



---

# **Contextual Information Retrieval in Research Articles: Semantic Publishing Tools for the Research Community**

M.A. Angrosh  
Stephen Cranefield  
Nigel Stanger

---

**The Information Science  
Discussion Paper Series**

Number 2011/06  
July 2011  
ISSN 1177-455X

## University of Otago

### Department of Information Science

The Department of Information Science is one of seven departments that make up the School of Business at the University of Otago. The department offers courses of study leading to a major in Information Science within the BCom, BA and BSc degrees. In addition to undergraduate teaching, the department is also strongly involved in post-graduate research programmes leading to MCom, MA, MSc and PhD degrees. Research projects in spatial information processing, connectionist-based information systems, software engineering and software development, information engineering and database, software metrics, distributed information systems, multimedia information systems and information systems security are particularly well supported.

The views expressed in this paper are not necessarily those of the department as a whole. The accuracy of the information presented in this paper is the sole responsibility of the authors.

### Copyright

Copyright remains with the authors. Permission to copy for research or teaching purposes is granted on the condition that the authors and the Series are given due acknowledgment. Reproduction in any form for purposes other than research or teaching is forbidden unless prior written permission has been obtained from the authors.

### Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper, or for subsequent publication details. Please write directly to the authors at the address provided below. (Details of final journal/conference publication venues for these papers are also provided on the Department's publications web pages: <http://www.otago.ac.nz/informationscience/pubs/>). Any other correspondence concerning the Series should be sent to the DPS Coordinator.

Department of Information Science  
University of Otago  
P O Box 56  
Dunedin  
NEW ZEALAND

Fax: +64 3 479 8311  
email: [dps@infoscience.otago.ac.nz](mailto:dps@infoscience.otago.ac.nz)  
www: <http://www.otago.ac.nz/informationscience/>

# Contextual Information Retrieval in Research Articles: Semantic Publishing Tools for the Research Community

M.A. Angrosh, Stephen Cranefield, Nigel Stanger  
*Department of Information Science, University of Otago, Dunedin, New Zealand*

## Abstract.

Over the last few years, the voluminous increase in the academic research publications has gained significant research attention. Research has been carried out exploring novel ways of providing information services using the research content. However, the task of extracting meaningful information from research documents remains a challenge. This paper presents our research work carried out for developing intelligent information systems, exploiting the research content. We present in this paper, a linked data application which uses a new semantic publishing model for providing value added information services for the research community. The paper presents a conceptual framework for modeling contexts associated with sentences in research articles and discusses the Sentence Context Ontology, which is used to convert the information extracted from research documents into machine-understandable data. The paper also reports on supervised learning experiments carried out using conditional probabilistic models for achieving automatic context identification.

Keywords: Semantic Publishing Models, Sentence Context Ontology, Linked Data Application, Conditional Random Fields, Maximum Entropy Markov Models, Citation Classification, Sentence Context Identification

## 1. Introduction

In recent years, there has been a dramatic increase in research output by scientists across the globe. A comparative analysis of published research during the years 1996-2002 and 2002-2008 by Research4Life [1], an organization which offers health, agriculture and environmental research for free or at a subsidized price to developing countries, observed 194% or 6.4-fold increase in articles published in peer reviewed journals. Furthermore, a recent report by UNESCO observes that developing countries more than doubled their annual spending on research and development activities between 2002 and 2007, from \$135 billion to \$274 billion, leading to a drastic increase in research output [2]. While this overwhelming increase in research output has certainly benefitted the research community, it has also

brought in its wake various challenges that need to be addressed in order to obtain optimum value from these invaluable resources. It is becoming increasingly difficult to keep abreast of research developments in one's fields due to the wide range of research outputs coupled with complexities of inter-disciplinary research activities.

Institutional digital libraries and information content providers are beginning to establish the required infrastructure for tracking research developments. However, there still remains a larger gap in the provision of information services using the content extracted from research documents. Current information search services for research content mainly based on bibliographic metadata are not intelligent enough to understand the meaning of the content in research documents and thus are not capable of providing services based on contextual data. Identifying

the limitations of current digital libraries, Shum et al. observed that none of the digital libraries were capable of providing information about the publications that support and challenge a given document and would not be able to trace the intellectual lineage for a given idea [3]. In order to offer such services, it is required to identify, extract and manage meaningful information about the content embedded in the research document, which is not readily available.

Besides providing information about the number of publications which support and challenge a given document, it would be beneficial to provide researchers with contextual information about how a given document is cited by other documents. Besides eliminating the tedious and time-consuming task of looking into each cited document to learn about the contexts in which the work is cited, such information services would also facilitate easier and more meaningful understanding of the cited work. In order to provide such information services, it is necessary to extract contextual information associated with sentences in research documents. This contextual information would also help in offering a wide range of services. For example, it would be possible to see the context of citation sentences in a given article in a single view. Furthermore, while it would also be possible to learn about citation contexts of works of individual authors, it would also pave ways for developing systems that could trace the intellectual lineage for a given idea. Against this backdrop, the present study is taken up for developing intelligent information systems based on the context associated with sentences in research articles.

Preliminary reports on this work have been published previously [4-7]. While [4] reports on modelling the contexts of sentences in related work sections of research articles and supervised learning experiments for context identification, [5] describes the ontological modelling of these contexts. The framework was extended to cover citation sentences appearing throughout the article and an ontology has been developed for modelling these contexts. Experiments with supervised learning methods were carried out and an information retrieval system was developed that used SPARQL queries. The work has been submitted as a journal paper and is currently under review [6]. Further, recently we have submitted a conference paper that proposes an argumentation map and uses the same for information retrieval [7]. The paper is currently under review.

The key focus of this paper is to develop a linked data application that provides intelligent information services using the extracted information from re-

search article. To this end, we begin by proposing a framework for defining the contexts associated with sentences in research articles. We proceed to develop Sentence Context Ontology for modelling these contexts. We present our experiments carried out with supervised learning methods for achieving the task of sentence context identification. Finally, we describe the linked data application developed for volumes published in the European Semantic Web Conference (ESWC) series.

## 2. Related Work

In order to achieve the key objective of this study i.e., to develop intelligent information systems using the context associated with sentences in research articles, we use techniques from the following different areas.

1. Citation Content Analysis
2. Machine Learning
3. Semantic Web Initiative for modelling Scientific Discourse

The following sections describe the prior work in these areas.

### 2.1. Citation Context Analysis

In recent times there has been a lot of interest in identifying and using the citation context for providing information services. We categorize the research work in this field into the following three areas: citation classification schemes, automatic extraction of citation contexts and using citation contexts for information retrieval.

#### 2.1.1. Citation Classification Schemes

Several studies have focused on identifying the reasons for citations in research articles. As early as 1965, Garfield identified fifteen different reasons for authors to cite other works [8]. Based on an analysis of 30 research articles in theoretical high energy physics, Moravcsik and Murugesan proposed a classification scheme consisting of four categories [9]. Nanba and Okumara presented a simplified citation classification scheme involving three categories [10]. Recently Teufel et al. presented an annotation scheme for classification of citations involving twelve categories [11].

### 2.1.2. Automatic Extraction of Citation Contexts

Nanba and Okumara used cue phrases for extracting ‘citing areas’ in research papers. The citing areas are defined as a succession of sentences that have a connection with the sentence that includes the citation in the paragraph [10]. The study created the citing area corpus by hand and applied n-word gram analysis to this corpus. The study developed 160 rules for automatic determination of citation types. The rules were based on 84 cue phrases extracted from the corpus.

Nanba et al. described methods for classifying research papers using citation information [12]. The authors proposed bibliographic coupling using citation types that identified problems or gaps in cited works as an effective way of classifying research papers.

Garzone and Mercer presented an automated citation classifier, which involved a pragmatic grammar consisting of 195 lexical matching rules and 14 parsing rules that was developed based on cue words extracted from a citation and its location in the article [13]. Pham and Hoffmann developed a Knowledge Acquisition Framework for Tasks in Natural Language (KAFTAN), capable of acquiring cue phrases for classifying citations [14]. Mercer and Marco extended the work of Garzone and Mercer [13] to propose the use of fine-grained cue phrases within citation sentences for classifying these citation sentences [15].

Teufel et al. presented an annotation scheme and employed machine learning techniques for achieving automatic classification of citation sentences following the annotation scheme [11].

Kaplan et al. experimented with co-reference chains for extracting citations from research papers and achieved 7-10% precision as compared to the cue-phrase-based technique. The study created a corpus of citations comprising citing papers for four cited papers [16].

There have also been several studies using conditional probabilistic models such as Conditional Random Fields (CRFs) and Maximum Entropy Markov Models (MEMMs) for extracting information related to citations, which are discussed in the following section.

### 2.1.3. Using Citations Contexts for Information Retrieval

Nanba and Okumara investigated the automatic generation of a review article based on citation information and relationships [10]. The study developed a prototype using citation relationships. The system identified the citing areas and the type of citing relationships and used this information for citation-based topical clustering of papers.

Nanba et al. [12] extended the prototype of Nanba and Okumara [10] by including support for classifying research papers based on citation types. Ritchie et al. conducted experiments using terms from citations for scientific literature search [17]. The authors used terms used by citing documents to describe a document, in combination with the terms of the document itself. The authors found that the combination of terms yielded better retrieval performance than standard indexing of the document terms alone.

### 2.2. Machine Learning Experiments using Conditional Probabilistic Models

The present study views the task of context identification as a sequential classification problem. The sequential classification is achieved by using conditional probabilistic models such as Conditional Random Fields (CRF) and Maximum Entropy Markov Models (MEMMs). Various experiments have been carried out using these models for extracting bibliographic and citation information from documents.

Le et al. used Hidden Markov Models and Maximum Entropy Markov Models for identifying citation types [18]. The authors noted that this method of using finite state machines required neither user interactions nor explicit knowledge about cue phrases and thus provided flexibility for extension. Feng and McCallum used CRFs for extracting various common fields from the headers and citations of research papers [19]. Hirohata et al. employed CRFs for identifying rhetorical roles in scientific abstracts. They carried out experiments to classify sentences in scientific abstracts into four sections – objective, methods, results and conclusions and achieved an accuracy of 95.5% per sentence and 68.8% per abstract [20]. French et al. used CRFs for automatic extraction of brain region mentions in neuroscience literature. Using a rich feature set derived from morphological, lexical, syntactic and contextual information, the study showed that CRFs performed well compared to dictionary methods [21].

Zou et al. conducted experiments using CRFs and Support Vector Machines (SVMs) for locating and parsing bibliographic references in HTML medical

articles [22]. While a CRF was used to model the word sequence, the SVM was focused on classifying individual words in the references. The study noted that both the classifiers achieved about 97% accuracy at chunk level. Gao et al. have developed a parser of bibliographic information in Chinese electronic books using CRFs [23]. Lopez used CRFs for extracting bibliographical references in patent documents [24]. The author observed that CRFs achieved better performance compared to rule-based algorithms by reducing the error rate by 75%. Council et al. have developed ParsCit – an open source tool, which besides identifying reference strings, identifies their citation contexts. The tool uses a trained CRF model for labelling the token sequences in the reference string [25]. Zhang et al. employed CRFs for extracting bibliographic fields such as author, title, journal, year from citations. Using a subset of open-access PubMed Central articles, the study achieved an overall 97.95% F-Score [26].

### 2.3. Semantic Web Initiatives for Modelling Scientific Discourse

Researchers in the field of the Semantic Web have also shown interest in modelling scientific discourse. The SWAN project (Semantic Web Application in Neuromedicine) has developed the SWAN Ontology – a knowledge schema for personal and community organization and annotation of scientific discourse [27]. The SWAN Ontology includes the Citations Ontology for defining a set of entities useful for referencing scientific publications [28]. The Bibliographic Ontology (bibo) was developed for defining the various constructs of bibliographic data [29]. CiTO – the Citations Typing Ontology – was developed for describing the nature of reference citations in research articles [30]. Groza et al. (2007) have proposed the SALT – Semantically Annotated LaTeX – framework for annotating research documents [31]. As part of the framework, the study combines three ontologies – the Document ontology, the Rhetorical ontology and the Annotation ontology for achieving this task.

## 3. The Rationale and Contributions for this Study

### 3.1. Why an application based on citation contexts?

Though there have been several studies on identifying citation contexts and using this information for

providing information services, there still does not exist a robust application that fully exploits the citation context information. The key focus of this study is to develop systems for automatic context identification and demonstrate the use of this information through developing a robust application. In order to achieve this we define a framework defining contexts associated with citation sentences and non-citation sentences. The justification for defining our own set of contexts is provided in the following section.

### 3.2. Why another set of citation contexts?

Even though there are different citation classification schemes available as mentioned in Section 2, the present study resorted to defining its own set of citation contexts as explained in Section 4. The available classification schemes are developed for specific disciplines and create difficulties in applying them to other disciplines. White observes that most of these classification schemes are idiosyncratic and are hard to code, resulting in difficulties for using them across literatures [32]. The citation contexts identified in the present study resulted after manually analyzing 331 citation sentences from 20 research articles selected from the Lecture Notes in Computer Science (LNCS) collection at springerlink.com [33], which formed our training dataset. The process of defining contexts also included identifying features present in each of these citation sentence that would justify the defined context for a given citation sentence. Thus, based on the presence of these features the new set of citation contexts was evolved. We explain in Section 6, the various features defined in the study. Further, besides defining contexts for citation sentences, we also define contexts for non-citation sentences. The contexts for non-citation sentences were defined after manually analyzing 838 sentences extracted from the training set of 20 research articles. Section 4 describes in detail the different contexts defined for sentences in our study. The proposed framework also facilitates in developing our Sentence Context Ontology for deriving machine-understandable data. The justification for the Ontology is provided in the following section.

