

ORGANISING KNOWLEDGE: OK! WHAT OUR CATALOGUES AND METADATA HAVE TO DO WITH THE SEMANTIC WEB AND LINKED DATA

Simon Rooks (*Multi-Media Archivist, BBC*) & Guy Maréchal (*Senior Adviser: Titan & Memnon*)

1. Introduction

At the 2009 IASA Conference in Athens, the Cataloguing and Documentation Committee was re-worked as the 'Organising Knowledge' (OK) Task Force. The label 'Organising Knowledge' was engendered by Chris Clark (British Library) whose presentation, with that of Ingrid Finnane (National Library of Australia), amounted to a call to adopt a new perspective on how we create, enhance, manage, link and share metadata about our collections and, crucially, to understand and harness the possibilities of the semantic web. There is an array of sometimes bewildering techniques and practices now widely established including linking data, tagging, user comments, collection building, recommendation and rating. What can they offer institutions such as those represented in IASA, and how do we effectively utilise our knowledge and expertise?

Beyond inspirational papers and a new label, what is next? Those responsible for cataloguing and documentation in established institutions face major challenges in this area. One is to understand better the landscape of resource discovery, navigation, knowledge as brands and products, user behaviours and how contextualising our metadata as knowledge can promote discovery. We need also to engage with leaders in these fields who look curiously and sometimes hungrily at our professionally constructed datasets and aspire to unlock their value.

Following presentations by Guy Maréchal at the 2010 IASA Conference, the Executive Board asked that the OK Task Force progress in these areas with the aim of one or more tutorials at the 2011 IASA in Frankfurt, covering both conceptual and technical issues. Planning for the Frankfurt sessions is well under way and the OK Task Force has prepared several proposals for the Organizing Committee which may be found in the program either as a paper presentation or a tutorial:

- Introduction to the semantic technologies and Linked Open Data (by Guy Maréchal [Senior adviser, Titan & Memnon])
- Opportunities and needs of the semantic technologies and taxonomies for the cultural sector (by Fran Alexander [Taxonomy Manager, Information and Archives, BBC])
- Easy empowering of your cultural data into linked, enriched and structured semantic assets (by Guy Maréchal)
- The migration strategy to reach persistence in small and medium collections (by Guy Maréchal)

2. Illustration of the change by way of a simple example

The usual way of cataloguing and documenting of media assets is to fill in a metadata template for each of the assets and then to store it in a database. The list of metadata depends on the nature of the asset (book, sound recording ...), of its cultural domain and other classification and sector rules. Three of the entries are very general: the 'name of the assets', the 'name of the contributor' and the 'hyperlink to the file' representing the asset.

The well known "*Eine kleine Nachtmusik*" has been composed by Mozart. According to the XML, Dublin Core and METS syntaxes, these metadata could look like:

```
<dc:name>Eine kleine Nachtmusik</dc:name>
<dc:contributor>Mozart</dc:contributor>
<mets:file ID="FILE_W002" ADMID="TMD_W002" MIMETYPE="audio/wav"
GROUPID="GW003" SIZE="1" CHECKSUMTYPE="MD5" CHECKSUM="the_md5_file_
checksum here">
```

```
< mets:FLocat LOCTYPE="URL" xlink:href="file://root/path/subdir/
S_2069-B-01-W3.ogg" />
</mets:file>
```

Anybody with a minimum of music knowledge will understand that it is meant that the composer is Wolfgang Amadeus Mozart [1756 – 1791] and that the music involved is the usual name of the Serenade identified K.525! Everybody should also forget about the hyperlink and simply assume that a file coded in the “ogg” format is available representing the audio recording.

From the Information Technology perspective it is precisely the reverse: “Mozart” is simply and not more than a string of characters, and “Eine kleine Nachtmusik” another one! But, through the complex hyperlink, the IT has what is required for presenting you with the beautiful sound of Mozart’s music!

