Enriching a News Portal with Semantic Information: An Entity-Based Approach *

Stefano Bocconi
Elsevier Labs
Radarweg 29, 1043 NX Amsterdam, The Netherlands
e-mail: Stefano.Bocconi@gmail.com

Angela Fogarolli
University of Trento
Via Sommarive 14, 38050 Trento, Italy
e-mail:angela.fogarolli@unitn.it

Abstract

In this paper we describe the production and consumption of linked data in the scenario of the Italian news agency ANSA portal. The goal of the use-case is to provide viewers of a news item with background information and links to related news articles contained on the portal. This information enrichment process is entity-based: ANSA news archive is analyzed using Name Entity Recognition, and each detected entity is annotated with a unique identifier. These identifiers are obtained using the Entity Name Server developed within the scope of the OKKAM European project. Subsequently the news are published on the portal using RDFa and linked to Sig.ma, a semantic search engine that provide background information harvested from sources such as DBpedia and links to additional news sources. The presented project has the potential to contribute to Linked Data by creating and publishing a large quantity of entities and assertions about them coming from the ANSA news archive.

1. Introduction

This paper describes the ongoing implementation of an experimental portal for the Italian news agency ANSA[1]. ANSA's business goal in this project is to provide viewers of a news item with background information and links to related news articles contained on the portal, thereby increasing the traffic to the site.

In order to provide both additional background information and links to related news sources, we adopt an entity-based approach. This means that entities contained in news items such as people, organizations or locations act as entry points for a user to gain additional knowledge about them. Background information comes from publicly accessible sources such as DBPedia[2](Auer et al. 2007), as well as from private sources (e.g. information contained in ANSA's subscriber-only knowledge-base). A necessary step to provide all the available information about the same entity from different source is to consolidate that information, which means detecting that different fragments of information actually refer to the same entity.

The OKKAM project[3] (described further in this paper) aims at facilitating this effort by assigning a unique identifier to each entity and providing a service (the Name Entity Server) that can be queried to retrieve such identifiers. Each identifier is associated with a brief entity description (the entity profile), which contains also references to other source of knowledge where the entity has been described. The advantage of using a unique identifier when encoding information about an entity is that several fragments can be unambiguously joined when they refer to the same entity. In this way, Information Retrieval, the necessary step before information consolidation, has ideally 100% precision and recall if an identifier is used instead of relaying on an entity's name (homonyms and synonyms) or other (ambiguous) keywords.

In the news portal in exam, information exists in three formats: initially ANSA's news articles appear in a textual format; subsequently this information is encapsulated in a more structured way through NewsML (International Press Telecommunications Council 2009) and finally some parts (i.e. the information about the entity extracted from the news) are encoded in RDF (using RDFa (W3C 2008)) when the article is published. Moreover, the linked information contained in public and private knowledge-bases which can be navigated starting from the news portal is encoded in RDF.

The task of harvesting public and private online RDF knowledge-bases and consolidate the information is fulfilled by Sig.ma [4] (Oren et al. 2008), a semantic search engine. Sig.ma allows to semantically aggregate data by crawling sites which use RDF, RDFa or Microformats, and consolidate and index this information.

From Sig.ma's point of view, ANSA's experimental portal is both a producer and a consumer of information: entities contained in the news item are linked to background information contained in Sig.ma, but at the same time Sig.ma harvests RDFa content embedded in online ANSA news articles, thereby increasing the amount of information it knows.

The rest of the paper is organized as follow: in Section 2. we present the business rationale of the project, and in

*This work is supported by the FP7 EU Large-scale Integrating Project OKKAM - Enabling a Web of Entities (contract no. ICT-215032).

[1] http://www.ansa.it/
[2] http://dbpedia.org/
[3] http://okkam.org/
[4] http://sig.ma/

Section 3. we illustrate the OKKAM project of which this project is an use case. In Section 4. we describe the actual process of semantic publishing and linking to background information, and in Section 5. some of the current development directions. Finally, we discuss related work in Section 6. and provide some conclusions in Section 7.

