

# Enabling Case-Based Reasoning on the Web of Data

Benjamin Heitmann and Conor Hayes

firstname.lastname@deri.org  
Digital Enterprise Research Institute  
NUI Galway  
Ireland

**Abstract.** While Case-based reasoning (CBR) has successfully been deployed on the Web, its data models are typically inconsistent with existing information infrastructure and standards. In this paper, we examine how CBR can operate on the emerging Web of Data, with mutual benefits. The expense of knowledge engineering and curating a case base can be reduced by using Linked Data from the Web of Data. While Linked Data provides experiential data from many different domains, it also contains inconsistencies, missing data and noise which provide challenges for logic-based reasoning. CBR is well suited to provide alternative and robust reasoning approaches. We introduce (i) a lightweight CBR vocabulary which is suited for the open ecosystem of the emerging Web of Data, and provide (ii) a detailed example of a case base using data from multiple sources. We propose that for the first time the Web of Data provides data and a real context for open CBR systems.

## 1 Introduction

Case-Based Reasoning (CBR) has made great inroads on the Web as a means of automated technical support and product recommendation. However, it still is open to the criticism made by Kitano and Shimazu that its focus on domain specific problems has led to closed and inflexible data models, typically isolated from other information infrastructures and standards [1]. With the subsequent growth of the Web, there have been several attempts by the CBR community to standardise the different parts of CBR's knowledge containers to enable interoperability and wider dissemination of case-based technology [2–6]. Several iterations of the XML-based Case-Based Markup Language (CBML) [2, 5, 7, 6] were developed between 1998 and 2004, though practical adoption of the language was low. Concurrent work on distributed and multi-casebase reasoning, while assuming case-base interoperability, did not specify how that might be achieved (see [8] for an overview). Lack of adoption may have been partly due to the cumbersome manner in which the structure and the semantics of the case data were described in XML. We suggest that a greater impediment was the lack of real data and real contexts in which distributed case-based reasoning would have been required.

In approximately the same time period, rule and logic-based reasoning has gone through a renaissance in close association with the **Semantic Web** research initiative. However, the particular strengths of case-based reasoning - its suitability for weak theory domains, its relative robustness to noise and its ability to

incorporate techniques from machine and statistical learning make it an ideal alternative reasoning paradigm for the Semantic Web, an observation that has been made by Tim Berners-Lee [9]. Furthermore, the constraints for wider adoption of case-based reasoning standards have significantly been loosened through several features of the rapidly emerging **Web of Data**, a highly practical offshoot of Semantic Web research. Firstly, the awkward requirement to define case semantics has been delegated to the hundreds of integrated Linked Data vocabularies already deployed. Secondly, there is an enormous amount of data in structured, semantically annotated format already on the Web. What is missing is a way of creating flexible case views of this data so that reasoning with cases can emerge as standard reasoning paradigm on the Web of Data.

In this paper, we describe how we can create case views of **Linked Data**, which would enable the authoring of and querying of distributed case data on the Web. Linked Data refers to a set of principles for publishing and connecting structured data on the Web, thus forming a Web of Data [10]. While existing CBR systems have been developed under the assumption of a closed ecosystem, the Web of Data represents an open ecosystem, where each data source can evolve and change independently [11]. During the course of this paper, we will explain how Linked Data can be used to provide structured data for the different knowledge containers of a CBR system.

The main contributions of this paper are: (i) the introduction of a lightweight CBR vocabulary which is suited for the open ecosystem of the emerging Web of Data, and (ii) a detailed example of a case base using data from multiple sources.

The rest of the paper is structured as follows: In section 2 we discuss related work and explain our approach for providing a CBR view on Linked Data. Section 3 explains the necessary background about the Linked Data principles and the standards used for the Web of Data, as well as Linked Data sources for experience mining. Section 4 explains the benefits of using Linked Data as sources for the different knowledge containers of a CBR system. Then in section 5 we introduce our lightweight CBR vocabulary, and describe an example case base from the music domain with data from multiple sources in section 6. Finally, section 8 discusses our contributions and concludes the paper.