### 3.3. Why Sentence Context Ontology

While there have been efforts for building ontologies for modelling scientific discourse, these ontologies have focused on specific entities in research articles. For example, the focus of the SWAN ontology is to model research statements and research ques-

tions [27]. The Bibliographic Ontology provides a more formal way of describing bibliographic details of documents [29]. Besides providing for modelling bibliographic details, the Citations Typing Ontology (CiTO) takes one step further to include different reasons for citations in research documents [30]. The key focus of this study is to identify contexts associated at sentence level in research documents and use this information for providing intelligent information services. However, to the best of our knowledge, there doesn't exist an ontology which describes the contexts associated with different types of sentences in research documents.

Therefore we developed the Sentence Context Ontology for modelling contexts associated with sentences. We explain in this paper the conceptual basis on which the ontology is developed and illustrate how the ontology is used for developing intelligent information retrieval tools for the research community.

### 3.4. Key Contributions of this paper

The following forms the key contributions of this paper:

1. We propose a framework for defining contexts associated with sentences in research articles. The framework is described in Section 4.
2. We developed the Sentence Context Ontology based on the above framework. The ontology is described in Section 5.
3. We carried out machine learning experiments using the labels resulting from the framework for achieving automatic identification of contexts associated with sentences. The details of these experiments are provided in Section 6.
4. We developed a linked data application for research papers published in the proceedings of the European Semantic Web Conference (ESWC) series. Section 7 provides details of the linked data application and explains the unique services provided by this application

## 4. Identifying Contexts associated with Sentences in Research Articles – Conceptual Framework

A research article can be viewed as a collection of different sections appropriately placed in relation to

each other for presenting the author's research work. The individual sections in the article are a collection of different paragraphs, with each paragraph comprising a set of sentences. Sentences in research articles can be broadly categorized into two different types – citation sentences and non-citation sentences. While citation sentences point to an external publication for various reasons for expressing the author's ideas, non-citation sentences have their own meanings and contexts associated with them. The present study distinguishes between citation sentences and non-citation sentences based on the following definition.

Citation sentences are defined as those sentences that have a reference to a published or unpublished source. Specifically, this is an expression in the sentence that points to an entry in the bibliographic references section of the article for the purpose of acknowledging the cited work. This expression can either be a numeric expression such as '[1]', '[1, 2]' etc. or author names used in the sentence for referring to the cited work. For example, in the sentence 'Toulmin proposed the ...', the word 'Toulmin' is the name of the author and is used to refer to the cited work. Non-Citation Sentences are defined as those sentences that do not have any expressions as defined above.

Instead of considering all sentences of an article, the present study limits its focus only to those paragraphs that have citation sentences. We assume that these paragraphs are sufficient to provide a rich representation of the article that can be used for delivering unique information services. Nanba and Okumara identify passages with citation sentences as 'citing areas' and note that these passages provide a summary of the cited paper from the current author's viewpoint [10]. Further, a citation sentence is always associated with one or more sentences in the article. Furthermore, even though a single citation sentence alone may appear as a paragraph in the article, such a sentence is usually related to either the preceding or the following paragraph in the article.

Figure 1 shows our framework for modelling contexts of sentences in paragraphs with citation sentences in research articles. The framework is developed based on the generic rhetorical pattern observed in these paragraphs. As seen in the Figure, citation sentences with different contexts (light shaded blocks) are either preceded or followed by non-citation sentences (dark shaded blocks) with different contexts. The study defines the following contexts associated with non-citation sentences and citation sentences in research articles.

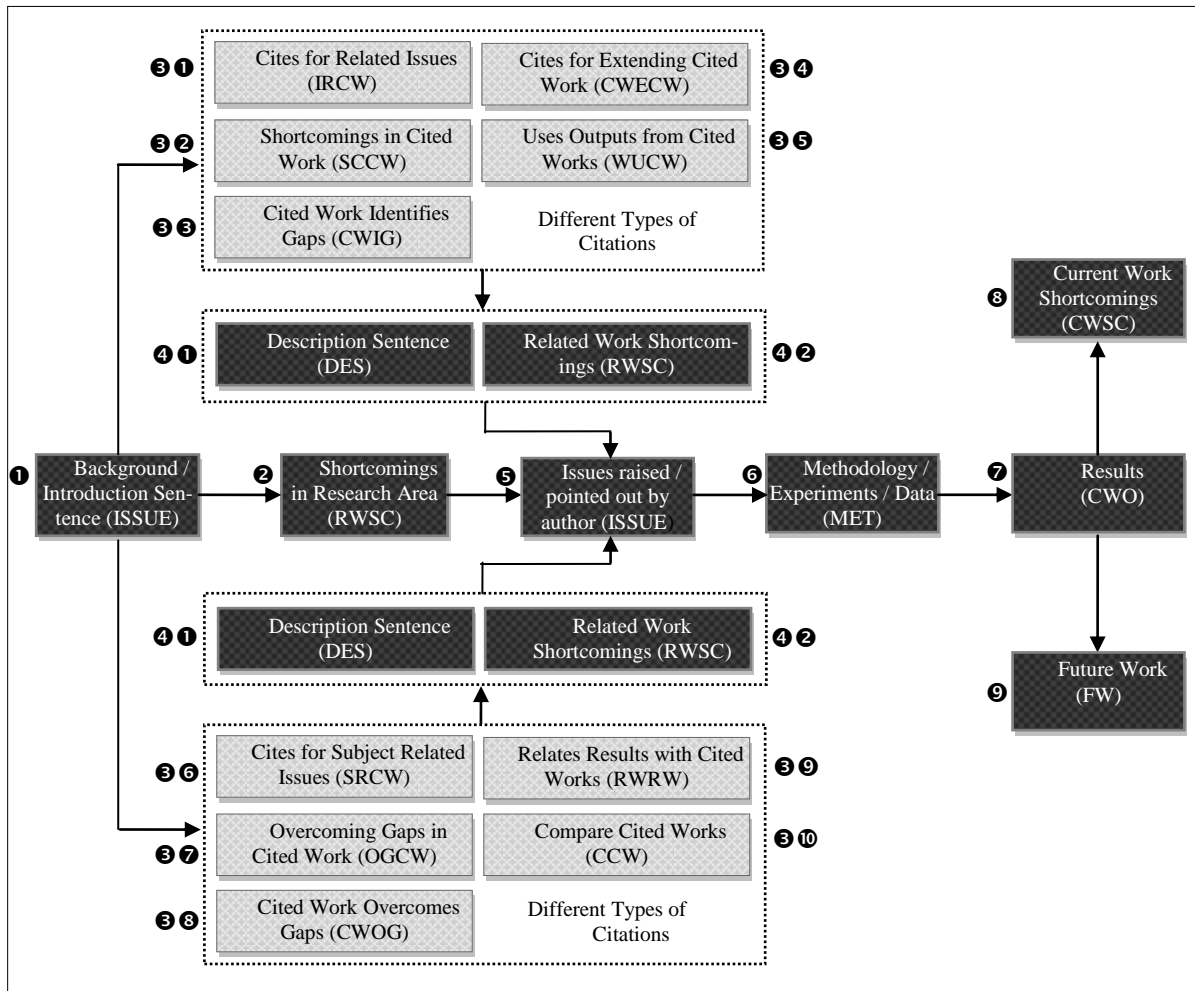


Figure 1: Contexts of sentences in paragraphs with citation sentences in research articles

■ Non-Citation Sentence    □ Citation Sentence

#### 4.1. Contexts associated with Non-Citation Sentences

A variety of contexts could be associated with a non-citation sentence. For example, it could be an introduction sentence, introducing the reader to the research ideas addressed in the article or a background sentence providing the background of the research ideas. It can also be a shortcoming sentence, identifying gaps in the research area or a cited work. The dark shaded blocks in Figure 1 identify the different contexts associated with non-citation sentences are defined as follows:

##### 4.1.1. Issue Sentences (ISSUE)

The study considers different types of sentences as Issue sentences. This facilitates in having control over labels for sentences which otherwise would result in difficulties in carrying out machine learning experiments for context identification. Sentences with the following characteristics are considered as Issue sentences.

##### Background/Introduction sentences (Block 1)



These sentences are used to introduce the reader to the research article or provide background about issues addressed in the article and generally appear at the start of the paragraphs.

Such sentences generally precede citation sentences and are considered as issue sentences as they denote an introductory or background issue against which the author uses a cited work to progress his writing or argument.

#### *Issues raised or pointed out by the author (Block 5)*

These sentences denote an issue pointed out by the author in relation to the cited work and generally follow a citation sentence.

Generally in a research article, the author after citing a related work, points to issues of his interest. Such sentences fall into this category. While these sentences follow citation sentences can further form preceding issue sentences for the citation sentence that follow them. These sentences are characterized as issue sentences in this study.

#### *4.1.2. Shortcoming Sentences (RWSC)*

Shortcoming sentences are defined as those sentences that identify research gaps or shortcomings in the ideas dealt in the paper. These sentences form an important component in developing the author's argument. The study distinguishes between two different types of shortcoming sentences:

#### *Shortcomings in research area (Block 2)*

These sentences identify shortcomings or gaps in the research area being addressed in the research article and generally precede a citation sentence.

#### *Shortcomings in cited work (Block 42)*

These sentences identify shortcomings or gaps in the cited work used by the author in the research article and generally appear after a citation sentence.

#### *4.1.3. Description Sentences (DES) (Block 41)*

These sentences further describe the cited work used in the article and generally follow a citation sentence.

#### *4.1.4. Methodology Sentences (MET) (Block 6)*

These sentences denote the methodology used in the research article

#### *4.1.5. Current Work Outcome Sentences (CWO)*

These sentences denote the outcomes or results of the current paper.

#### *4.1.6. Future Work Sentences (FW) (Block 7)*

These sentences denote the future work that could follow on from the current paper.

#### *4.1.7. Current Work Shortcoming Sentences (CWSC) (Block 8)*

These sentences denote the shortcomings of the current paper.

#### *4.2. Contexts associated with Citation Sentences*

The contexts associated with citation sentences reflect the reason for referring to the cited work in the research article. The following defines the different contexts that are identified for citation sentences in the present study.

#### *4.2.1. Cites for Related Issues (IRCW) (Block 31)*

These are citation sentences, where the author uses the cited work to refer to issues in the research area of the research article.

#### *4.2.2. Shortcomings in Cited Work (SCCW) (Block 32)*

These are citation sentences wherein the author identifies shortcomings in the cited work.

#### *4.2.3. Cited Work used for Identifying Gaps (CWIG) (Block 33)*

These are citation sentences where the author uses cited work for identifying gaps in the research area addressed in the article.

#### *4.2.4. Current Work Extends Cited Work (CWEWCW) (Block 34)*

These are citation sentences wherein a statement is made about how the current work extends the cited work.

#### 4.2.5. Uses Outputs from Cited Works (WUCW) (Block 35)

These are citation sentences wherein the author refers to the outputs used from the cited work.

#### 4.2.6. Cites for Subject Related Issues (SRCW) (Block 36)

These are citation sentences, where the cited work is mainly used to refer to subject related works dealt in the research article.

#### 4.2.7. Overcomes Gaps in Cited Works (OGCW) (Block 37)

These are citation sentences wherein the current paper makes claims about overcoming the gaps identified by the current paper in the cited work.

#### 4.2.8. Cited Work Overcomes Gaps (CWOOG) (Block 38)

These are citation sentences wherein the author references cited works which overcome the identified gaps identified in the current paper.

#### 4.2.9. Results with Related Work (RWRW) (Block 39)

These are citation sentences wherein the results of the current paper are compared with the cited works.

#### 4.2.10. Compare Cited Works (CCW) (Block 310)

These are citation sentences wherein the results or works of the cited works are compared.

### 5. Sentence Context Ontology

While in the previous section, we defined various contexts that could be associated with a given sentence, it is also necessary to define relations between these sentences. For example, in the sample paragraph provided in Figure 2, each sentence is related to the adjacent sentences. If we specifically consider the second sentence in the paragraph, we would notice that the third sentence is a shortcoming sentence identifying shortcomings in the cited works, cited in the second sentence. Also, the second sentence has a preceding and following citation sentence in sentence 1 and sentence 4 respectively. Further, each of these citation sentences is related to a specific cited work, the details of which are provided in the references section of the article. In order to model these relations, we propose the Sentence Context Ontology, which forms our vocabulary for modelling contexts of sentences in research articles.

There is a large amount of literature addressing the problem of automated composition of web services. However, most of the approaches address composition at the functional level (see, e.g. [12, 4]), and much less emphasis has been devoted to the problem of process-level composition. Different planning approaches have been proposed to address the problem of on-the-fly composition, from HTNs [17] to regression planning based on extensions of PDDL, to STRIPS-like planning for composing services described in DAML-S [15]. However, none of these techniques addresses the problem of composing web services with conditional outputs, non-nominal outcomes, and with process models describing interaction protocols that include conditional and iterative steps. In [8, 11, 7], the authors propose an approach to the automated composition of web services based on a translation of DAML-S to situation calculus and Petri Nets. Also in these papers, however, the automated composition is limited to sequential composition of atomic services, and composition requirements are limited to reachability conditions.