2. The business case for a news portal semantic enrichment

The semantic enrichment process described in this paper corresponds to current trends in news production. In fact, according to the International Press Telecommunication Council (International Press Telecommunications Council 2009), pag. 85:

> "*Increasingly, news organisations are using entity extraction engines to find "things" mentioned in news objects. The results of these automated processes may be checked and refined by journalists. The goal is to classify news as richly as possible and to identify people, organisations, places and other entities before sending it to customers, in order to increase its value and usefulness.* "

As expressed in the citation, before being delivered to customers, news needs to go through a phase of entity extraction. In our case, entities are extracted and linked to knowledge bases containing information about them. These sources can be publicly available or restricted according to the kind of subscription a customer has. In Fig. 1 the IPCT information flow is described: when a news is created (News Object), it is examined in order to extract entities. This operation uses a knowledge-base of entities that can disambiguate what entity has been detected and provide its identifier. In our project, the Entity Name System (ENS[5]) is used for entity disambiguation and entities identifiers (More details about the ENS are given in the next section). In case an entity has been found, the news can be offered to customers enriched with references to the entities detected (e.g. the identifiers). These references can be resolved to get more information by accessing the news agency's knowledge-base. In the format devised by the IPTC, information about related entities is grouped in Knowledge Items ((International Press Telecommunications Council 2009), pag. 85). A customer may have (partial) access to Knowledge Items depending on the business model, since the knowledge-base constitutes an asset for a news agency which needs to be maintained, e.g. when new entities are found in the news and need to be added. This last task is taken care of by a documentation team, which can be the journalists themselves or other dedicated staff.

3. An entity-based approach: the OKKAM project

Unique identifiers for our entity-based approach are provided by the tools developed within the OKKAM project.

[5]The Entity Name System is the core of the OKKAM project

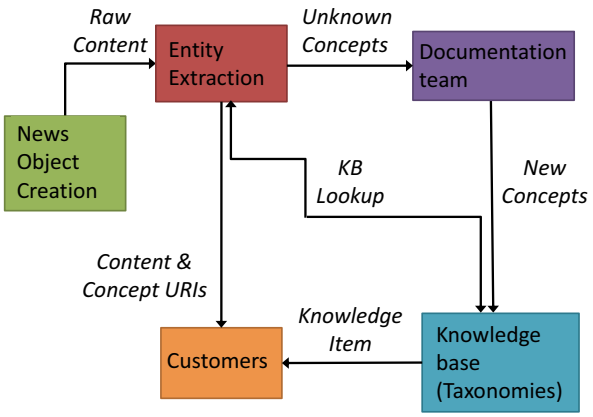

Figure 1: The information extraction process for news items, according to the IPTC

OKKAM is an European FP7 project running from beginning 2008 till mid 2010, with the vision to "*enable the Web of Entities, a global digital space for publishing and managing information about entities, where every entity is uniquely identified, and links between entities can be explicitly specified and exploited in a variety of scenarios*"(Bouquet et al. 2008).

The OKKAM project provides two main assets:

- the Entity Name Server (ENS)
- OKKAM-aware tools that exploit the services of the ENS

The ENS is a distributed infrastructure that allows to store and retrieve entities, i.e. their profile and their OKKAM identifiers. For this use case we consider entities of type person, organization, location, artifact or event. The ENS can not be considered as a knowledge-base, in the sense that the information stored in an entity's profile serves the purpose to make retrieval of the entity's identifier possible. More extensive information about an entity must be stored in external knowledge-bases and can be linked from the OKKAM entity profile.

The ENS can be queried using an entity's attributes to retrieve the entity's OKKAM identifier, or using the identifier to retrieve the entity's profile. Moreover, both queries can return all the alternative identifiers an entity has. This last functionality is especially important in fields where identifiers already exist and are used by a particular community. The ENS establishes a link between different information sources by mapping different identifiers onto each other so that it is possible to integrate different descriptions for the same entity. The ENS itself does not join different descriptions (provided by the alternative identifiers) but simply refers to those descriptions in different data sources. What the ENS does is stating which identifiers refer to the same entity.

In case of OKKAMization, which is the process of assigning an OKKAM identifier to an entity, first the entity is searched by querying the ENS. A query has a particular

syntax which describes through a list of attributes and values the characteristics of an entity (entity profile). If the entity exists, then the entity profile containing its unique OKKAM identifier is retrieved; otherwise, a new entity can be created with the information provided in the query.