The fundamental intention of the semantic technologies is to ensure, by construction, the **interoperability** of applications and navigations through the expression of the relations existing between representations of concepts and their instances with their characteristics - or as Tim Berners Lee put it: the task is to provide “*information that has well-defined meaning, hence better enabling computers and people to work in cooperation*” (Berners-Lee, Hendler and Lassila, 2001). That representation is usually expressed according to a combination of standards languages (using the XML syntax) of the W3C, in particular, the RDF [Resource Description Framework] and the OWL [Ontology Web Language]. The RDF is a general-purpose language for representing information in the Web. The OWL language ensures the definition of the ontologies and of the instances of the classes. A specific language and protocol has been standardized for the searches: SPARQL.

A specialized textual syntax has been designed for expressing the instances of the triples. It is called “Turtle”. It allows RDF graphs to be completely written in a compact and natural text form, with abbreviations for common usage patterns and datatypes. Turtle provides levels of compatibility with the existing N-Triples and Notation 3 formats as well as the triple pattern syntax of SPARQL. Obviously, the final users are not exposed directly to these languages - the Graphical User Interfaces hide them, and radically new ways of navigation and querying are emerging which are user friendly.

The semantic technologies allows keeping the current representations according to the usual cataloguing and documentation rules expressed using the well known DC / MARC / MODS / ... models [collectively referred to as “**Flat**” models]. The semantic models [collectively referred as “**Rich**” models] can hook and integrate the ‘Flat’ models.

In the example, for the semantic technologies, the representation of Mozart is a resource, being an instance of the class of things called “Physical person”. Figure 1 illustrates the approach.

The rectangles represent “Resources”, being identified. The upper rectangle has the class “Physical person”. The relation “**is an instance of**” is expressed by the green arrow.

The middle rectangle represents the resource carrying the representations and properties of Mister Wolfgang Amadeus Mozart as an instance of the class “Physical person”. It ‘owns’ the lower rectangle representing the existence of the resource and its associated properties, including the relation “is an instance of”, linking it to the class “Physical person”. The instance inherits all of the characteristics of “Physical person”.

The other lower rectangles represent the files representing Mozart: in the example, the .xml file could carry the classical ‘Flat’ model according to the Dublin Core of the structural representation of his life (Date of born; ... ; marriage; ... death); the .odt file could carry a bibliography; the .jpg file could carry the scan of a painting representing him; ...

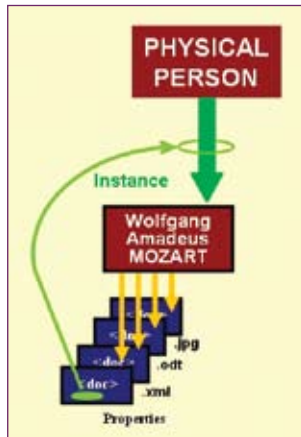


Figure 1: Illustration of the relation 'Is an instance of'

Any of the relations could, through the Web, link data present in distinct databases: this is what is called "Linked Open Data" [LOD]. The OWL definition of the class "Physical person" is in one semantic database [FOAF for example] while its instances, including you and "Wolfgang Amadeus Mozart" who could be described in 27 independent semantic databases linked by LOD and aliases. A network of related data is constructed.

The same construction could be used for expressing the process of Mozart "Composing 'Eine kleine Nachtmusik'". It could be said this is an instance of a resource of the class "Event".

Similar methodology is applied for constructing other types of relations, like expressing that the resource "Composer" inherits the characteristics of "Role" through the relation "specialises". Figure 2 illustrates an excerpt of a possible semantic modelling of the example.