■ Citation Sentence in Question      ■ Shortcoming Sentence      ■ Preceding and Following Citation Sentence

Figure 2: Example Paragraph from Pistore et al.[34]

SENTCON, the Sentence Context Ontology, is an ontology for describing the context of sentences in scientific research articles with a specific focus on citation sentences and their adjacent sentences. Though SENTCON has been initially designed for

application to research articles published in the Lecture Notes in Computer Science series, it can easily be extended for other domains. SENTCON is developed using the Web Ontology Language (OWL) [35].

In order to use various properties of research articles, SENTCON imports the Bibliographic Ontology [29] with a namespace <http://purl.org>. The ontology was developed using Protege 4.0.2, the ontology editor and knowledge-base framework [36]. The

ontology is shown in diagrammatic form in Figure 3. The following section provides details of SENTCON.

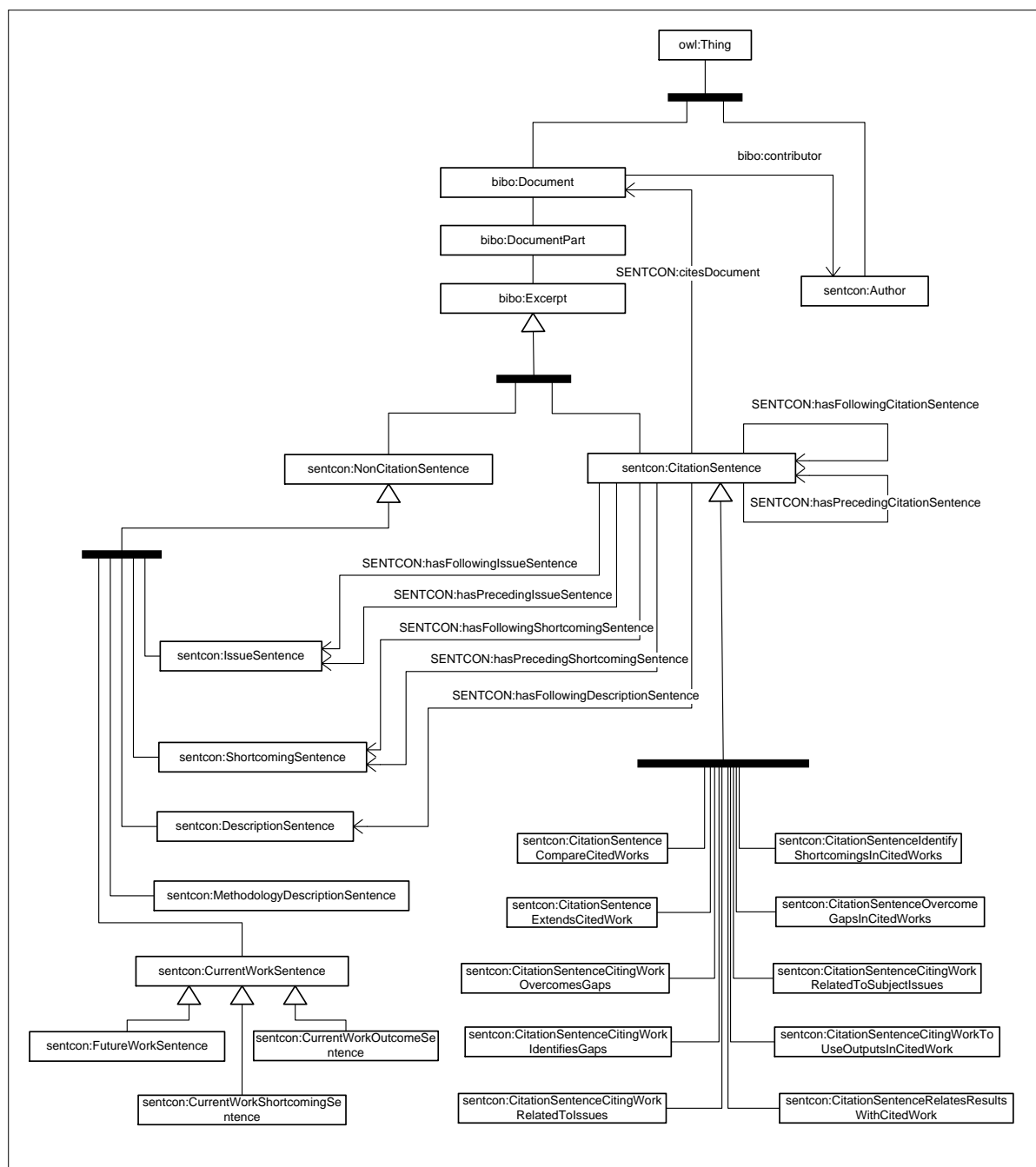


Figure 3: Sentence Context Ontology

### 5.1. SENTCON – Scope and Usage

The primary purpose of SENTCON is to facilitate modelling contexts of sentences in research articles, with a key focus on citation sentences and their adjacent sentences and to publish this in machine-readable formats such as Resource Description Framework (RDF). Figure 4 shows a schematic diagram resulting from modelling the sample paragraph provided in Figure 2 using the SENTCON ontology.

The key classes of SENTCON are the Citation Sentence class, the Non-Citation Sentence class and the Author Class. The Citation Sentence class and the Non-Citation Sentence classify various contexts associated with sentences in research articles as de-

scribed in Section 3 and the Author class defines authors of published articles and cited articles. The Sentence Class and the Non-Citation Sentence Class are defined as subclasses of the bibo:Excerpt class which is defined as ‘a passage selected from a larger work’ in the Bibliographic Ontology.

The bibo:Excerpt class is a subclass of the bibo:DocumentPart Class, which in turn is a subclass of bibo:Document Class in the Bibliographic Ontology. The key classes of SENTCON are as shown in Table 1.

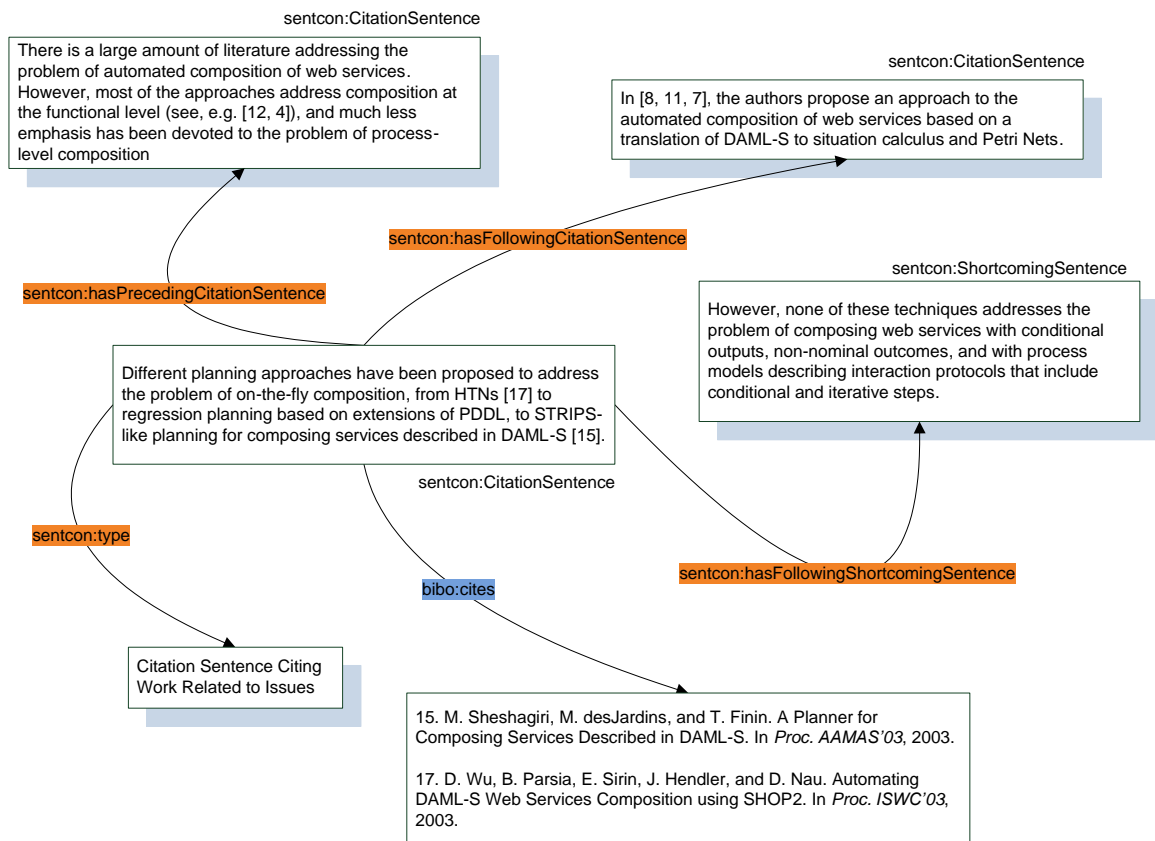


Figure 4: Modelling a Sample Paragraph using Sentence Context Ontology

Table 1: Key classes of SENTCON

Class	Membership Condition
sentcon:CitationSentence $\subseteq$ bibo:Except	<i>A citation sentence in research article</i>
sentcon:NonCitationSentence $\subseteq$ bibo:Except	<i>A non-citation sentence in research article</i>
sentcon:Author	<i>Authors associated with research article; instances include both citing authors and cited authors</i>

### 5.2. The sentcon:CitationSentence Class

The sentcon:CitationSentence class defines various subclasses for describing different contexts associated with citation sentences in research articles .

The subclasses of Citation Sentence class are listed in Table 2. Each of these subclasses characterize a specific context as explained earlier in Section 3

Table 2: Subclasses of sentcon:CitationSentence Class

Subclasses of Citation Sentence class	Description
sentcon:CitationSentenceCompares CitedWorks	<i>Citation sentences that compare cited works</i>
sentcon:CitationSentenceExtends CitedWork	<i>Citation sentences that extends current work with cited works</i>
sentcon:CitationSentenceCitesWorks IdentifyingGaps	<i>Citation sentences that cite works which identify gaps in the research area addressed in the article</i>
sentcon:CitationSentenceCitesWorks OvercomingGaps	<i>Citation sentences that cite works which overcome the identified gaps</i>
sentcon:CitationSentenceCitesWorks RelatedToIssues	<i>Citation sentences that cite works related to issues addressed in the research article</i>
sentcon:CitationSentenceIdentifies ShortcomingsInCitedWork	<i>Citation sentences that identifies shortcomings or research gaps in the cited work</i>
sentcon:CitationSentenceOvercomeGaps InCitedWork	<i>Citation sentences that state how the current work overcomes shortcomings or research gaps identified in the cited work</i>
sentcon:CitationSentenceCitesWorks RelatedToSubjectIssues	<i>Citation sentences that cites works related to subject issues addressed in the research paper</i>
sentcon:CitationSentenceUsesOutputs InCitedWork	<i>Citation sentences that discuss how the current work uses outputs from the cited work</i>
sentcon:CitationSentenceCompares ResultsToCitedWork	<i>Citation sentences that compares results of the current work to the cited work</i>

SENTCON defines various properties for relating instances of Citation Sentence class. These sentences characterize the relations between citation sentences and non-citation sentences in research articles.

Table 3 lists various properties of Citation Sentence class.

Table 3: Properties of sentcon:CitationSentence Class

Property	Domain	Range
bibo:cites	sentcon:CitationSentence	bibo:Document
sentcon:hasFollowingCitationSentence	sentcon:CitationSentence	sentcon:CitationSentence
sentcon:hasPrecedingCitationSentence	sentcon:CitationSentence	sentcon:CitationSentence
sentcon:hasFollowingIssueSentence	sentcon:CitationSentence	sentcon:IssueSentence
sentcon:hasPrecedingIssueSentence	sentcon:CitationSentence	sentcon:IssueSentence
sentcon:hasFollowingShortcomingSentence	sentcon:CitationSentence	sentcon:ShortcomingSentence
sentcon:hasPrecedingShortcomingSentence	sentcon:CitationSentence	sentcon:ShortcomingSentence
sentcon:hasFollowingDescriptionSentence	sentcon:CitationSentence	sentcon:DescriptionSentence

### 5.3. The sentcon:NonCitationSentence Class

The sentcon:NonCitationSentence class defines various classes for describing contexts associated with non-citation sentences in research articles

Table 4 lists the various subclasses of non-citation sentence class

Table 4: Subclasses of sentcon:NonCitationSentence Class

Subclasses of Non-Citation Sentence class	Description
sentcon:IssueSentence	<i>Non-Citation sentences that identify the issues addressed in the research paper. These could be either background issues or issues raised by the author of the article.</i>
sentcon:ShortcomingSentence	<i>Non-Citation sentences that refer to shortcomings or research gaps, which could be either in the related research area or the cited work in the research article.</i>
sentcon:DescriptionSentence	<i>Non-Citation sentences that further describe the earlier cited work.</i>
sentcon:MethodologyDescription Sentence	<i>Non-Citation sentences that refer to the methodology adopted in the research article</i>
sentcon:CurrentWorkOutcomeSentence	<i>Non-Citation sentences that refer to the outcome or results of the current paper</i>
sentcon:FutureWorkSentence	<i>Non-Citation sentences that refer to potential future work following from the current paper</i>
sentcon:CurrentWorkShortcoming Sentence	<i>Non-Citation sentences that refer to the shortcomings or research gaps in the current paper</i>

## 6. Machine Learning Techniques for Context Identification

We report in this section, the experiments carried out with Maximum Entropy Markov Models (MEMMs) and Conditional Random Fields (CRFs) using sixteen different labels resulting from the framework described in Section 3.