## 2 Related work

Despite the conventional focus on reasoning from a single case base, the idea that case knowledge can be distributed in several places has been recurrent in CBR research from its earliest days. Kolodner [12], Barletta and Mark [13] and Redmond [14] all proposed approaches in which case snippets were linked to form a full case episode. Redmond [14] argued that a case-based reasoner cannot know in advance which parts of different cases may be useful during a problem-solving episode. Therefore, storing cases as a linked network of snippets allowed for more flexible retrieval possibilities than their storage as monolithic cases.

While the approach we propose focuses on developing case views of the emerging Web of Data, we draw inspiration from the ideas in this early work. Kitano and Shimazu's proposal [1] for a less domain-fixated role for CBR led to several research initiatives focusing on how the then new XML standards could be used to develop standard case representation formats [2–4]. Several instantiations of Case-Based Mark-up Language (**CBML**) were subsequently developed with the

aim of facilitating case base interoperability and wider dissemination and integration of case-based systems on the Web [5, 7, 6]. With the emergence of XML schemas, Hayes and Cunningham proposed a version of CBML that wrapped existing XML domain data to produce domain-compliant case views [5]. The work in this paper shares this perspective. Subsequently, Coyle et al. produced versions of CBML capable of representing hierarchical and object-oriented case data [7] as well as similarity knowledge [6]. Unfortunately, despite concurrent research on distributed/multi-case based reasoning, none of the CBML research gained much traction. This may have been partly due to the cumbersome manner in which XML encoded the structure and the semantics of the case data. Another impediment was the lack of real data and real contexts in which distributed case-based reasoning would be required.

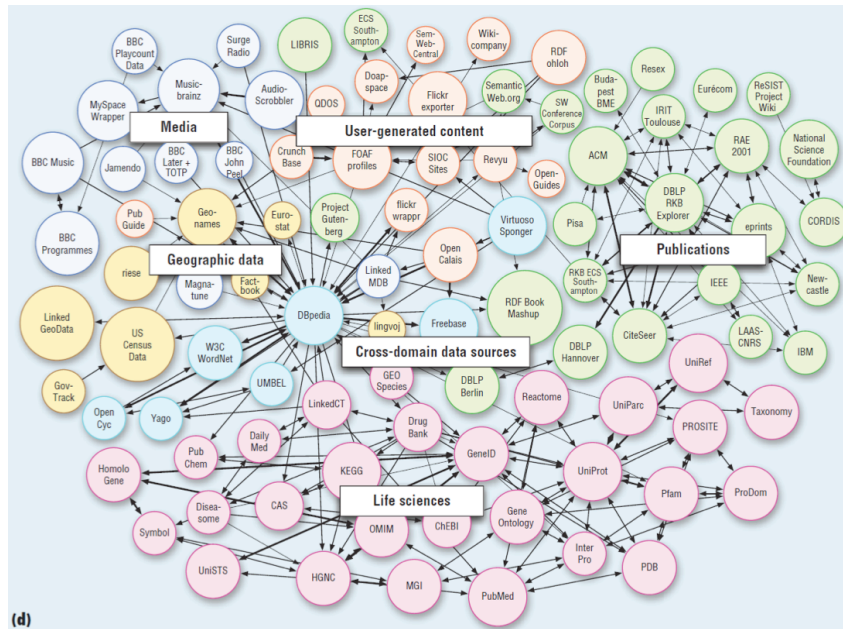
In this paper we suggest that the emergence of the Web of Data, not only provides data, but a context and real requirement for reasoning with cases. As we discuss in the next section, the Web of Data is described using the Resource Description Format (RDF). There has been some work on an RDF-based Case Markup Language, **CaseML** [15]. However, it suffers from the same problems as the earlier versions of CBML in that it shoehorns existing data into a rigid case format, thus hampering interoperability with existing vocabularies and schemas. Furthermore, it subscribes to a view of a closed RDF-based system, which can integrate data from specific controlled sources, but which does not expect to operate on open data from the Linked Data cloud. In contrast, we propose a flexible, lightweight case vocabulary for creating case views on the Web of Data, an open data environment where new vocabularies are regularly added.