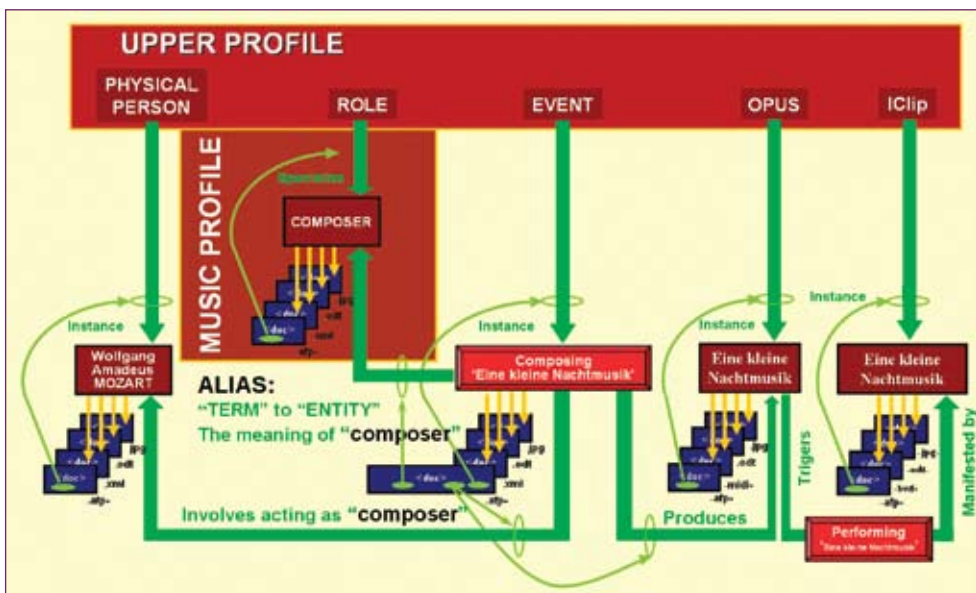



Figure 2: Illustration of semantic modelling



The set of very general classes is called “Upper profile”. The possibility of expressing sets of classes dedicated to specialized domains is illustrated by the “Music profile”. The resource “Performing ‘Eine kleine Nachtmusik’” could also be an instance of the class “Event” representing the performance and its recording in a concert hall. The performing event will produce the resource, being an instance of the class “Logical clip” to which the audio recordings — represented by a set of files coded according to the Broadcast Wave Format [BWF] — and the ‘Flat’ set of metadata according to the usual cataloguing rules can be attached.

Notice that the link to “Composer” is more complex than other links and that through ‘alias’ the Flat models could be combined with Rich models.

It is very important to notice that the representations of a physical resource (a painting, for example) are themselves resources (a JPEG file, for example) identified independently of the identification of the physical resource.

A very interesting introduction to the Linked Open Data is available as a video conference by Tim Bernes Lee at http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html. The video can also be downloaded in .mp4 format.

3. The layers of representation of knowledge

The representation of knowledge has to be seen independently from the point of view of humans and of the ICT [Information and Communication Technologies]. Each of the levels could be empowered by the next higher levels: human also by a higher cultural and social education and ICT by training, trials, validations or human error corrections.

- 3.1 **Textual representation** The textual expression of knowledge is very powerful for producing knowledge and for accessing by humans. The interoperability is ensured between individuals sharing the same culture, the same language and having common social repositories and locators. The complexity and the richness of the grammar rules, of the syntaxes, of the poetry; the voluntary multiple, evocative and ambiguous meanings; the games between sound and sense; . . . ; open doors to utterances that are above knowledge. These expressions can be stored in a persistent manner with no loss of information, but they are not normalised and have poor precision and recall capabilities. This is the level of the Web-1.
- 3.2 **Tagging** The tagging has a low threshold. In most cases it is sufficient and even fun for the single human user. The interoperability is ensured between individuals and machines through simple standards. It offers moderate precision on large databases but remains with a poor precision on the meanings attached to the tags. The control of the consistency is limited. This is the level of the Web-2.
- 3.3 **Taxonomies and Thesaurus** This level offers a very high precision but, by nature, is long, difficult and tedious to maintain and is hardly scalable. The interoperability is ensured as long as no changes occur. Level 3, combined with level 1 could be very powerful. This is the level of ‘Web-2’ with data mining enrichment tools. Retrieval services like Google, Yahoo have demonstrated the power but simple searches could generate thousands of hits.
- 3.4 **Semantic** The semantic expressions of knowledge are very powerful for producing or accessing knowledge by ICT, but the modes of representation of the knowledge, in a way suitable for human understanding, remain a research area. The precision and scalability are without limits. The interoperability is ensured for all the situations where formal modelling could apply. The recall and retrieval is optimum: the thousands of hits of level 3, become focused to only pertinent and serendipity hits. In concrete trials in large semantic databases, we have often obtained only 30 replies (with 20 or more pertinent), while for the same searches at level 3, millions of replies