### 6.1. Maximum Entropy Markov Models (MEMMs)

MEMMs are variants of Hidden Markov Models (HMMs), wherein the observed state data is conditioned over observations instead of building a joint model of observation and states. HMMs are observed to suffer from two key problems: (a) they do not provide for incorporating features that allow richer representation of observations and (b) they follow a traditional approach of employing a generative model for solving a conditional problem with given observations. MEMMs were introduced for solving these problems [37]. MEMMs encode the probability distribution  $P_s(s|o)$ : the probability of making the transition to  $s$  from  $s'$  and observing  $o$ .

The maximum entropy distribution is a conditional exponential model of the form

$$P_s(s|o) = \frac{1}{Z(o,s)} \exp\left(\sum_a \lambda_a f_a(o,s)\right)$$

where  $\lambda_a$  are parameters to be estimated from the training data,  $f_a(o,s)$  are binary feature functions that capture important relations between the state and the observed sequence and  $Z(o,s)$  is the normalizing factor that makes the distribution sum to one across all next states  $s$ .

### 6.2. Conditional Random Fields (CRFs)

MEMMs are observed to suffer from label bias problems [38]. In order to overcome these problems CRFs were introduced. These are undirected graphical models that define a single log-linear probability distribution over label sequences given an observation sequence [38]. The structure of the graph in a CRF encodes independence relationships between labels and not the observations. This graphical structure also facilitates a functional form of the distribu-

tion. This function combines several different terms known as *clique potentials* into a single product, wherein each term forms a subset of the variables drawn from the full model.

More formally, let  $G$  be an undirected graph with edge set  $E$  and vertex set  $V$ . The conditional probability of the labels given the observations in a CRF factors according to the following equation

$$P(Y|X) = \frac{1}{Z_x} \prod_t \psi_t(y_{t-1}, y_t, X)$$

The normalization constant is computed by summing over all possible label sequences  $Y'$ , which is tractable for linear chain structures using dynamic programming:

$$Z_x = \sum_{Y'} \prod_t \psi_t(y_{t-1}, y_t, X)$$

Conditional Random Fields use a particular functional form for their clique functions:

$$\psi_t(y_{t-1}, y_t, X) = \exp(w^T f(y_{t-1}, y_t, X))$$

where  $w$  is a real-valued weight vector and  $f$  is a vector of feature functions. The weights  $w$  are the model parameters that are estimated during the training phase.

### 6.3. Feature Definition

The following are the three different kinds of features defined for our study:

#### 6.3.1. Citation Features

Citation features indicate whether a given sentence is a citation sentence. The distinction is made based on the presence of a citation reference in the sentence. References to citations in our dataset drawn from the LNCS collection is made using terms such as '[1]', '[11]', '[1, 11, 12]'. The application uses regular expressions for identifying the presence of these terms and decides whether the given sentence is a citation sentence or not.

In addition, references to citations can be made using referenced names of authors listed in the references section in the sentence without using the number reference. In such cases, the application looks for author names and terms such as 'et al.', to

decide about the status of the sentence. Thus, a feature ‘sentHasCitation’ is added to indicate that a given sentence is a citation sentence. A second citation feature ‘prevSentHasCitation’ is also defined to indicate that the previous sentence is a citation sentence.

### 6.3.2. Section Features

We defined various section features for indicating the section of the article to which the sentence belonged. In order to define section features, we adopted the following criteria. The content of research article was divided into three categories: the Introduction Block, the Body Block and the Conclusion Block. The sections of the article with headings ‘Introduction’, ‘Related Work’, ‘Overview’, ‘Motivation’ were considered under Introduction Block, the sections of the article under the heading ‘Conclusions and Future Work’ were considered under Conclusion Block. The other sections were considered under the Body Block\*.

\* It needs to be noted that the Related Work section in the article may appear anywhere in the article. Irrespective of its position, this section is considered under the Introduction block.

This demarcation is made in order to differentiate between citation sentences referring to research issues and subject issues dealt in the paper. Thus, the features ‘sentSec=Intro’, ‘sentSec=BGR’, ‘sentSec=RelWork’, ‘sentSec=Conc’ were defined to indicate that sentences belong to the Introduction, Background, Related Work and Conclusion sections of the paper respectively. A feature ‘sentSec=Sub’ is defined for sentences which do not belong to the above sections. This feature indicates that these sentences belong to the Body block, which represents the core subject of the paper.

### 6.3.3. Term Features

We followed a generalization strategy for defining term features for sentences. This involved identification of terms and phrases that indicated the context and meaning of the sentence. We defined eight categories as listed in Table 5 for identifying different kinds of terms. Accordingly 396 terms and phrases belonging to different categories were identified in the training dataset. Additionally these terms were also identified in the test dataset i.e. the ESWC collection and a total of 717 terms were identified in both training and test dataset. The details of the number of terms identified in each of these categories are provided in Table 5.

Table 5: Categories of Terms defined under Generalization Strategy

Category	Description	Example Terms	Terms Identified	
			(Training Dataset)	(ESWC + Training Dataset)
Connecting Terms (CT)	Terms or phrases that indicate relations between sentences. These terms, usually connect a sentence with its preceding sentence.	They, Therefore, According to these, For this purpose, Furthermore	51	54
Shortcoming Terms (SCT)	Terms or phrases that describe the shortcomings or gaps.	Nevertheless, performance suffers, perform poorly, are not studied	92	308
Methodology Terms (MET):	Terms or phrases that describe the methodology adopted or followed in the paper.	we consider, we use, we assume	87	88
Result Terms (RES)	Terms or phrases that describe the results achieved either by the current paper or the cited paper.	we will show, we discover, we summarize	80	118
Future Work Term (FWT)	Terms or phrases that describe the future work of the paper	future work, we plan to extend, will be investigated	34	49



Overcoming Gap Terms (OGT)	Terms or phrases that describe the characteristic of overcoming the identified gaps or shortcomings	enhanced, superior, promising, improved, better potential	23	49
Identifier Terms (IDT)	Terms or phrases that identify gaps or shortcomings in the related work or the cited work.	as shown, observations in, according to	15	35
Extending Terms (EXT)	Terms or phrases that discuss extending the current work with cited work.	builds on previous work, Similar to	04	04
Comparing Terms (COM)	Terms or phrases that mention comparison studies.	compared, evaluated	10	12
<b>Total number of terms identified using generalization strategy</b>			<b>396</b>	<b>717</b>

The process of feature selection using citation features, block features and term features resulted in 15 different features for sentences as listed in Table 6.

Table 6: Features defined for Sentences in Research Articles

Feature	Description
<b>Citation Features</b>	
sentHasCitation	Sentence has citation
prevSentHasCitation	Previous sentence has citation
<b>Block Features</b>	
sentSec=Intro	Sentence belong to the Introduction section of the article
sentSec=BGR	Sentence belongs to the Background section of the article
sentSec=RelWork	Sentence belongs to Related Work section of the article
sentSec=Sub	Sentence belongs to Body block
sentSec=conc	Sentence belongs to Conclusion block
<b>Term Features</b>	
sentHasTerm=CT	Sentence contains a connecting term or phrase
sentHasTerm=SCT	Sentence contains a shortcoming term or phrase
sentHasTerm=MET	Sentence contains a methodology term or phrase
sentHasTerm=RES	Sentence contains a resulting term or phrase
sentHasTerm=FWT	Sentence contains a future work term or phrase
sentHasTerm=OGT	Sentence contains a overcoming gap term
sentHasTerm=IDT	Sentence contains an identifier term
sentHasTerm=EXT	Sentence contains an extending term
sentHasTerm=COM	Sentence contains a comparing term

#### 6.4. Dataset

The dataset was developed from 20 research articles selected from the LNCS collection at springer-link.com [33]. The training set of 20 research articles yielded 250 paragraphs with citation sentences, which resulted in 1162 sentences. Each paragraph was represented as a sequence of sets of features and was manually assigned one of the labels signifying its context.

#### 6.5. Training CRFs and MEMMs

A 10-fold cross validation was performed. Mallet, a Java-based package that provides an implementation of linear chain CRF and MEMM algorithms, was used for training CRF and MEMM [39]. In the case of CRFs, we used two different types of CRF structures: a first-order linear chain and a combination of first-order and zero-order. While a first order linear chain uses distinct copies of features for each transition from state  $y_{t-1}$  to state  $y_t$ , zero-order features are dependent only on the current state  $y_t$ . Our earlier experiments with first-order linear chains show that first-order linear chain performs poorly for states which appear less often in the training dataset [4]. Therefore, we experimented using a combination of both zero-order and first-order features. Zero-order features can prove to be useful when used in addition to the first-order features, particularly in situations where sequences do not occur enough times in the training data. They provide a ‘back-off’ capability i.e., a source of information to use when the main source is not available.

#### 6.6. Results

The results of the classifier are tabulated in Table 7. While an accuracy of 93% was obtained using first and zero-order features in CRFs, a lower accuracy of 89% was obtained using first-order features alone in CRFs. The accuracy decreased to 68% with MEMMs. Further, while MEMMs failed completely for four different classes, CRFs with first-order features failed with two classes and CRFs using both first and zero-order features failed for one class respectively. The experiments show that a CRF with first order and zero-order features provide a suitable classifier for our classification task.

Table 7: Results of the Classifier

Label	Accuracy: 93.37%			Accuracy: 89.58%			Accuracy: 68.33%		
	CRF – 1 <sup>st</sup> & 0 Order			CRF – 1 <sup>st</sup> Order			MEMM		
	P	R	F	P	R	F	P	R	F
FW	1.00	0.96	0.98	1.00	0.93	0.96	1.00	0.64	0.78
CWO	0.96	1.00	0.98	0.92	0.90	0.91	0.00	0.00	0.00
ISSUE	0.97	0.97	0.97	0.93	0.96	0.95	0.79	0.93	0.86
METH	0.91	0.98	0.94	0.90	0.92	0.91	0.57	0.66	0.61
WUCW	0.94	0.92	0.93	0.90	0.94	0.92	0.86	0.84	0.85
SRCW	0.88	0.98	0.92	0.84	0.91	0.88	0.78	0.65	0.70
IRCW	0.89	0.96	0.92	0.81	0.95	0.87	0.31	0.54	0.39
RWSC	0.90	0.94	0.92	0.86	0.94	0.90	0.75	0.85	0.80
DES	0.89	0.89	0.89	0.85	0.86	0.85	0.75	0.59	0.66
CWIG	0.91	0.68	0.78	1.00	0.50	0.66	0.66	0.12	0.21
SCCW	0.86	0.67	0.76	0.77	0.50	0.60	1.00	0.14	0.25
CWOG	0.77	0.73	0.75	0.66	0.42	0.51	1.00	0.10	0.19
CWECW	1.00	0.50	0.66	1.00	0.25	0.40	0.00	0.00	0.00
CWSC	0.62	0.45	0.52	0.66	0.18	0.28	0.00	0.00	0.00
RWRW	1.00	0.12	0.22	0.00	0.00	0.00	0.00	0.00	0.00
CCW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

P = Precision; R = Recall; F = F-Score

## 7. Developing the Linked Data Application: Extracting and Generating RDF Data

The term ‘linked data’ coined by Tim Berners-Lee refers to a style of publishing and interlinking structured data on the web [40]. The basic characteristics of linked data are to use an RDF data model to publish data and use RDF links to interlink data from different sources [41]. In this section, we explain about the process of extracting information from research articles and generating the RDF data.

### 7.1. Extracting and Preparing the Data

The task of sentence context identification involves extraction of different types of data from research articles. This includes the following activities:

- Obtaining and processing PDF documents for extracting sentences and other information from the articles
- Identifying features and keywords in sentences
- Parsing the reference sections of articles for obtaining information about cited documents

The architecture of the system developed in order to achieve these tasks is provided in Figure 5. As seen in the Figure, a number of components were developed for extracting information from research articles. The following sections explain briefly each of these components.

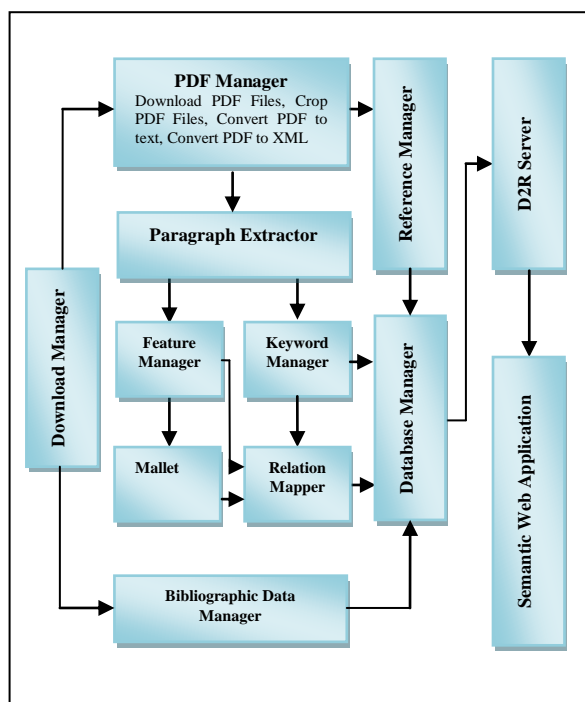


Figure 5: System Architecture of Information Extraction System

**Download Manager** – This module comprises XPath expressions and Java HTTP URLs and is used for web scraping Springerlink pages. The full-text documents and the bibliographic data from the ESWC collection is downloaded through our institutional license. While the full-text documents are stored in PDF form, the associated bibliographic data is stored in XML format.

**PDF Document Manager** – The PDF Document Manager handles PDF documents and extracts usable text for context identification. The following are the key functions of the PDF Document Manager.