The OKKAMization process is at the base of data integration. Data which refer to the same OKKAM identifier from different sources can be aggregated. A user or a software agent can then query those information sources using the relative identifier and then aggregate the retrieved information.

The OKKAM-aware tools include the Name Entity Recognition services that interact with the ENS and the OKKAM semantic data aggregator Sig.ma. Such tools help on one hand to create information sources where entities are associated with unique identifiers, and on the other hand to present users with information relative to an entity. Both cases will be discussed in the scope of their application to the construction of the experimental news portal.

4. The semantic enrichment process

In this section we describe the process that leads to publishing semantically enriched news on the portal. Firstly, we focus on the recognition and annotation of entities in the news. Secondly, we describe how the information about entities extracted in the previous task is connected and visualized. Finally, we relate our semantic enrichment process to the IPTC information flow described in Section 2. and discuss ways to control the quality aspects.

Producing enriched news

As a first step, ANSA news archive is analyzed using Name Entity Recognition (NER). NER is composed by an Entity Detection phase which parses the text and determine the entities, and an Entity Normalization phase which associates a detected entity with its identifier. In this project, NER tools are provided to the OKKAM infrastructure by a commercial company, Expert System[6], which uses a Semantic Network to perform part-of-speech analysis. The NER phase aims at the detection of people, organizations, locations and events inside the news articles. While Entity Detection is a standalone functionality completely developed by Expert System, for Entity Normalization (to provide entities with their ids) the Entity Name System (ENS) is queried to retrieve the OKKAM unique identifier for each detected entity.

News item are encoded in NewsML (International Press Telecommunications Council 2009), a news standard based on XML, and the output format of this phase is 'enriched' NewsML where the section containing the body of the news item has been annotated with OKKAM identifiers. Since this section is meant to be published online, it is encoded in XHTML (within the NewsML format). Annotations can therefore be included using the RDFa format (W3C 2008).

In the following examples, the identifiers id1 and id2 are placeholders for OKKAM identifiers, which look like 'http://www.okkam.org/ens/id1402baaa-c4f4-4553-a610-8e5b41de976c' for id1 and 'http://www.okkam.org/ens/id45432bf3-3a7e-4504-ad40-76d8da98b98f' for id2. When resolving an OKKAM identifier, the profile of the entity[7] is displayed.

Here follows a simple snippet of code of an annotated piece of news:

*(ANSA) - Rome, September 15 - Chamber of Deputies Speaker <**span about="id1" typeof="v:Person"**><**span property="v:name"**>Gianfranco Fini</**span**></**span**> on Monday sued the editor of Premier <**span about="id2" typeof="v:Person"**><**span property="v:name"**>Silvio Berlusconi</**span**></**span**>'s family daily while the Speaker's supporters tried to heal an apparent rift between the long-time political allies.*

The namespace is defined as xmlns:v="http://rdf.data-vocabulary.org/#", which is a small vocabulary developed by Google based on the vCard standard (F. Dawson 1998). In the context of the ANSA news portal, we use this vocabulary because Google announced to semantically support it. Therefore, if we encode news in a google-compliant way they will be more visible and easily retrieved by the Google search engine. This is very important for news agencies that always strive to be indexed by search engines.

Metadata about entities can be encoded using different metadata standards i.e. FOAF[8] and Dublin Core Metadata Initiative[9]. Ideally, we could also use microformats such as hCard, hCalender, since Sig.ma can index them, but ANSA requires annotations not to be visible to the user.

For the sake of clarification, the RDF triples contained in the example above have been extracted using a RDFa extractor which can be found at http://www.sindice.com/developers/inspector. This is the result of the extraction:

```
id1 rdf:type <v:Person>
id1 v:name "Gianfranco Fini"
id2 rdf:type v:Person
id2 v:name "Silvio Berlusconi"
```

Since we use the recognized vCard tags, this information should suggest to Google that the page contains two people who are called 'Gianfranco Fini' and 'Silvio Berlusconi'. Google for the moment ignores the rdf:about field (but it might support it in the future). This leads to the interpretation that the above RDF subjects are blank nodes, which is not a problem from an Information Retrieval point-of-view, since the goal is to retrieve relevant documents. On the other hand, from an information aggregation point-of-view, this does not allow to consolidate information about the same entities (unless using an inverse-functional property to denote the blank node). Sig.ma on the other side does consider the rdf:about field and keeps the reference to the document containing the statement as a source. Using an (OKKAM)

[6] http://www.expertsystem.net/

[7] The format (RDF, HTML) depends on content negotiation.