The Web Ontology Language (**OWL**) has been proposed for representing cases in an interoperable way. [16] discusses the general applicability of OWL for case representation in biology and medicine. [17] propose using C-OWL with contexts to model the perspectives of different viewpoints. These approaches do not address the cost of engineering the knowledge required for the case base and the OWL reasoning. Our approach is orthogonal, in that it focuses on authoring and curating the knowledge required to bootstrap a CBR system from the Web of Data.

### 3 Background: Linked Data and the Web of Data

The term **Linked Data** refers to a set of best practices for publishing and connecting structured data on the web [10]. Taken together, all Linked Data constitutes the Web of Data. While the World Wide Web provides the means for creating a web of human readable documents, the Web of Data aims to create a web of structured, machine-readable data. Much of this data can be used for the purpose of mining experiential data from the Web. By making data available which connects the experiences of users from different domains and communities, Linked Data has the potential to provide data which never before has been available for the purpose of experience mining.

The **Web of Data** utilises technologies from the Semantic Web technology stack: the Resource Description Framework (**RDF**) provides a graph based data model and the basic, domain independent formal semantics for the data model [18]; the **SPARQL** Query Language allows querying RDF data with graph patterns, and provides basic means for transforming RDF data between different schemata. In addition, technologies from the World Wide Web provide the fundamental



**Fig. 1.** Overview of Linking Open Data sources as of July 2009 with source types shown

infrastructure: Uniform Resource Identifiers (**URIs**) are used to provide globally unique identifiers for the data, and the HyperText Transfer Protocol (**HTTP**) is used for accessing and transporting the data.

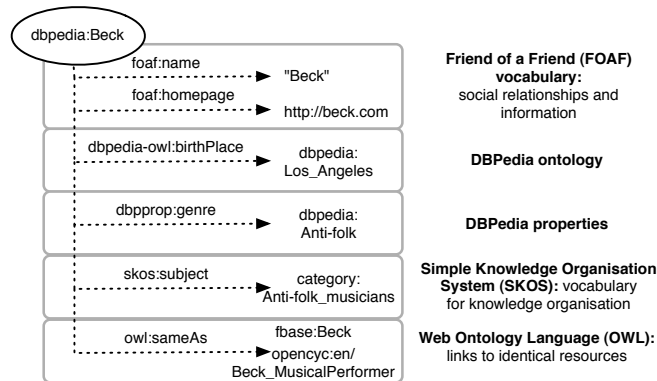
In order to build a single Web of Data, all data providers have to follow the same guidelines for publishing their data and connecting it to other data sources. These guidelines are provided by the “**Linked Data principles**” [10], which specify how to use the different standards of the Web of Data together:

1. Use URIs as names for things (and e.g. persons, places).
2. Use HTTP URIs so that people can look up and access those names via HTTP.
3. When someone looks up a URI, provide useful information, using the standards RDF and SPARQL.
4. Include links to other URIs, so that data about more things can be discovered.

The Linked Data principles have been adopted by an increasing number of data providers, especially from the Linking Open Data community project<sup>1</sup> (see Figure 1), which makes free and public data available as Linked Data. As of November 2009 this includes 13.1 billion RDF triples which are interlinked by 142 million RDF links.

The nucleus of the Linked Data cloud is formed by **DBpedia**, which extracts RDF from wikipedia topic pages, and thus provides authoritative URIs and RDF data about topics from any domain. Figure 2 shows how DBpedia makes use of different vocabularies to describe a resource.