- **Download PDF Files** – the full-text links of PDF documents provided by the Download Manager are used for obtaining PDF documents. The PDF documents are stored by their digital object identifiers (DOIs).

- **Crop PDF Files** – It is important to remove the header and footer information from PDF documents, as it might be difficult to identify these components after conversion to text format. This is carried out by cropping PDF files at a pre-designated margin.
- **Convert PDF to text to XML files** – The cropped PDF files are then converted to text and XML files. The processing of PDF files is performed using the batch processing mechanism of Adobe Acrobat.

**Paragraph Extractor** – The Paragraph Extractor extracts paragraphs with citation sentences from the text files. This Python module performs this activity by checking for presence of citations in paragraphs and accordingly obtains only those paragraphs with citations. The Extractor also identifies the section to which the extracted paragraph belongs. In order to achieve this, the PDF document bookmarks are used from the converted XML files, which provide section headings of the article. This is then used to identify the section to which the extracted paragraph belongs. The section information facilitates in defining section features for each sentence using the Feature Manager.

**Keyword Manager** – The Keyword Manager is responsible for extracting keywords from citation sentences obtained from the article. In order to achieve this, the system employs the `topia.termextract` python extraction library, which uses Parts-Of-Speech (POS) and simple statistical analysis for determining the terms and their strengths [42]. The Keyword Manager facilitates in building a list of keywords from citation sentences, which can then be used for information services.

**Feature Manager** – The Feature Manager receives paragraphs as text files from the Document Manager and is mainly responsible for generating features for each sentence in the text file. While this module uses NLTK [43] for carrying out sentence segmentation, it employs regular expressions for identifying the presence of different entities to generate features for a given sentence.

**Classifier** – The Classifier receives the features generated by the Feature Manager and uses it as test data to run against the classifier model obtained from the training data as discussed in Section 6. The

classifier returns labels for each of these feature sets, which indicate the context associated with sentences.

**Relationship Mapper** – The Relationship Mapper tags together the results obtained in the Document Manager, Feature Manager and the Classifier. Apart from this, the Relationship Mapper also links related sentences. For example if a shortcoming sentence follows a citation sentence, a relationship between these sentences is recorded.

**Reference Manager** – The Reference Manager is responsible for handling the bibliographic references of research articles. The functions of this module are as follows:

- **Extractor** – This extracts the reference section from the text files
- **Splitter** – Each of the individual references is identified from the extracted reference section
- **Term Identifier** – After identifying individual references, different terms in the reference such as author names, article title, article source are identified.

**Bibliographic Data Manager** – The primary function of the Bibliographic Data Manager is to handle the bibliographic details of research articles. The bibliographic details stored in XML format are handled by the module using lxml [44] – a library for working with XML and HTML in Python.

**Database Manager** – The Database Manager is responsible for storing data in the relational database management system. The system uses MySQL as the backend for storing data drawn from different sources.

**D2R Server** – The application uses D2R Server [45] for publishing linked data from the relational database. The mapping file of the D2R Server is appropriately configured in accordance with the SENTCON ontology for deriving RDF data.

**Semantic Web Application** – The resulting RDF data from the D2R server is used for developing the Semantic Web Application. The details of the application are discussed in the following section.

## 7.2. Data Model

Using the system described above, the research articles published in ESWC are processed to extract contextual information from these articles. Table 8 provides the details of the extracted data from the ESWC series. As seen from the table, presently, we have extracted data from five volumes published in the ESWC series. This provided a total of 241 articles<sup>^</sup>. Paragraphs with citation sentences were extracted from these articles and a total of 7242 sentences were extracted from these paragraphs. This included a total of 4175 citation sentences and 3067 non-citation sentences respectively. The details of different types of citation and non-citation sentences extracted in different volumes are provided in Table 9 and 10 respectively. Further, as may be seen in Table 8, a total of 12034 authors were extracted from cited works (citations) in these articles. Further, a total number of 4121 documents were cited by these articles.

---

<sup>^</sup> Few articles published in these five volumes were not considered for different reasons. While some were invited talks, few followed a different reference format. Some articles were not included due to unresolved errors while extracting information. We intend to resolve these errors.

Table 8: Details of Sentences extracted from ESWC volumes

Year	ESWC Volume	Articles	Sentences		Citations	
			Citation Sentences	Non-Citation Sentences	Authors	Cited Documents
2005	3552	42	805	570	2455	874
2006	4011	45	871	660	2277	844
2009	5554	76	1214	867	3513	1161
2010	6088	27	534	406	1564	519

2010	6089	51	747	564	2225	723
<b>Total</b>		<b>241</b>	<b>4175</b>	<b>3067</b>	<b>12034</b>	<b>4121</b>
<b>Total Number of Sentences</b>				<b>7242</b>		

Table 9: Details of Citation Sentences extracted from ESWC volumes

Year	ESWC Volume	Citation Sentences									
		A	B	C	D	E	F	G	H	I	J
2005	3552	286	355	31	8	87	0	32	0	6	-
2006	4011	306	336	28	39	83	0	62	0	11	6
2009	5554	463	438	30	66	137	0	61	0	15	4
2010	6088	231	167	7	37	48	0	38	0	7	3
2010	6089	347	240	10	41	55	0	46	0	7	1
<b>Total</b>		<b>1633</b>	<b>1536</b>	<b>106</b>	<b>191</b>	<b>410</b>	<b>0</b>	<b>239</b>	<b>0</b>	<b>46</b>	<b>14</b>
<b>Total Number of Citation Sentences</b>											<b>4175</b>

- A – Citation Sentence Cites Works Related to Issues
- B – Citation Sentence Cites Works Related to Subject Issues
- C – Citation Sentence Cites Works Identifying Gaps
- D – Citation Sentence Cites Works Overcoming Gaps
- E – Citation Sentence Identifies Shortcomings in Cited Work
- F – Citation Sentence Extends Current Cited Work
- G – Citation Sentence Uses Outputs in Cited Work
- H – Citation Sentence Overcome Gaps in Cited Work
- I – Citation Sentence Compares Results to Cited Work
- J – Citation Sentence Compares Cited Works

Table 10: Details of Non-Citation Sentences extracted from ESWC volumes

Year	ESWC Volume	Non-Citation Sentences					
		DES	RWSC	CWO	CWSC	FW	METH
	3552	159	237	101	32	11	30
	4011	162	282	128	45	12	31
	5554	206	386	157	52	17	49
	6088	91	177	76	27	8	27
	6089	139	247	98	40	10	30
<b>Total</b>		<b>757</b>	<b>1329</b>	<b>560</b>	<b>196</b>	<b>58</b>	<b>167</b>
<b>Total Number of Citation Sentences</b>							<b>3067</b>

- A – Description sentences
- B – Shortcoming sentences
- C – Current work outcome sentences
- D – Current Work Shortcoming Sentence
- E – Future Work Sentence
- F – Methodology Description Sentence

The data extracted from research articles are stored in different tables in the relational database as shown in Figure 6.

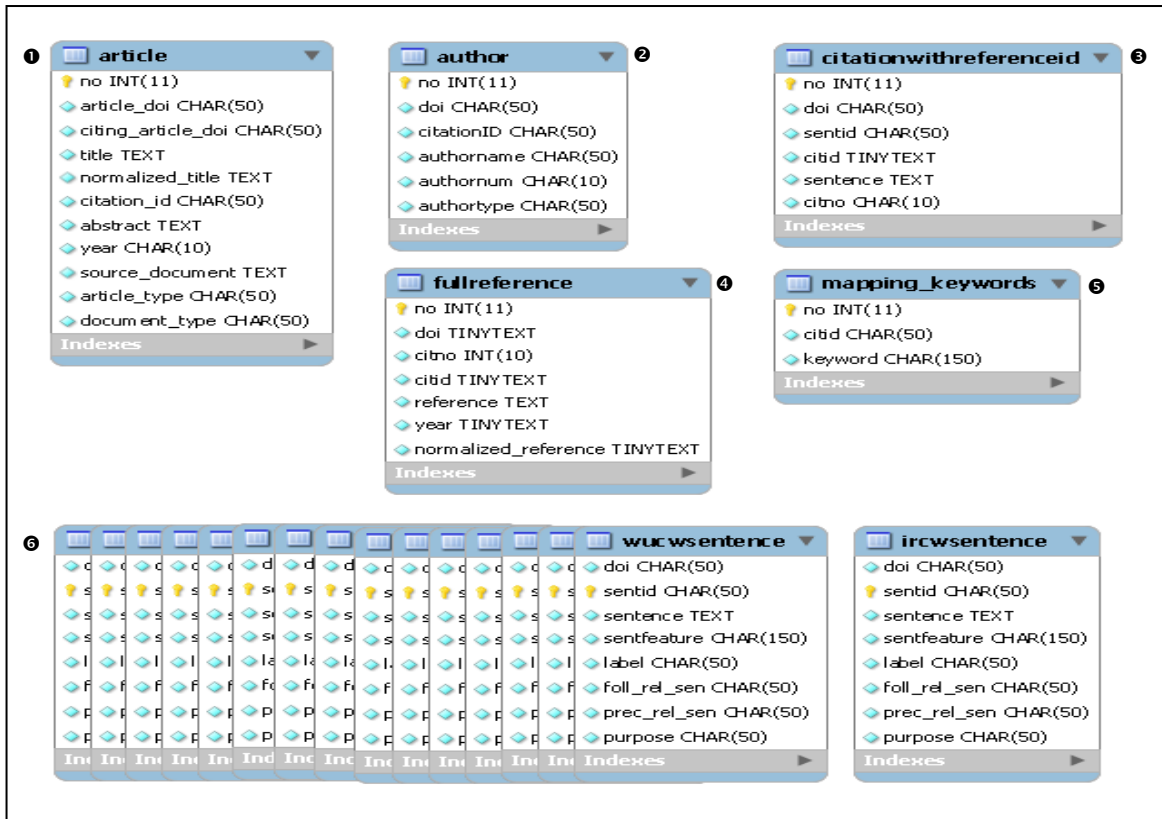


Figure 6: Different tables defined for storing data extracted from research articles

As may be seen in Figure 6, the article table (table 1) and the author table (table 2) hold data related to both citing articles as well as cited articles. A series of tables (tables indicated by 6) are created for storing each type of citation sentence separately. For example, the table titled ‘ircw sentence’ holds only citation sentences citing works related to issues. A table is also created for storing citation sentences with their reference id (table 3). This table is important as it creates a unique identifier for each of the references cited in the given sentence. For example, if there are two cited works in a given sentence, a unique identifier is created for each cited work and the citation sentence is associated with each identifier. Tables are also created for holding the full reference data (table 4) and the keywords associated with each citation sentence (table 5).

### 7.3. Generating RDF Data

#### 7.3.1. Choosing URIs

The use of D2R server for the data model described above facilitates in defining URIs for the database entities. The D2R Server assigns URIs for database entities using URI patterns [45]. As described above, the citationwithreferenceid (table 3) forms an important table as it holds data for each of the reference work cited in the given sentence, along with the citation sentence itself. The mapping file of the D2R Server is configured to define URIs for a citation sentence using this table. We also specify in the mapping file how data from different tables are related for a given citation sentence using properties defined in Sentence Context Ontology described in Section 5. The following provides details of the URIs defined for different entities in our application and also shows the resulting RDF data for these entities.

#### 7.3.2. URI for Citation Sentence

The pattern ‘citationwithreferenceid/@@citationwithreferenceid.num@@’ produces a relative URI ‘citationwithreferenceid/1001’ by in-

serting the number of the given citation sentence. Thus, the following URI:

‘https://info-nts-12.otago.ac.nz:8090/page/citationwithreferenceid/1001’ provides information about the citation sentence with number ‘1001’. The resulting RDF data for citation sentence numbered 1001 is as shown in Listing 1.

#### Listing 1

```
<rdf:Description rdf:about="citationwithreference/1001">
  <sentcon:purpose>Article Uses Cited Work for Research Issues</sentcon:purpose>
  <rdfs:label>Previous reports on our work contain additional details on the unsupervised miner [22], its application to a bio-medical corpus [21], and a qualitative evaluation [25].</rdfs:label>
  <sentcon:sentence>Previous reports on our work contain Additional details on the unsupervised miner [22], its application to a bio-medical corpus [21], and a qualitative evaluation [25].</sentcon:sentence>
  <bibo:doi>10.1007/11431053_38</bibo:doi>
  <rdf:type
    rdf:resource="sentcon/resource/SentenceCitingWorkRelatedToIssues"/>
  <sentcon:hasFollowingIssueSentence>To the best of our knowledge, so far only one other approach has been presented that addresses the quantitative and automated evaluation of an ontology by referring to its source corpus.</sentcon:hasFollowingIssueSentence>
  <dc:title>Lexically Evaluating Ontology Triples Generated Automatically from Texts</dc:title>
  <sentcon:keyword>Previous reports</sentcon:keyword>
  <sentcon:keyword>bio-medical corpus 1,</sentcon:keyword>
  <sentcon:keyword>qualitative evaluation</sentcon:keyword>
  <sentcon:keyword>unsupervised miner</sentcon:keyword>
</rdf:Description>
```

### 7.3.3. URI for Article Data

The application distinguishes between two types of articles: Citing Articles and Cited Articles. While citing articles are those articles published in the ESWC, cited articles are those that are cited in these articles. The article table in the database holds data related to both of these kinds and defining the mapping properties for the article table results in the URIs for article data.