were frequent. Navigation in semantic databases and LOD is fun and simple. This is the level of Web-3.

- 3.5 **Operational** The semantic opens the door to the capacity of computation, inference and operations through 'intelligent' agents. The associated technologies are partly available and already in use in targeted domains.

4. The fundamentals of the semantic web

In 2001, Tim Berners-Lee et al. introduced their vision of the Semantic Web, as an extension of the current Web, in which information has "*well-defined meaning, hence better enabling computers and people to work in cooperation*" (Berners-Lee, Hendler and Lassila, 2001). The most essential part of this next generation Web is content that is formally described via ontologies, metadata conforming to these ontologies, logic, and agents (Antoniou and van Harmelen, 2004). Many definitions of the term ontology exist. The most popular is by Gruber who defines an ontology as "an explicit specification of a conceptualization" (Gruber, 1993). This definition is further extended by Studer et al. to "formal, explicit specification of a shared conceptualization" (Studer, Benjamins and Fensel, 1998). *Conceptualization* refers to an abstract model of some part of the world which identifies the relevant concepts and relations between these concepts. Explicit means that the type of concepts, the relations between the concepts, and the constraints on their usage, are explicitly defined. *Formal* refers to the fact that the ontology should be machine readable. Finally, '*shared*' means that the ontology should reflect the understanding of a community and should not be restricted to the comprehension of only some individuals. By doing so, it captures consensual knowledge (Fensel, 2003). Ontologies occur in different degrees of formality, ranging from thesauri to richly axiomatic structures (McGuinness, 2003).

A huge momentum has recently been gained in Semantic Web research by the ongoing implementation of a vision of a *Web of Data* formulated by Tim-Berners Lee in which formerly fragmented data is connected and interlinked with each other based on the so-called Linked Data principles [Linked Data Principles <http://www.w3.org/DesignIssues/LinkedData.html>]. The so-called Linked Open Data (LOD) cloud, which represents a huge interconnected data set, has been steadily growing over the past few years.