<sup>1</sup> <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/>



**Fig. 2.** Example of a DBPedia resource with descriptions of the different vocabularies used for the properties

The following types of Linked Data sources provide data which could be used for mining experiences:

Major **search engines** such as Google and Yahoo have started to index RDF meta-data which is embedded in web pages. This can be directly used to expose structured data about user experiences and reviews to the search engines. Yelp and Amazon for instance are providing this data, so that e.g. the average rating for an item can be already displayed as part of the search results. **Social Web sites** such as LiveJournal make user generated content from their users available. This data is modelled after the principle of object centred sociality [19], connecting users indirectly into communities via objects of a social focus, such as musical artists.

**E-commerce** sites use the Good Relations vocabulary to describe their products and their features and prices, payment options, as well as store locations and opening hours. BestBuy has released such data about all of their stores in the US. **Broadcasters and news publishers** provide data about media content and usage. The DBTune project makes data about MySpace users and their connections to musical artists available. Other sources include the BBC's catalogue of broadcasts on TV and Radio, as well as play-count data for different artists played across all of the BBC stations.

## 4 A case-based view on the Web of Data

While CBR suffers less from the knowledge elicitation problem, it is still not immune from it. In this section we examine how Linked Data can help bootstrap the development of decentralised case-based systems. Customisation of each CBR system for a specific domain is possible via the selection of sources and instances from these sources and through the assignment of data to the four 'containers' which hold the knowledge used in a CBR system (as defined by Richter [20]):

1. **Vocabulary knowledge** consists of the semantic components that can be manipulated by a reasoning system.

2. **Case knowledge** consists of the ‘problem’ episodes or instances represented as cases that can be used to solve similar problems in the future.
3. **Similarity knowledge** represents the similarity measures which are used to match cases in a particular domain.
4. **Adaptation knowledge** is knowledge used to adapt the solution of the matching case for the target problem.

Capturing the knowledge for these containers represents a challenge in every CBR system because of the knowledge engineering costs required in their creation and curation. Vocabulary knowledge may need to be extracted and ordered, case structures may need to be designed and features selected, similarity measures may need to be designed and tested and adaptation strategies modelled. We propose that Linked Data can be used as a knowledge source to ease the creation of these containers in many CBR systems.

Pre-existing vocabulary knowledge is contained in the vocabularies and ontologies already deployed on the Web of Data. Some vocabularies just include class definitions and relations between classes, such as the Semantically-Interlinked Online Communities (SIOC) vocabulary<sup>2</sup>. Then there are topical hierarchies, such as the taxonomy of DBpedia categories. Finally there general knowledge ontologies expressed in OWL such as OpenCyc<sup>3</sup>.

Case knowledge is provided by the different data sources which are part of the Linking Open Data cloud, and which have been discussed in the background section. We will show detailed examples of using LD for case knowledge in section 6.

Adaptation knowledge still needs to be engineered for the specific domain of the CBR system. [17] gives an example of how OWL can be used to model adaptation logic. While similarity knowledge is currently not available as linked data, Coyle et al. [6] show how this knowledge can be modelled in XML, and this approach can be adapted for RDF and Linked Data.

## 5 CBR Vocabulary

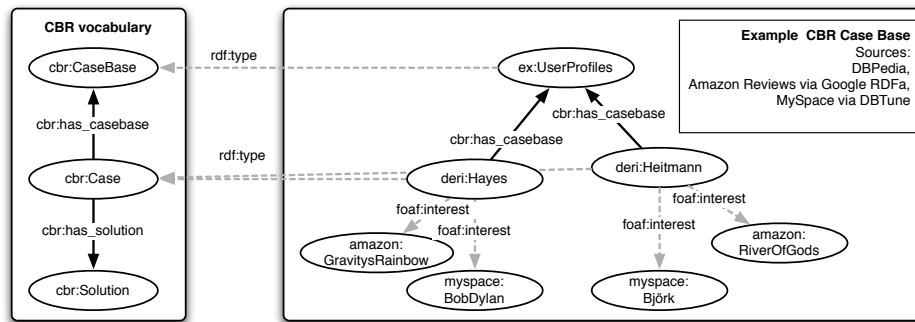
In order to use Linked Data for case based reasoning, we are proposing a new, light-weight CBR vocabulary to provide a CBR view on Linked Data. Figure 3 shows the full vocabulary on the left hand side. While CBML [21] and CaseML [15] have attempted to define all of the objects and properties which are broadly associated with CBR, we will just focus on the few main classes and properties which are unique to CBR: cases, solutions and case bases.