[8] http://www.foaf-project.org/

[9] RDF definition at http://purl.org/dc/terms/

Figure 2: Sig.ma results for entity 'Gianfranco Fini'

identifier in the about field, semantic search engines can easily aggregate all the statements about an entity. In our case, different statements about the same identifier can be aggregated when the user searches by identifier on Sig.ma.

Connecting to background info

Simply submitting the annotated pages to a semantic search engine (in our case Sig.ma, the predefined semantic aggregator of the OKKAM project) causes the search engine to extract the statements and record the ANSA news page as the source of them. A screenshot of Sig.ma interface is reported in Fig. 2. On the left part of the screen Sig.ma shows facts about the entities (i.e. RDF triples extracted from online sources such as DBpedia but also all RDFa-annotated pages crawled by Sig.ma), and on the right part of the screen the sources providing this information. When the user would search for 'Gianfranco Fini' or for id2, the statements would be displayed and the source (including the ANSA article) reported.

If a user clicks on any detected entity in the news page, a search for that entity is started on Sig.ma. In this way, the user can access the entity's profile, containing facts (such as birth date for a person), and relations with other entities (such as married to <*other entity*> for a person). As already mentioned, these relations are harvested by Sig.ma by crawling public sources such as DBpedia, as well as extracted by pages annotated with RDFa or Microformats. The source section on the right sides contains other news from the ANSA portal where the entity has been mentioned. The user has therefore the possibility to navigate from an entity's profile to news about that entity or to profiles of related entities.

Quality control and maintenance

When we compare the information flow described in Fig. 1 with the scenario explained above, the functionality provided by the OKKAM ENS supports the knowledge-base (KB) role by providing stable identifiers for the entities, with the advantage that external information using the same identifier (or one of the other alternative identifiers known to the ENS) can be merged together. The information contained in such an internal KB can be made available to subscribed customers.

Both the ENS and the KB need to be maintained and updated when new entities are found in news items. In fact, entity extraction is not always able to detect all the entities and link them to the correct information.

The extraction might:

1. Detect as an entity something which is not an entity (false positive)
2. Not detect as an entity something that is an entity (false negative)
3. Detect an entity but link it to the wrong identifier (e.g. in case of homonyms)
4. Detect an entity (possibly after human intervention) but being unable to link it since the entity is not present in the knowledge-base nor in the ENS

In all these cases there is a need for quality control that corrects the results of the extraction. In case 1,2 and 3 the correction can be performed by the journalist using an authoring tool, directly on the news item being annotated.

The last case requires information to be entered in the repository, and therefore a modification of the KB and the ENS. The envisioned workflow is that the journalist can request a new identifier for a new entity, to annotate that entity in the news item, and this identifier is immediately created, with minimum profile information in the ENS. At a later stage, when time constraints are less stringent, more information can be entered in the profile and in the KB by the journalist or by a documentation team. The cost of such a documentation team is justified by the return of investment of selling the KB services to customers, since the news agency knowledge-base constitutes a valuable asset on its own. On the other hand, the advantage of this approach is that the entity repository is augmented to contain entities that might not be present in other public sources, as we say in the following.

5. Ongoing efforts

This project can significantly contribute to publish semantic data, considering that news items contain a large number of entities which are not "famous"' enough to get coverage by enciclopedyc initiative such as DBPedia. The fact that Sig.ma harvests these news makes these entities publicly searchable for the Semantic Web community.

Here we describe some directions we are currently experimenting.

Creating information

Up till now, in the scenario we presented, RDFa annotations are used merely as a way to notify Sig.ma (and Google) that a page contains information about a certain entity, without providing more additional information than the name and the type of the entity (e.g. v:Person).