The pattern ‘article/@@article.num@@’ produces a relative URI ‘article/1001’ by inserting the number of the respective article. Thus, the following URI:

‘https://info-nts-12.otago.ac.nz:8090/page/article/1’ provides information about the article with number ‘1’. The result-

ing RDF data for the citing article numbered 1 and cited article numbered 1001 is as shown Listing 2 and Listing 3 respectively. The different properties defined for both citing articles and cited articles can be seen in these listings.

#### Listing 2

```
<rdf:Description rdf:about="article/1">
  <rdfs:label>Automatic Location of Services</rdfs:label>
  <dc:title>Automatic Location of Services</dc:title>
  <sentcon:year>2005</sentcon:year>
  <sentcon:documentType>Citing Article</sentcon:documentType>
  <rdf:type rdf:resource="sentcon/resource/article"/>
  <dc:creator>Uwe Keller</dc:creator>
  <dc:creator>Ruben Lara</dc:creator>
  <dc:creator>Holger Lausen</dc:creator>
  <dc:creator>Axel Polleres</dc:creator>
  <dc:creator>Dieter Fensel</dc:creator>
  <bibo:abstract>The automatic location of services that fulfill a given need is a key step towards dynamic and scalable integration. In this paper we present a model for the automatic location of services that considers the static and dynamic aspects of service descriptions and identifies what notions and techniques are useful for the matching of both. Our model presents three important features: ease of use for the requester, efficient pre-filtering of relevant services, and accurate contracting of services that fulfill a given requester goal. We further elaborate previous work and results on Web service discovery by analyzing what steps and what kinds of descriptions are necessary for efficient and usable automatic service location. Furthermore, we analyze intuitive and formal notions of match that are of interest for locating services that fulfill a given goal. Although having a formal underpinning, the proposed model does not impose any restrictions on how to implement it for specific applications, but proposes some useful formalisms for providing such implementations.</bibo:abstract>
  <bibo:doi>10.1007/11431053_1</bibo:doi>
  <sentcon:authorType>Citing Author</sentcon:authorType>
</rdf:Description>
```

#### Listing 3

```
<rdf:Description rdf:about="article/1001">
  <rdfs:label>Efficient semantic matching</rdfs:label>
  <dc:title>Efficient semantic matching</dc:title>
  <sentcon:year>2005</sentcon:year>
  <sentcon:documentType>Cited Article</sentcon:documentType>
  <rdf:type rdf:resource="sentcon/resource/article"/>
  <sentcon:normalizedReference>Efficient semantic matching</sentcon:normalizedReference>
  <sentcon:fullReference>14. F. Giunchiglia, M. Yatskevich, and E. Giunchiglia. Efficient semantic matching. In Proceedings of ESWC, 2005</sentcon:fullReference>
  <sentcon:citID>10.1007/11431053_21_14</sentcon:citID>
  <sentcon:citationSentence>As a matter of fact, [14] shows, that when we have conjunctive concepts at nodes (e.g., Images, Europe), these matching tasks can be resolved by the basic DPLL procedure in polynomialtime; while when we have full proposition concepts at nodes (example, Images, Computers Internet), the length of the original formula can be exponentially reduced by structure preserving transformations.</sentcon:citationSentence>
  <sentcon:authorType>Cited Author</sentcon:authorType>
  <dc:creator>M Yatskevich</dc:creator>
  <dc:creator>F Giunchiglia</dc:creator>
  <dc:creator>E Giunchiglia</dc:creator>
  <sentcon:citingArticleTitle>Web Explanations for Semantic Heterogeneity Discovery</sentcon:citingArticleTitle>
</rdf:Description>
```

## 8. CIRRA – Contextual Information Retrieval in Research Articles – Semantic Web Application using the Linked Data

We explain in this section the semantic web application which uses linked data for providing value

added information services for the research community.

The architecture of the application is shown in Figure 7.

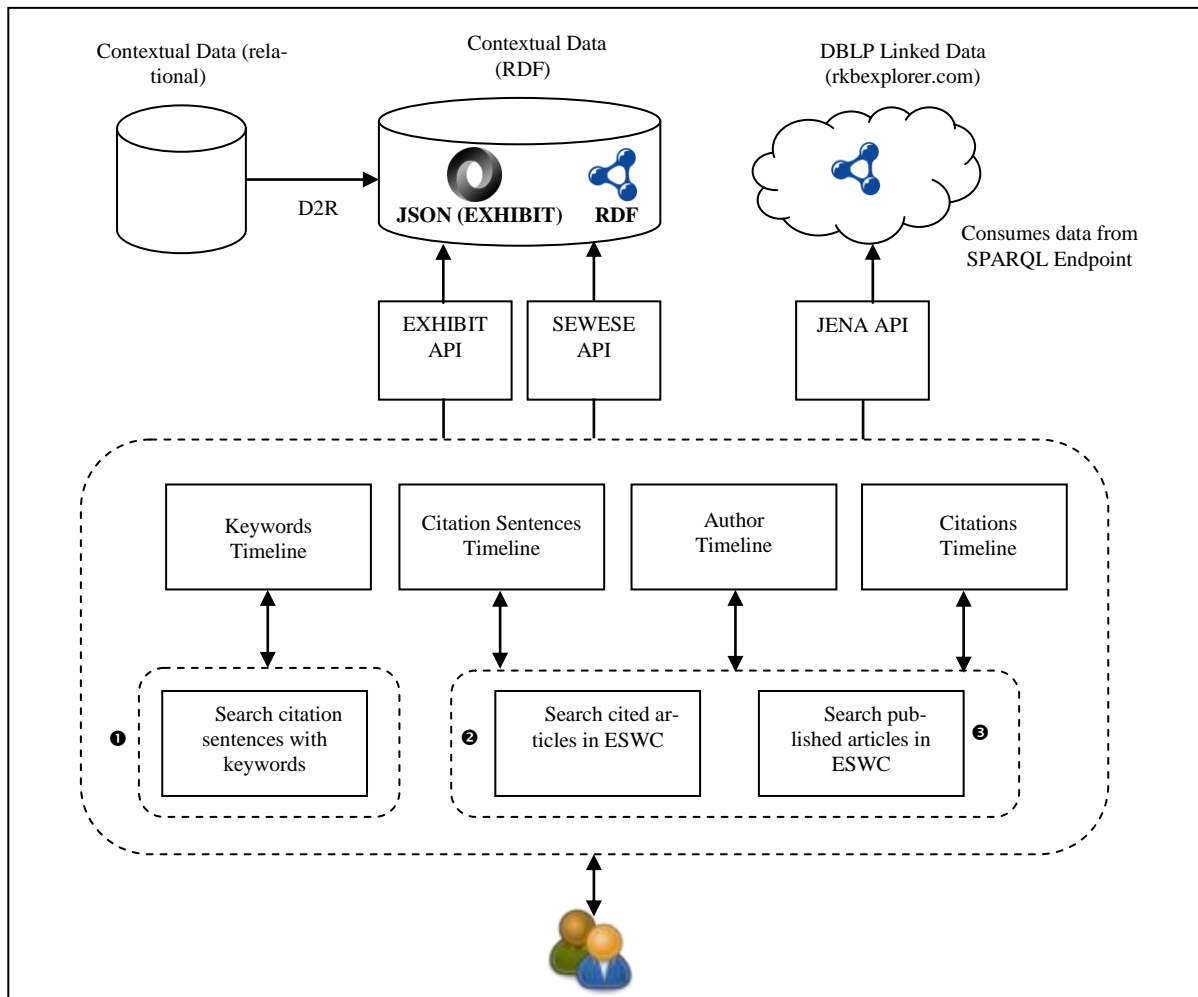


Figure 7: Architecture of CIRRA – A Semantic Web Application for Contextual Information Retrieval in Research Articles

The application uses Semantic Web Semantic Tags (SEWESE) [46] and the EXHIBIT API [47] for querying the RDF data resulting from the process described above. The use of these APIs facilitates the development of interactive user interfaces. SEWESE uses SPARQL queries for querying the RDF data. However The Exhibit API requires the RDF data to be converted into the JSON (Exhibit) [48] format before querying the data. The data is

therefore converted into the JSON (Exhibit) format using the Babel service provided by Exhibit [49]. The application uses Timeline [50], a web widget provided by Exhibit for visualizing temporal data in rich user interfaces. The application allows a user to search and browse the contextual data in the following three ways:

1. Search published articles in ESWC



2. Search articles cited by articles in ESWC
3. Search citation sentences using keywords

### 8.1.1. Search Published Articles in ESWC

The application supports searching published articles in ESWC. The interface facilitates keyword search, which retrieves published titles in ESWC along with author information from DBLP linked data available at the SPARQL endpoint (<http://rkbexplorer.com>). Figure 8 provides a screenshot of the results obtained for the search term ‘Se-

mantic Web’, retrieved from the SPARQL endpoint. As seen in Figure 8, retrieved titles from ESWC along with author information for the searched term are displayed. Further, for each of these titles, the interface also provides a link titled “View Contexts of Cited Works in the Article” (indicated by the label 1 in Figure 8), which allows users to navigate to ‘Citation Sentences Timeline’. This timeline allows users to view the contexts of all citation sentences in the selected article. The Citation Sentences Timeline is explained later in this section.

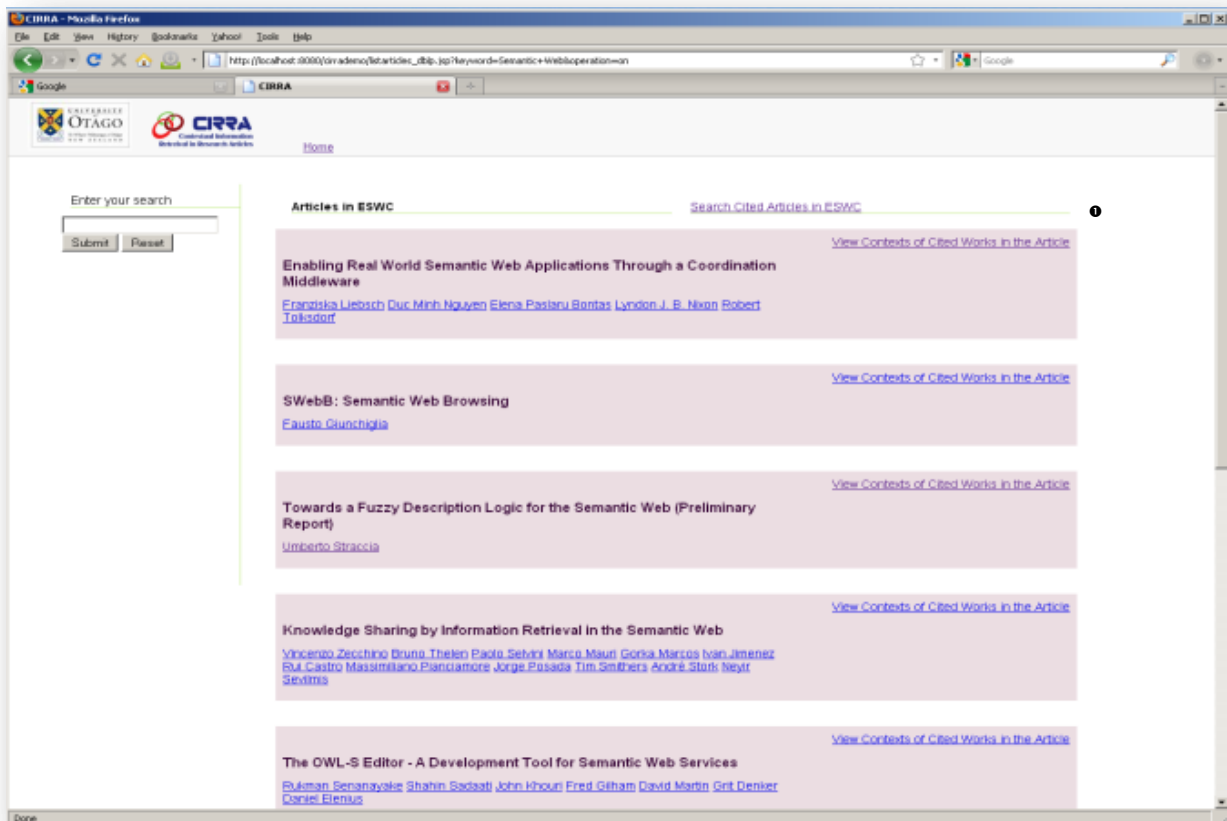


Figure 8: List of retrieved articles published in ESWC Collection

The application uses SPARQL queries for obtaining title and author information from the SPARQL endpoint at rkbexplorer.com. The SPARQL query shown in Listing 4 is used for retrieving titles; year and web address of each retrieved title for the keyword ‘Semantic Web’ from the ESWC collection. The web address is further used to retrieve authors for each retrieved titles. The SPARQL query shown in Listing 5 uses the web address for retrieving authors of individual title. The application uses Jena, A Java framework for building Semantic Web applications [51]. The application mainly uses Jena for interacting with the SPARQL Endpoint through the use of SPARQL queries.