This very simple vocabulary allows us to create cases by linking to different objects, properties and concepts from different data sources and vocabularies from the Web of Data. This gives us the ability to reuse case snippets from a massive and increasing data store which includes many different communities and domains. This allows us to build CBR systems on top of all this newly available data.

- We define the classes `cbr:Case` and `cbr:Solution` in order to allow a differentiation between a query and potential results or recommendations. Note that a resource can be an instance of both classes at the same time.

<sup>2</sup> <http://rdfs.org/sioc/spec/>

<sup>3</sup> <http://sw.opencyc.org/>



**Fig. 3.** The CBR vocabulary and a case base with experiential data for user profiling. Properties from the CBR vocabulary are displayed with solid lines, all other properties are displayed in dashed lines.

- Cases are connected to solutions via the property `cbr:has_solution` which can have cases as subject and solutions as object (`cbr:has_solution` has `rdfs:domain cbr:Case` and `rdfs:range cbr:Solution`). Note that usage of the `cbr:has_solution` property is optional, as not every CBR setting requires a solution to be designated in advance.
- Cases are grouped together via the `cbr:CaseBase` class, which can be used to represent the different types of knowledge available to a CBR system.
- Cases are associated to case bases via the property `cbr:has_casebase`, which has `rdfs:domain cbr:Case` and `rdfs:range cbr:CaseBase`.

**Modelling decisions:** As the Web of Data is an open ecosystem, applications should not make assumptions about the vocabulary used by data sources and they should take frequent changes of sources into account [11]. This has influenced several modelling decisions for our CBR vocabulary: (a) **intentional simplicity** characterises our vocabulary and lowers the barriers for reusing and integrating external data. This leads us to model problems as cases, in order to simplify reuse of external domain data. (b) **reuse of existing properties** and data types, as we leave the definition of the properties which represent the actual features of the data to externally defined vocabularies and ontologies, as appropriate to the domain of a source. (c) **simplified inheritance** and inferencing, as the formal RDF semantics follow the open world assumption, which is only superficially related to object oriented programming. Resources are instances of multiple classes as defined by the properties which they use. Consider as an example, that by using the `cbr:has_solution` property, its subject automatically becomes a `cbr:Case` and its object becomes a `cbr:Solution`.

## 6 Example: a case base with experiential data for user profiling

In this example we show how to create an example case base by using our CBR view on the Web of Data. This involves three steps: (1) discovering and aggregating relevant data, (2) conversion of external data to a common format, and (3)

authoring of the case base. The scenario for the example case base is to aggregate data for user profiles and to use the profiles for personalised recommendations.

**First step: discovering and aggregating relevant data.** In order to discover relevant data, different approaches can be used [22]: SPARQL queries to specific data sources, Linked Data search engines such as *Sindice.com* or a custom built crawler. The experiential data for our example comes from Amazon.com reviews and MySpace user profiles. We access the Amazon data via the *Sindice.com* search service, the MySpace data is accessed via the SPARQL endpoint provided by the *DBTune.org* project.

**Second step: conversion of external data.** After discovering the data, it needs to be converted to RDF as a common format. Amazon exposes its data about reviews as RDFa in its HTML pages. This data can automatically be converted to RDF, see [22] for details. Listing 1.1 shows the data with one review of the book “River of Gods”, obtained because the user Benjamin Heitmann wrote the review.

Listing 1.2 shows data from MySpace about the musician “Bob Dylan”, obtained because user Conor Hayes lists him as a favourite artist. *DBTune.org* provides a SPARQL endpoint which directly returns RDF data.