In this sense, news articles only 'consume' (through Sig.ma) information about entity coming from other sources such as DBpedia, without publishing additional information that these articles contain. This is not necessarily so, as in fact two different scenarios are possible:

- annotations describe additional information about the entity
- annotations describe events that relate different entities

An example of the first case is the following:

* The Italian prime minister *

This indicates that the entity identified by id2 ('Silvio Berlusconi') is the Italian prime minister (has role).

An example of the second case is the following:

**

Thereby indicating an event that happened between id1 and id2. Currently we are trying to detect events as standalone terms, and not as relations between for example an actor and a subject. We are therefore investigating in what measure roles and relations can be automatically detected in the text by Name Entity Recognition services.

When HTML is used as the container for RDF as in the RDFa annotations, structural elements of the web page can be preserved while defining and carrying arbitrary vocabularies (such as microformats or metadata standards).

In the previous examples we have seen the use of rdf:subject or rdf:predicate inside an HTML span property; but the RDF elements can be included and combined in other HTML elements, e.g. the previous example could also be expressed with the same meaning in the following way:

* The Italian prime ministerSilvio Berlusconi *

or

* Silvio Berlusconi The Italian prime minister *

Here we report some guidelines we developed for creating linked data using OKKAM identifiers for the ANSA News Portal:

1. Declare the prefix mappings example include:

```
xmlns:okkam="http://models.okkam.org/ENS-core-vocabulary.owl"
xmlns:dc="http://purl.org/dc/terms/"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:v="http://rdf.data-vocabulary.org/"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema"${}
```

2. Use the tag rdf:about for describing an entity though the OKKAM unique identifier.
3. Use the Google-supported vocabulary whenever the property is covered and use otherwise other metadata standards.
4. Validate the RDF by checking the resulting triples using SINDICE extractor [10].

Entity consolidation on the Web

To exploit the full power of entity-based information aggregation, information harvested from different sources needs to be consolidated, assigning the right data to the right entity. Sig.ma is an entity centric semantic search engine, which means that the result from a search using attributes that refer to an entity is theoretically a set of distinct entities E_i, $i = 0..n$, with 0 in case there is no such entity, 1 if the attributes are unambiguous, so when there is only one occurrence for the searched entity (for example when searching by OKKAM id), and n if there is ambiguity.

In the example cited above, searching for "Gianfranco Fini" should return a set of data grouped into different entity profiles (one of more in case there are homonyms). One of these profile must be the same profile as returned by a search for id1 (Gianfranco Fini's OKKAM id). Sig.ma is in the process of implementing this consolidation phase. Several methods to decide whether two entity description actually refer to the same entity are currently investigated. One such approach (Stoermer and Bouquet 2009) combines probabilistic as well as ontological methods for deciding whether two records describe the same entity, taking into account intensional and extensional aspects of the entities at hand. The approach aims at general-purpose usefulness with special focus on web information systems, and bases on empirical findings about what are commonly used entity types, and how they are usually described.

Assigning unique identifiers to entities is an important step to be able to consolidate entities. Identifiers namely provide a means for authors to express knowledge and relate it unambiguously to a particular entity. Subsequent integration of different information sources can exploit the author's judgment instead of having to rely on error-prone algorithms. This is also called *a-priori* integration (Bouquet, Stoermer, and Bazzanella 2008) and is supported by the OKKAM ENS. Since the ENS is an open, public service which can be invoked by any application in which entities are mentioned, any authoring tool can implement this *a-priori* integration.

6. Related Work

Several initiatives have targeted news publishing trying to enrich a viewer's experience of news reading. Many of them come from commercial companies, the most known being probably Thompson Reuters' OpenCalais[11]. OpenCalais offer Name Entity Recognition services for free through a Web and API interface. They also offer stable identifiers, but they

[10] Sindice is also a tool partially developed inside the OKKAM project. The triple extractor can be found at: `http://www.sindice.com/developers/inspector`

[11] `http://www.opencalais.com/`

are not providing any way to obtain from attributes of an entity that entity's identifier, and they do not attempt entity consolidation.

Evri[12] is a commercial company that categorizes and extracts entities from text, especially news. They offer an interesting search interface that resemble a SPARQL endpoint above unstructured data. They have no notion of stable identifiers, which makes it hard to connect to their knowledgebase.

Another interesting example of semantic enrichment of news is iGlue, currently under development, but with an online demo[13] that shows entities detected in a news item. When the user clicks on one entity, they get a pop-up window with the entity profile.