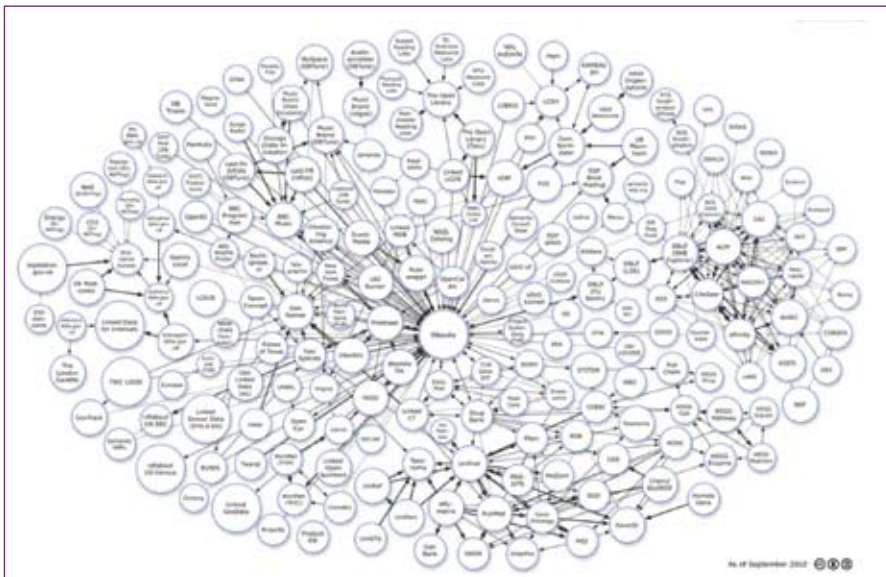


Figure 3: The linked data bubble

In early 2007 the LOD community project was launched within the W3C Semantic Web Education and Outreach group. It bootstraps the *Web of Data* by publishing datasets using the Resource Description Framework (RDF), the metadata model primarily used on the Semantic Web. RDF enables automated software to store, exchange, and use machine-readable information distributed throughout the Web, in turn allowing users to deal with the information with greater efficiency and certainty. Currently, the LOD project includes more than 200 different datasets, ranging from rather centralized ones, such as DBpedia, a structured version of Wikipedia, to those that are very distributed, for example the FOAF-o-sphere. The current LOD cloud contains data from diverse domains such as people, companies, books, scientific publications, films, music, television and radio programs, genes, online communities, statistical or scientific data (Bizer, Heath and Berners-Lee, 2009). Datasets were contributed both by researchers as well as by industry.

The *key success factor* of the LOD movement is the simplicity of its underlying principles:

1. All items should be identified using URIs (Uniform Resource Identifiers);
2. All URIs should be dereference able;
3. When looking up an URI, it should lead to more (linked) data;
4. Links to other URIs should be included in order to enable the discovery of more data.

The cornerstone of the Web is the systematic way of writing hyperlinks to Web pages using a uniformed syntax and protocol. The pages receive an URL [Uniform Resource Locator] and the protocol is "HTTP" [HyperText Transfer Protocol]. But resources can also be uniquely named independently of their location: this is the URN.

The URI concept relates to both ways of identifying resources. Clever ways of organising the naming and locating of resources have been elaborated with structuring and universality. An interesting example is the 'Cool URIs' concept see: [<http://www.w3.org/Provider/Style/URI>]. In some cases, the use of resolvers allows the management of the links between the URN and the URL of a unique semantic resource.

5. The impact on cultural organisations and on archives

- 5.1 **General** The current cataloguing and documentation rules are at level 3 of the representation of knowledge. This has all its advantages but also its limitations as introduced at section 3. One of its main advantages is that it constructs a **hierarchical** structure. The semantic level constructs a structure. This means that archiving becomes complex! Isolating a consistent set of resources implies defining consistent scissoring rules which are beyond the scope of this introductory paper but will be presented during the tutorials proposed for the IASA conference in Frankfurt (September 2011). At the semantic level, archiving and operations have to be disjointed. In particular, the implementations of the OAIS model have to include an extra persistence protocol. The SIP, AIP and DIP constructs and the PI and PDI could fuse into one representation model called "Autonomous Semantic Object".
- 5.2 **Structuring between the assets** The semantic approach allows very easy implementation of highly powerful Conceptual Reference Models such as the FRBR, Cidoc-CRM, FRBR-oo. They also allow the links between logical and physical resources and the modelling of the processes (as represented by arrows in FRBR).