#### Listing 4

```
PREFIX id: <http://dblp.rkbexplorer.com/id/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX akts: <http://www.aktors.org/ontology/support#>
SELECT distinct ?title ?year ?webaddress ?title1 WHERE {
  ?paper rdf:type akt:Book-Section-Reference .
  ?paper akt:has-title ?title .
  ?paper akt:has-date ?publishedyear .
  ?publishedyear akts:year-of ?year .
  ?paper akt:has-web-address ?webaddress .
  ?paper akt:article-of-journal ?journal .
  ?journal akt:has-title ?journaltitle .
  FILTER
  ((?journaltitle = "ESWC" || ?journaltitle = "ESWC (1)"
  || ?journaltitle = "ESWC (2)") && regex(?title, 'Semantic
  Web'))
}
```

#### Listing 5

```
PREFIX id: <http://dblp.rkbexplorer.com/id/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX akts: <http://www.aktors.org/ontology/support#>

SELECT distinct ?authors WHERE {
  ?paper rdf:type akt:Book-Section-Reference .
  ?paper akt:has-author ?author .
  ?author akt:full-name ?authors .
  ?paper akt:has-web-address ?webaddress .
  FILTER
  (?webaddress =
  "http://dx.doi.org/10.1007/978-3-642-13486-9_1")
}
```

#### *Citation Sentences Timeline*

The interface displaying the retrieved titles and author information from ESWC collection also provides a link for each article (indicated by the label 1 in Figure 8) for viewing the contexts of citation sentences. This allows the user to navigate to the ‘Citation Sentences Timeline’. This timeline allows users to view the contexts of citation sentences for the selected article in a single view. Figure 9 provides a screenshot of the Citation Sentences Timeline which displays the citation sentences along with their contexts on the timeline.

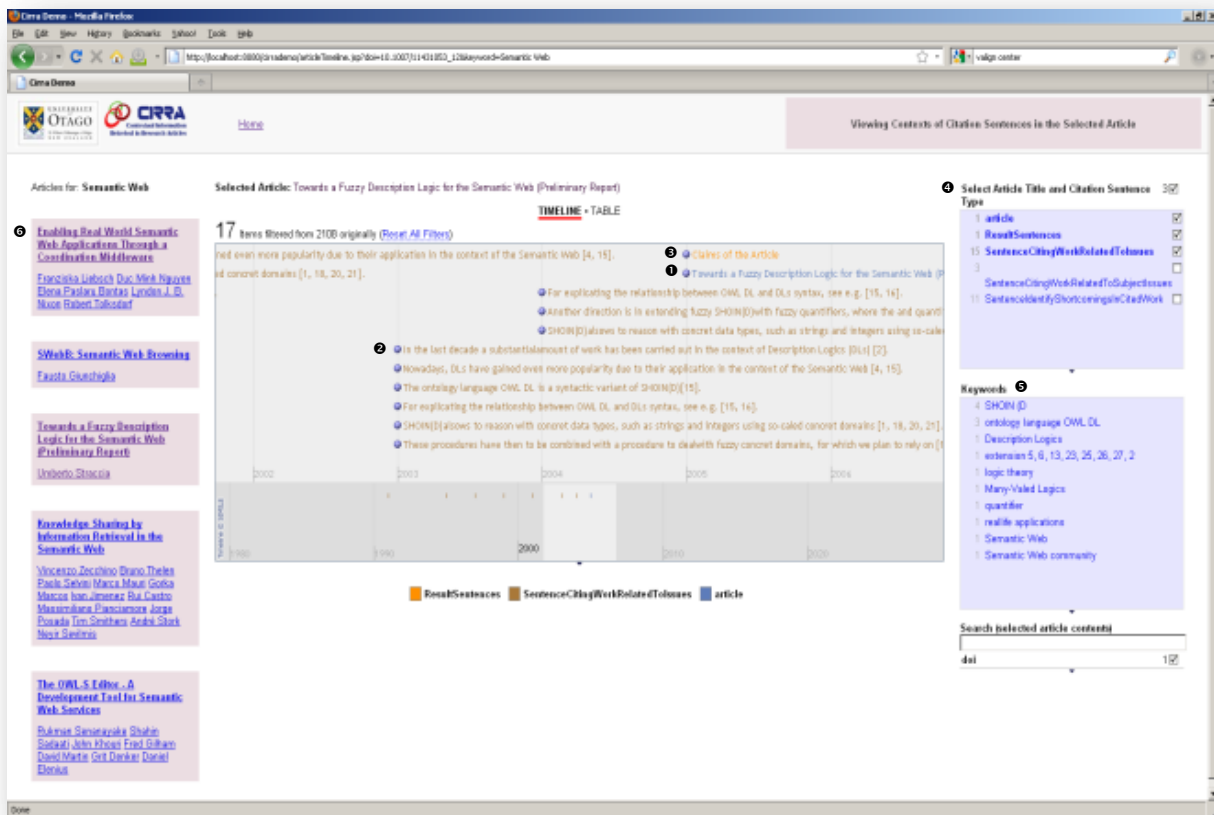


Figure 9: Screenshot of the Citation Sentences Timeline displaying citation sentences of the selected article

The citation sentences timeline allows for horizontally moving the timeline with the year of publication as the reference point. This provides a good interface for viewing the citation sentences of the article, placed with respect to the citations' year of publication on the x-axis of the timeline.

The following are the key features provided by the citation sentences timeline:

*View details of the selected article on the timeline*

The citation sentences timeline, which displays all citation sentences on the timeline also provides for viewing the bibliographic details of the selected article on the timeline. The selected article title ap-

pears on the timeline (indicated by label ❶ in Figure 9), which can be clicked to display the bibliographic details in the lens view. The details of the article title, authors and the abstract are displayed in the lens. Figure 10 shows the screenshot of the details of the article displayed for a given article. The names of the authors displayed in the lens are hyper-linked, and when clicked, these links navigate the user to the author timeline, which provides details about all works of an author across the collection. The author timeline is explained in detail later in this section.

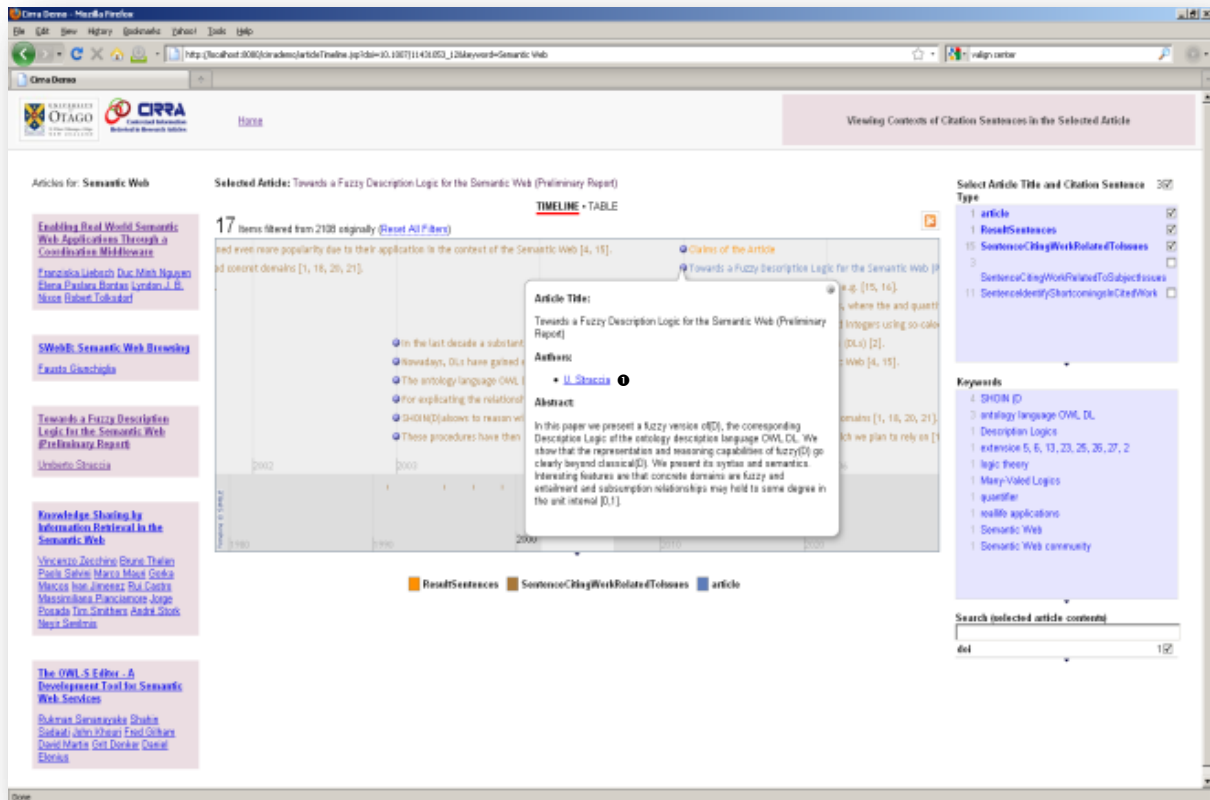



Figure 10: Viewing article details on the timeline

### View contexts of citation sentences on the timeline

The citation sentences timeline displays citation sentences of the article on the timeline, placed according to the year of publication of the cited work (indicated by the label  in Figure 9). Each of these citation sentences has a context, which is defined according to the subclasses of Sentence Context Ontology and is distinguished in the timeline by the use of different colours. Each of these citation sentences is clickable and when clicked provides details of the sentences related to citation sentence. The following are the different types of associated sentences displayed (if available) when the user clicks on a citation sentence:

1. Preceding issue sentence

2. Preceding shortcoming sentence
3. Preceding citation sentence
4. Following description sentence
5. Following shortcoming sentence
6. Following issue sentence
7. Following citation sentence

The properties relating the citation sentence class to non-citation sentence classes are used for relating the associated sentences to the citation sentence, and these relations are used for displaying associated information on the timeline. Figure 11 provides a screenshot of the citation sentences timeline, where a citation sentence is clicked for viewing the associated sentences.

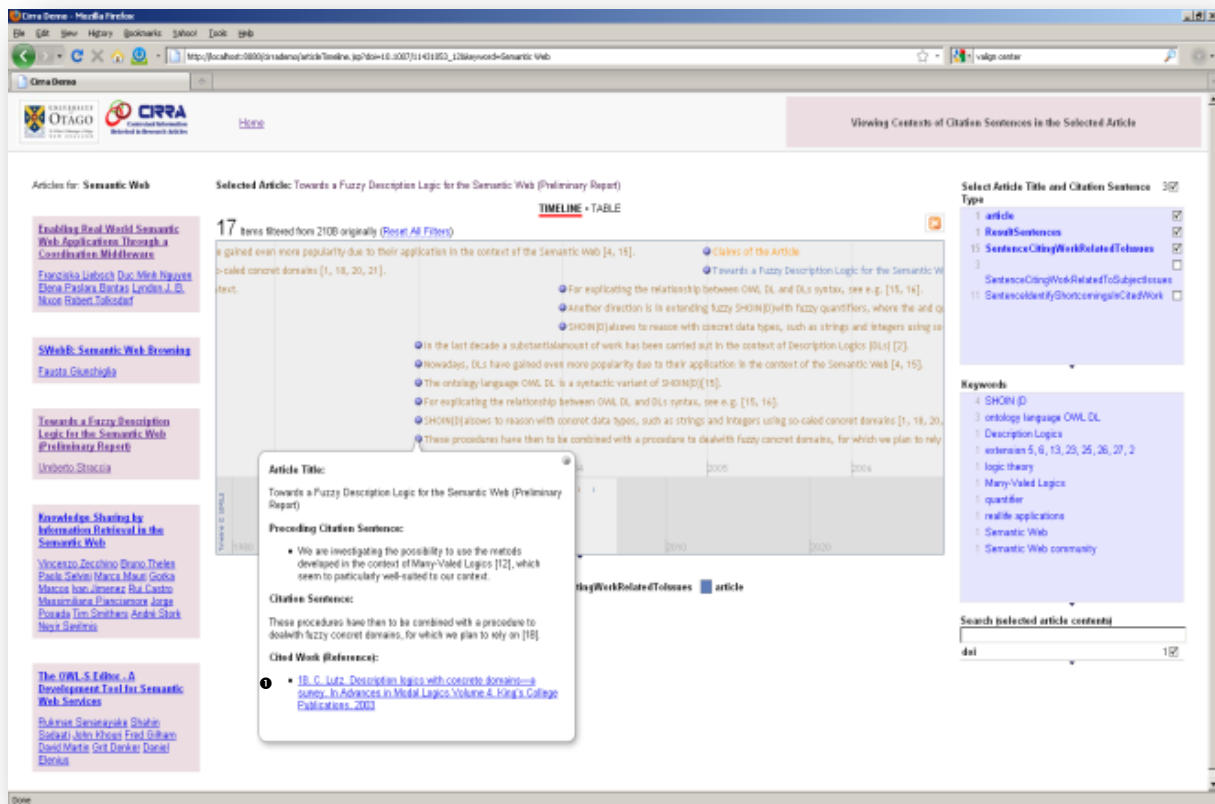


Figure 11: Screenshot of viewing associated sentences of a selected citation sentence

As seen in Figure 11, the associated sentences of a citation sentence are displayed in the lens view, which pops up, when the user clicks on the citation sentence. The lens also displays the full reference of the cited work used in the citation sentence. Clicking on the reference would navigate the user to citations timeline which provides details about how the selected cited work is cited across the entire collection. The citations timeline is explained later in this section.

#### View result sentences of the article on the timeline

The citation sentences timeline also facilitates in viewing sentences in the article that characterize results or outcomes of the selected article. A link titled ‘Claims of the Article’ appears on the timeline (indicated by the label ③ in Figure 9) and allows the

user to view all result sentences of the article. Figure 12 provides a screenshot of the timeline showing result sentences of the selected article.

#### Select article or citation sentence type

The interface displaying citation sentences of the article allows users to filter information on the timeline. The first facet on the right hand side (RHS) of the timeline with caption ‘Select Article Title and Citation Sentence Type’ (indicated by the label ④ in Figure 9) allows user to select the title or the required citation sentence types for display on the timeline. For example, if a user is interested in viewing only those citation sentences used to discuss research issues, he/she can select accordingly in the first facet to control the display of citation sentences on the timeline.