On the identifiers field, Hepp et.al (Hepp, Siorpaes, and Bachlechner Sept Oct 2007) first proposed the use of Wikipedia URI as an identification of resources in knowledge management. DBPedia now offers a way to automatically harvest Wikipedia identifiers, but does not cover the vast variety of entities contained in large news archive (especially the not-so-famous characters in news items).

A very similar initiative to the one presented in this paper has been performed at the BBC. With the goal of better connecting BBC online resources and documents to each other, (Kobilarov et al. 2009) use Name Entity Recognition to map entities to unique identifiers (DBpedia terms). The main difference with their approach is that they use DBpedia to disambiguate possible matches and as an identifier system, while in our approach the disambiguation is performed by the semantic network of the extractor and by the ENS. Beside the technical difference, their approach relies on DBpedia covering all terms, which is not always the case when terms are not so popular, as mentioned before.

7. Conclusions

In this paper we describe the ongoing implementation of an experimental portal for the Italian News agency ANSA, within the scope of the EU project OKKAM. We report the business rationale for such a project and the technical infrastructure in place to first annotate with, and then index, semantic information in news item. Further, we discuss how this effort is positioned in the broader scenario of semantic publishing for the Linked Data Community, especially from the point of view of our approach, which is entity-based. We also argue that two steps are very important for the production and consumption of semantic information, namely assigning unique identifiers to entities and consolidate information harvested from the Web about the same entity.

Acknowledgements

The authors would like to thank the Italian news agency ANSA for providing and working on the project presented in this paper, the software company Expert System for providing the Name Entity Recognition software and the OKKAM project partners for building the infrastructure used in this project.

[12] Evri.com

[13] http://iglue.com:8082/demo1/query.nytimes.com/gst/index.html

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. G. 2007. Dbpedia: A nucleus for a web of open data. In Aberer, K.; Choi, K.-S.; Noy, N. F.; Allemang, D.; Lee, K.-I.; Nixon, L. J. B.; Golbeck, J.; Mika, P.; Maynard, D.; Mizoguchi, R.; Schreiber, G.; and Cudré-Mauroux, P., eds., *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, 722–735. Springer.

Bouquet, P.; Stoermer, H.; Niederee, C.; and Mana, A. 2008. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, number CSS-ICSC 2008-4-28-25 in CSS-ICSC, 554–561. IEEE Computer Society.

Bouquet, P.; Stoermer, H.; and Bazzanella, B. 2008. An Entity Name System (ENS) for the Semantic Web. In *The Semantic Web: Research and Applications. Proceedings of ESWC2008.*, volume Volume 5021/2008 of *Lecture Notes in Computer Science*, 258–272. Springer Berlin / Heidelberg.

F. Dawson, T. H. 1998. vcard mime directory profile, ietf rfc 2426. Technical report, The Internet Engineering Task Force (IETF).

Hepp, M.; Siorpaes, K.; and Bachlechner, D. Sept.-Oct. 2007. Harvesting wiki consensus: Using wikipedia entries as vocabulary for knowledge management. *Internet Computing, IEEE* 11(5):54–65.

International Press Telecommunications Council. 2009. Guide for implementers.

Kobilarov, G.; Scott, T.; Raimond, Y.; Oliver, S.; Sizemore, C.; Smethurst, M.; Bizer, C.; and Lee, R. 2009. Media meets semantic web — how the bbc uses dbpedia and linked data to make connections. In *ESWC 2009 Heraklion: Proceedings of the 6th European Semantic Web Conference on The Semantic Web*, 723–737. Berlin, Heidelberg: Springer-Verlag.

Oren, E.; Delbru, R.; Catasta, M.; Cyganiak, R.; Stenzhorn, H.; and Tummarello, G. 2008. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies* 3(1).

Stoermer, H., and Bouquet, P. 2009. A novel approach for entity linkage. In Zhang, K., and Alhajj, R., eds., *Proceedings of IRI 2009, the 10th IEEE Internationational Conference on Information Reuse and Integration*, 151–156. IEEE Systems, Man and Cybernetics Society.

W3C. 2008. Rdfa in xhtml: Syntax and processing, w3c recommendation.