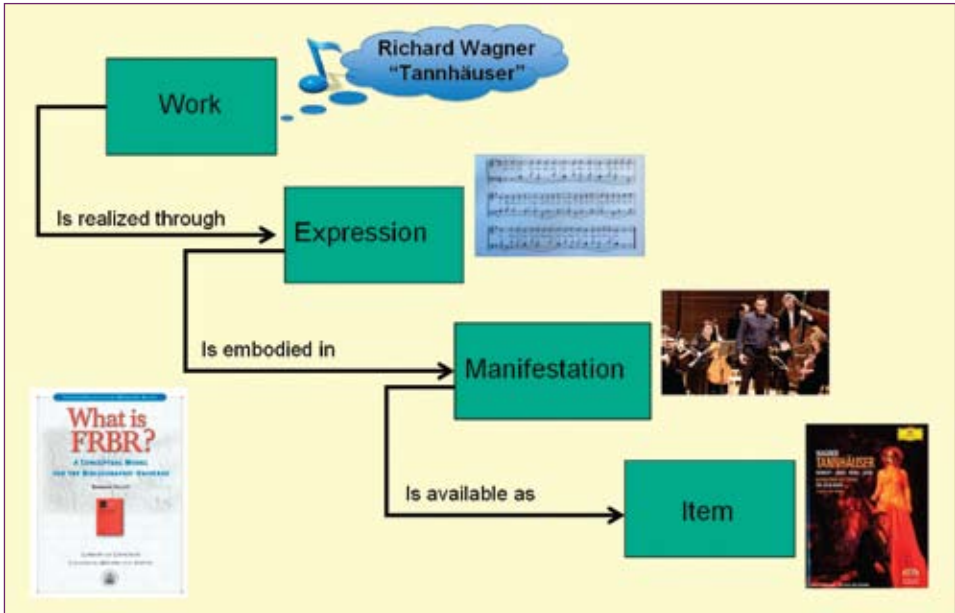


Figure 4: Illustration of the FRBR model

5.3 **Structuring within the assets** The capacity of structuring within the assets is one of the most innovative and powerful capabilities of the semantic approach. The W3C standard, called “fragments”, has defined a normalised way of expressing it: each fragment within one media asset can be defined as a resource. This is already in operation for the semantic modelling of TV news and of Interviews: half an hour of News could become 7,000 ESE [Elementary Semantic Elements]. They structure the news by subjects; they attach and synchronise the transcriptions of what is said and their translations; they identify the speaking persons or presentations on the video; they annotate according to taxonomies and thesauri; they construct exports according to international standards such as NewML-G2 and many other possibilities.

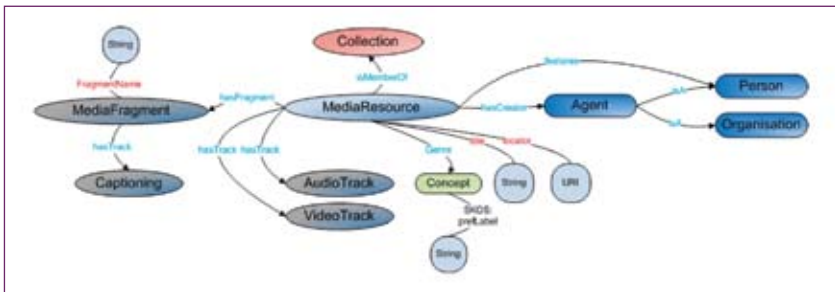


Figure 5: Example of RDF representation with fragments

5.4 **Enhancements and Enrichments** The ‘Flat’ models can be represented in semantic databases without changing them but by giving them access through semantic searches: that process is called ‘**enhancement**’. In turn, they can be enriched (as said in the News example) by structuring, finding and creating LODs, by transcriptions, translations, synchronisations and other ‘**enrichments**’.

A typical example of such a process has been implemented by Memnon Archiving Services in its IPI-Solutions cluster of functions plugged-in to its semantic database ISIS.