#### Listing 1.1. A book review from Amazon

```
amazon:RiverOfGodsReview1 review:itemreviewed "Book: River of Gods, Author: Ian McDonald" .
amazon:RiverOfGodsReview1 review:summary "Als construct a black hole to escape the AI police" .
amazon:RiverOfGodsReview1 review:rating "5" .
```

#### Listing 1.2. Data about a musical artist from MySpace

```
myspace:BobDylan rdf:type http://purl.org/ontology/mo/MusicArtist .
myspace:BobDylan foaf:name "Bob Dylan" .
myspace:BobDylan myspace:genreTag http://purl.org/ontology/myspace#Folk%20Rock .
```

**Third step: authoring of the case base.** After aggregating and converting external data, the entities which should be part of the case base can be selected. This can be either done manually or automatically through the domain logic of the application.

Figure 3 shows the final graph of the example case base. Only the solid lines represent properties from the CBR vocabulary. These properties are created through associating the entities representing the users *deri:Hayes* and *deri:Heitmann* to the case base during the authoring phase. Listing 1.3 contains all the RDF that is required to define the example case base.

#### Listing 1.3. RDF triples for the example case base

```
deri:Hayes cbr:has_casebase ex:UserProfiles .
deri:Hayes foaf:interest myspace:BobDylan .
deri:Heitmann cbr:has_casebase ex:UserProfiles .
deri:Heitmann foaf:interest amazon:RiverOfGodsReview1 .
```

## 7 Discussion

In previous work [23], we have shown how Linked Data can be used to ease the bottleneck in knowledge acquisition for recommender systems based on collaborative filtering. In a certain sense, we are now proposing to use Linked Data for Case Based Reasoning out of the same reasons: We are proposing new ways to lower

the entry barriers to implementing and deploying CBR systems, by using Linked Data as a short-cut for creating the case base, and for lowering the expense of the required knowledge engineering.

In our example, we show how a user profile can be developed by selecting objects and associated properties and relations from structured data sources currently existing on the Web. On the DBpedia data source alone, rich data exists for at least 3.4 million things, out of which 1.5 million are classified in a consistent Ontology, including 312,000 persons, 413,000 places, 94,000 music albums, 49,000 films, 15,000 video games, 140,000 organisations, 146,000 species and 4,600 diseases. The DBpedia data has meta data and abstracts for these 3.4 million entities in up to 92 different languages; It has 1,460,000 links to images and 5,543,000 links to external web pages; 4,887,000 external links into other RDF datasets, 565,000 Wikipedia categories, and 75,000 YAGO categories.<sup>4</sup>

Such data suggest the possibility of collecting rich experience trails of users and then using these to find similar profiles or to recommend objects of different types. While this is probably the easiest example we can envisage, there are rich possibilities for reasoning on case views of this data and for integrating case-based reasoning as a standard reasoning paradigm for Semantic Web data.

**Limitations:** In this paper we focus on case and vocabulary knowledge, as it is readily available for many different domains on the Web of Data today. However, we do not address approaches for engineering of adaptation and similarity knowledge, though we believe that here too knowledge costs can be reduced by examining the pre-existing domain concepts and relations on the Web of Data. We also do not discuss the different reasoning approaches which can be used as part of the CBR process. Instead we focus on acquiring the data which is required as the basis for the reasoning process in the first place. Finally, while the Web of Data provides many different sources of experiential data, this also can incur an implementation overhead as obtained data might be noisy or inconsistent.

## 8 Conclusion

In this paper we have suggested that there is an open opportunity for reasoning with cases of experiential data from the emerging Web of Data. Previous attempts to standardise case representation have failed because of a lack of real data and context for distributed case-based reasoning. The Web of Data now presents such a context. We provide an overview of the Linked Data principles and describe how Linked Data can facilitate the authoring of experiential case data. We present a realistic example of how case-based user profiles can be authored from data currently available from multiple sources. Such cases have the advantage of having rich features, allowing for many different types of matching and retrieval methods. While our example is simple, it demonstrates the rich possibilities for reasoning on case views of this data and for integrating case-based reasoning more extensively within Semantic Web applications.

