | 0,935 |
| 15 | 20/07/1994 | 0,932 |
| 16 | 21/07/1994 | 0,932 |
| 17 | 22/07/1994 | 0,937 |
| 18 | 25/07/1994 | 0,936 |
| 19 | 26/07/1994 | 0,934 |
| 20 | 27/07/1994 | 0,936 |
| 21 | 28/07/1994 | 0,94 |

**Figure 5** – Dollar exchange rate in Brazil – July 1976-February 2012. source: Fundação Getúlio Vargas

The triplification of this dataset, using the Xcel2RDF plug-in, took 48 seconds. Figure 6 illustrates some of the output triples.

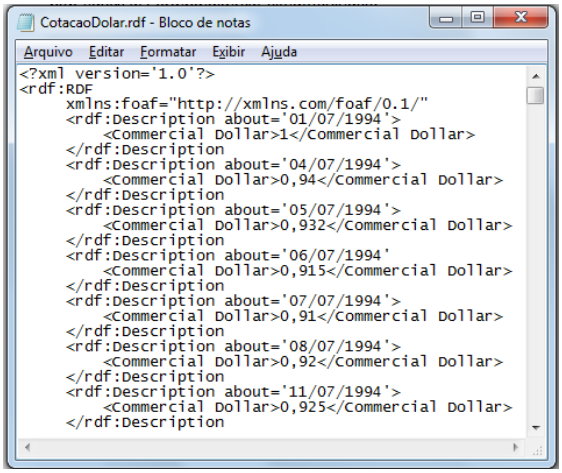


**Figure 6** - Example RDF triples from the conversion of the Dollar exchange rate in Brazil

The above cases studies indicate that, in order to generate reliable and error free triples, it is fundamental to count with high quality spreadsheets. Ideally, before starting the triplification process, the spreadsheets would undergo a verification and, if needed, a sanitation process. The triplification process, it is important to note, does not introduce errors, but rather faithfully represents the original data.

## V – TECHNICAL CHALLENGES

Among the difficulties we encountered while developing the proposed plug-in, we would like to call attention to the complexity associated to the identification of groupings. In the case of multi-dimensional spreadsheets, information may grouped using several distinct criteria, both in lines and columns, generating a very large of possible combinations.

RDF vocabulary reuse is highly advisable, to promote interoperability and to help identify links to additional resources in the Linked Open Data cloud. This, of course, can only be achieved with adequate libraries and tool support to help identify possible matches to existing vocabularies. We worked with a few, domain specific vocabularies, but future plans include the incorporation of an ontology matching tool and the construction of an RDF vocabulary library.

Finally, our greatest challenge was the development of an algorithm that takes into consideration the structure, i.e., and the way data was organized in the spreadsheets, to enhance performance in tasks related to grouping identification and discovery.

## VI – CONCLUSION

The use of Semantic Web standards, RDF in particular, is important to secure interoperability in the Web of Data.

Triplification tools, however, still lack usability and extensibility.

In this paper, we introduced Xcel2RDF, an Excel plug-in that enables non-experts to triplify spreadsheets. The resulting RDF triple sets may then be used for different purposes, such as the creation of data mashups, i.e., the merge of data from different data sources, in order to produce comparative views of combined data [16].

Xcel2RDF is extensible and offers the possibility of including additional support to other RDF Vocabularies. This is particularly interesting to facilitate interoperability.

We believe that the development of tools such as Xcel2RDF promotes interoperability, fosters the democratization of information and represents an important advance in the popularization and adoption of Semantic Web standards.

## ACKNOWLEDGMENT

This work was partly supported by CNPq, under grants 301497/2006-0, 305824/2010-4 and 475717/2011-2, and by FAPERJ under grants E-26/170028/2008.

## REFERENCES

[1] Breitman, K., Casanova, M.A., and Truszkowski, W. *Semantic Web: Concepts, Technologies and Applications*. London: Springer, 2006. v. 1. 337 p.

[2] Microsoft Excel Add-in documentation. Available at: http://msdn.microsoft.com/en-us/library/aa140990(v=office.10).aspx

[3] Salas, P., Breitman, K., Viterbo, J. and Casanova, M.A. Interoperability by Design Using the Std-Trip Tool: an a priori approach. In: Proc. 6th International Conference on Semantic Systems 2010 (ISEMANTICS' 10).

[4] Auer, S., Dietzold, S., Lehmann, J., Hellmann, S. and Aumueller, D. Triplify: lightweight linked data publication from relational databases. In: Proc. 18th international conference on World Wide Web. Madrid, Spain: ACM. (2009): 621-630.

[5] Stats2RDF tool. Available at: http://aksw.org/Projects/Stats2RDF.

[6] Cambridge Semantics Anzo Express tool. Available at: http://www.cambridgesemantics.com/products/anzo-express.

[7] XLWRap tool. Available at: http://xlwrap.sourceforge.net/.

[8] FOAF vocabulary. Available at: http://xmlns.com/foaf/0.1/.

[9] DataCube Vocabulary. Available at: http://dvcs.w3.org/hg/gld/raw-file/default/data-cube/index.html

[10] Casanova, M.A., Breitman, K., Brauner, D.F. and Marins, A. Database Conceptual Schema Matching. Computer (Long Beach), v. 40 (2007): 102-104.

[11] He, B. et al. Accessing the Deep Web. Comm. of the ACM 50(5): 94–101.

[12] Lu, J. et al. Learning Deep Web Crawling with Diverse Features. In: Proc. 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '09 (2009): 572–575.

[13] Buneman, P. and Tan, W.-C. Provenance in databases. In: Proc. 2007 ACM SIGMOD International Conference on Management of Data, New York, NY, USA. ACM Press, (2007): 1171-1173.

[14] Simmhan, Y.L., Plale, B. and Gannon, D. A survey of data provenance in e-science. SIGMOD Record 34(3): 31-36.

[15] Omitola, T., Gibbins, N. and Shadbolt, N. Provenance in Linked Data Integration. In: Future Internet Assembly, Ghent, Belgium (2010): 16-17.

[16] Accar, S., Alonso, J. and Novak, K. (eds.) Improving Access to Government through Better Use of the Web. W3C Interest Group, 12 (May 2009).

[17] Bell, G., Hey, T. and Slazay, A. Beyond the Data Deluge Science (March 2009): 1297-1298

[18] Hey, Tansley, Tolle (Eds) The fourth Paradigm - Microsoft Research (2009).

[19] Erling, O. and Mikhailov, I. RDF support in the Virtuoso DBMS. In: Proc. 1st Conference on Social Semantic Web, volume P-113 of GI-Edition - Lecture Notes in Informatics (LNI), Bonner Kollen Verlag (Sept. 2007).

[20] Bizer. C. and Seaborne, A. D2RQ - treating non-RDF databases as virtual RDF graphs. In: ISWC2004 (posters), (Nov. 2004).

[21] Kinsella, S., Bojars, U., Harth, A., Breslin, J.G. and Decker, S. An Interactive Map of Semantic Web Ontology Usage. In: Information Visualisation, 2008. IV '08. 12th International Conference (9-11 July 2008): 179-184.

[22] Leme, L.A.; Casanova, M.; Breitman, K.; Furtado, A. OWL schema matching. J. Braz. Comp. Soc. – Springer Verlag 16(1): 21-34.

[23] Bizer, C., Heath, T., Ayers, D. and Raimond, Y. "Interlinking Open Data on the Web"; Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria (May 2007).