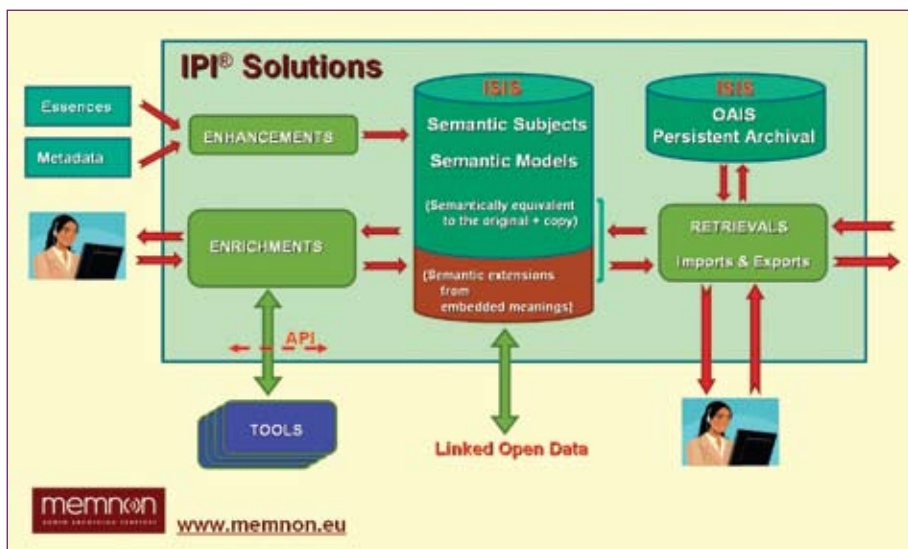


Figure 6: Example of a 'Semantic system' with enhancements and enrichments

5.5 Configuration and Rights management of the assets The management of the existence, states and stages of the assets of their archival, exchanges, sharings, destructions, releases and similar, can be implemented using the semantic approach. A Conceptual Reference Model for Configuration and Rights management is currently being finalised under the acronym: AXIS-CRM.

Acknowledgements

The authors would like to thank Jean-Pierre Evain [evain@ebu.ch] and Tobias Bürger [tobias.buerger@gmail.com] who gave them the authorisation to include excerpts of their paper, 'Semantic Web, linked data and broadcasting' in *EBU Technical Review – 2011 Q1*. It concerns the section 4 and figure 5.

They would like also to thank Roger Roberts [rro@rtbf.be president of the non-profit association TITAN] and Michel Merten [michel.merten@memnon.eu CEO of Memnon Archiving Services] for their continuous involvement and commitment to the research and implementation of products and services suitable for producing 'native semantic' contents and for 'persistent archiving' in a semantic context.

Bibliography

- Evain, JP and T. Bürger. 2011. 'Semantic Web, linked data and broadcasting'. *EBUT Review – 2011 Q1*.
- Berners-Lee, T., J. Hendler, and O. Lassila. 2001. 'The semantic web' in *The Scientific American*, 284(5):28-37, May.
- Antoniou, G. and F. van Harmelen. 2004. *A Semantic Web Primer*. MIT.

Gruber, T. R. 1993. 'A translation approach to portable ontology specifications' in *Knowledge Acquisition*, 5(2):199-220, June.

Studer, R., R. Benjamins, and D. Fensel. 1998. 'Knowledge engineering: Principles and methods' in *Data & Knowledge Engineering*, 25(1-2):161-198, March.

Fensel, D. 2003. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, Berlin.

McGuinness, D. L. 2003 'Ontologies Come of Age' In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, (eds), *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press.

Bizer, C., T. Heath, and T. Berners-Lee. 2009. 'Linked Data -The Story So Far' in: Heath, T., Hepp, M., and Bizer, C. (eds.). *Special Issue on Linked Data, International Journal on Semantic Web and Info Systems (IJSWIS)*

Useful links

Video http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

Linked Data Principles <http://www.w3.org/DesignIssues/LinkedData.html>

<http://www.w3.org/2008/WebVideo/Annotations>

<http://www.w3.org/TR/2010/WD-mediaont-10-20100608/>

http://tech.ebu.ch/docs/tech/tech3293v1_2.pdf

<http://tech.ebu.ch/tvanytime>

http://www.iptc.org/site/News_Exchange_Formats/NewsML-G2/

W3C WebIDL <http://www.w3.org/TR/WebIDL/>