**Acknowledgements:** The work presented in this paper has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

---

<sup>4</sup> These statistics are taken from <http://en.wikipedia.org/wiki/DBpedia>

## References

1. Kitano, H., Shimazu, H.: The Experience Sharing Architecture: A Case Study in Corporate-Wide Case-Based Software Quality Control. In: *Case-Based Reasoning: Experiences, Lessons, & Future Directions*. Leake, DB (Ed.) pp. 235-268 (1996)
2. Hayes, C., Cunningham, P., Doyle, M.: Distributed CBR using XML. In: *Proceedings of the KI-98 Workshop on Intelligent Systems and Electronic Commerce*, Citeseer (1998)
3. Shimazu, H.: A textual Case-Based Reasoning system using XML on the World-Wide Web. *Advances in case-based reasoning* (1998) 274–285
4. Watson, I., Gardingen, D.: A distributed case-based reasoning application for engineering sales support. In: *International Joint Conference on Artificial Intelligence*. Volume 16., Citeseer (1999) 600–605
5. Hayes, C., Cunningham, P.: Shaping a CBR view with XML. (*Case-Based Reasoning Research and Development*) 722–722
6. Coyle, L., Doyle, D., Cunningham, P.: Representing Similarity for CBR in XML. (*Advances in Case-Based Reasoning*) 155–164
7. Coyle, L., Hayes, C., Cunningham, P.: Representing cases for cbr in xml. *Expert Update* **6**(2) (2003) 7–13
8. Plaza, E., McGinty, L.: Distributed case-based reasoning. *The Knowledge engineering review* **20**(03) (2006) 261–265
9. Berners-Lee, T., Hall, W., Hendler, J., O’Hara, K., Shadbolt, N., Weitzner, D.: A framework for Web Science. *Foundations and Trends in Web Science* **1**(1) (2006) 76–77
10. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data—the story so far. *Journal on Semantic Web and Information Systems* (2009)
11. Oren, E., Heitmann, B., Decker, S.: ActiveRDF: embedding Semantic Web data into object-oriented languages. *Journal of Web Semantics* (2008)
12. Kolodner, J.: Retrieving events from a case memory: A parallel implementation. In: *Proceedings of a Workshop on Case-Based Reasoning*. Volume 233. (1988) 249
13. Barletta, R., Mark, W.: Breaking cases into pieces. In: *Proceedings of Case-Based Reasoning Workshop*. (1988) 12–17
14. Redmond, M.: Distributed cases for case-based reasoning: Facilitating use of multiple cases. In: *Proceedings of AAAI*. Volume 90. (1990) 304–309
15. Chen, H., Wu, Z.: On Case-based Knowledge Sharing in Semantic Web. In: *International Conference on Tools with Artificial Intelligence*. (2003) 200–207
16. Bichindaritz, I.: Mémoire: Case Based Reasoning Meets the Semantic Web in Biology and Medicine. In: *European Case Based Reasoning Conference*. (2004)
17. d’Aquin, M., Lieber, J., Napoli, A.: Decentralized Case-Based reasoning for the Semantic Web. *International Semantic Web Conference* (2005) 142–155
18. Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., Horrocks, I.: The semantic web: The roles of XML and RDF. *IEEE Internet computing* **4**(5) (2000) 63–73
19. Bojars, U., Passant, A., Cyganiak, R., Breslin, J.: Weaving SIOC into the Web of Linked Data. In: *Linked Data on the Web Workshop*. (2008)
20. Richter, M.: The knowledge contained in similarity measures. *Invited Talk at ICCBR* **95** (1995)
21. Hayes, C., Cunningham, P.: Shaping a CBR View with XML. In: *International Conference on Case-Based Reasoning*. (1999) 468–481
22. Heitmann, B., Kinsella, S., Hayes, C., Decker, S.: Implementing semantic web applications: reference architecture and challenges. In: *Workshop on Semantic Web Enabled Software Engineering*. (2009)
23. Heitmann, B., Hayes, C.: Using linked data to build open, collaborative recommender systems. In: *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*. (2010)