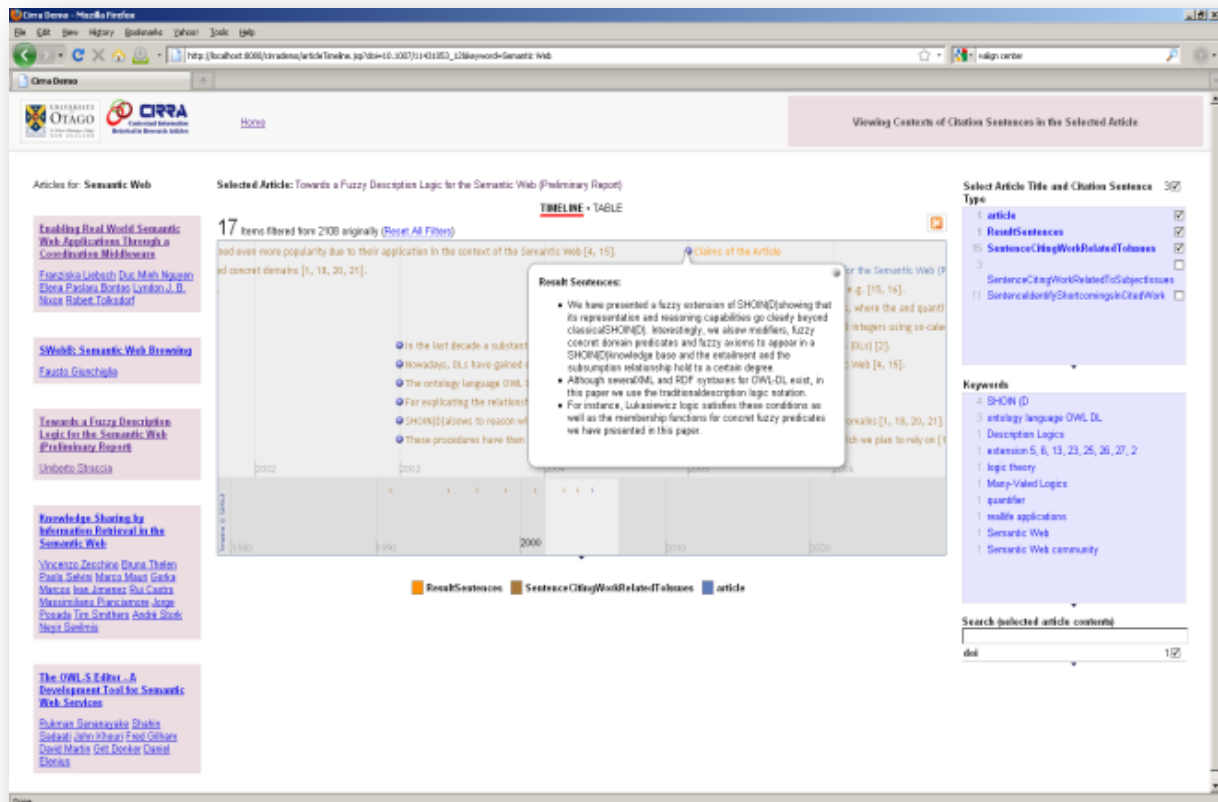


Figure 12: Citation sentences timeline displaying result sentences of the selected article

### Display citation sentences with specific keywords

The timeline also facilitates in displaying citation sentences with specific keywords. The second facet on the RHS of the timeline with the caption ‘Keywords’ (indicated by the label ⑤ in Figure 9) allows the user to filter citation sentences and choose to display only those citation sentences with the selected keywords.

### Select other articles to view its citation sentences

The interface also allows users to select other articles retrieved for the search term for viewing their citation sentences on the timeline. The article titles listed on the left hand side of the screen can be selected for displaying their citation sentences on the timeline (indicated by the label ⑥ in Figure 9).

### Navigation to the author and citations timelines

The interface also allows the user to navigate to the author timeline and citations timeline. The author names that appear when viewing the details of an article (indicated by the label ① in Figure 10) and the full reference that appears when viewing the full details of a specific citation sentence (indicated by the label ② in Figure 11) can be clicked to navigate to author and citations timelines respectively.

### Author Timeline

The author timeline forms an important feature of the application. This timeline facilitates in learning about the works of a selected author. The application currently distinguishes between a ‘citing author’ and a ‘cited author’. While citing authors are those who have published articles in ESWC, cited authors are those who have been cited in the published articles. The author timeline shows the titles of both published and cited works of the selected author. The application provides different lens views for

citing authors and cited authors. Figure 13 provides a screenshot of the author timeline showing the

works of a cited author.

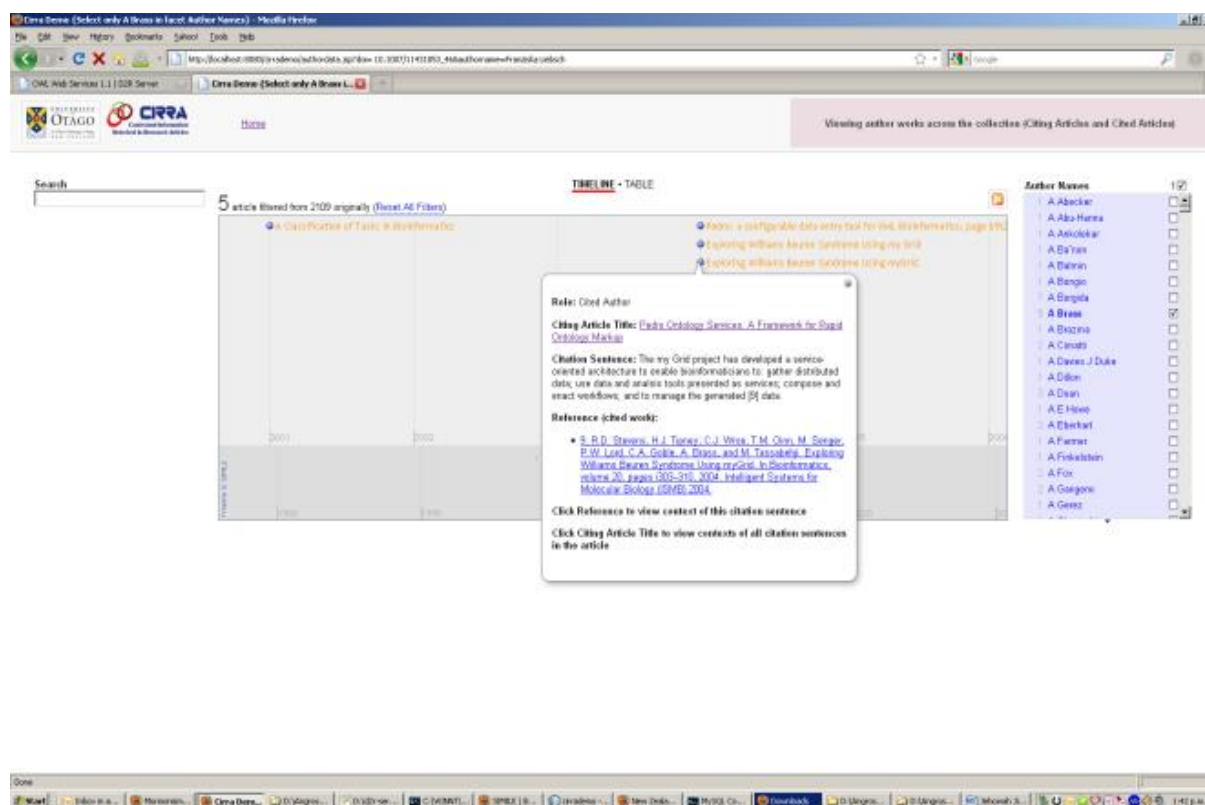


Figure 13: Screenshot of author timeline viewing details of cited work

As seen in Figure 13, the titles of the author’s works are displayed on the timeline and are placed according to their year of publication. For example, if the author’s work is published in the year 2000 and is cited two times, the title would appear twice on the timeline placed at year 2000. Each of these titles is hyperlinked and, when clicked, provides the following details in the lens view:

1. Role of the Author – Identifies the author as ‘Cited Author’
2. Citing Article Title – Shows the title of the article citing the selected work
3. Citation Sentence – Shows the citation sentence where the current work is cited

The user interested in learning more about the citation sentence can click on the full reference pro-

vided which would navigate the user to the citations timeline, where he can view the full details in which the selected work is cited. The citations timeline is explained later in this section.

The lens provides a different set of details with regard to work published by authors. Figure 14 provides a screenshot of the author timeline showing the published works of the selected author. The published titles of the authors are placed according to their year of publication on the timeline. Each of these titles is hyperlinked and when clicked, provides the following details in the lens view:

1. Role of the Author – Identifies the author as ‘Citing Author’
2. Citing Article Title – Shows the title of the article published by the author

3. Authors – Shows all the authors of the published work
4. Abstract – Shows abstract of the published work

The article title displayed in the lens view is clickable and when clicked, navigates the user to article timeline, where the user can view the contexts of all citation sentences as explained earlier in this section.

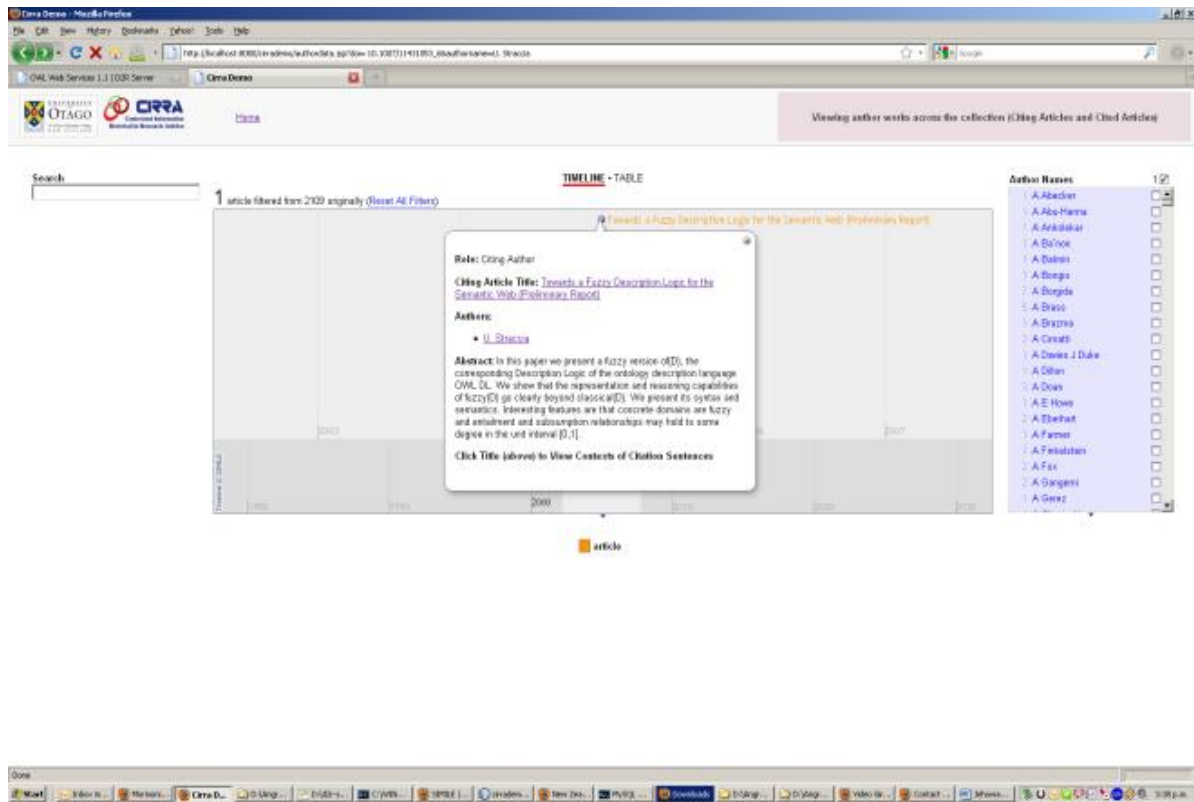


Figure 14: Screenshot of author timeline viewing details of published work

The other features of the author timeline are explained below:

The facet on the RHS of the screen provides a listing of both citing and cited authors in alphabetical order. The number preceding the authors in the facet indicates the total number of works of the author across the collection. This includes both published works and cited works.

The interface provides a search box on the LHS of the screen, which can be used by the user for searching the authors and article titles on the timeline.

#### Citations Timeline

The citations timeline forms an important feature of the application. The interface displays the different contexts in which the selected cited work is cited by different articles across the ESWC collection. Figure 15 provides a screenshot of the citations timeline.

As seen in Figure 15, the selected cited work is cited four times by three different articles and each of these contexts are displayed on the timeline, distinguished by the use of different colours. In order to achieve this functionality, the application creates a normalized title of each cited article and searches for this across the collection and displays the results. The citation sentences are placed according to the year of publication of the citing article on the timeline, in order to provide information about the year



when the work was cited. For example, a citation sentence placed on the year 2005 indicates the work was cited in the year 2005.

The first facet on the RHS with caption ‘Selected Reference’ indicates the normalized reference title currently selected.

The second facet on the RHS shows the full reference of the cited work as cited in the citing article. The numbers preceding the full reference in the facet indicate the number of times the cited work is cited in the article. The third facet shows other available references along with the number of times they have been cited across the collection.

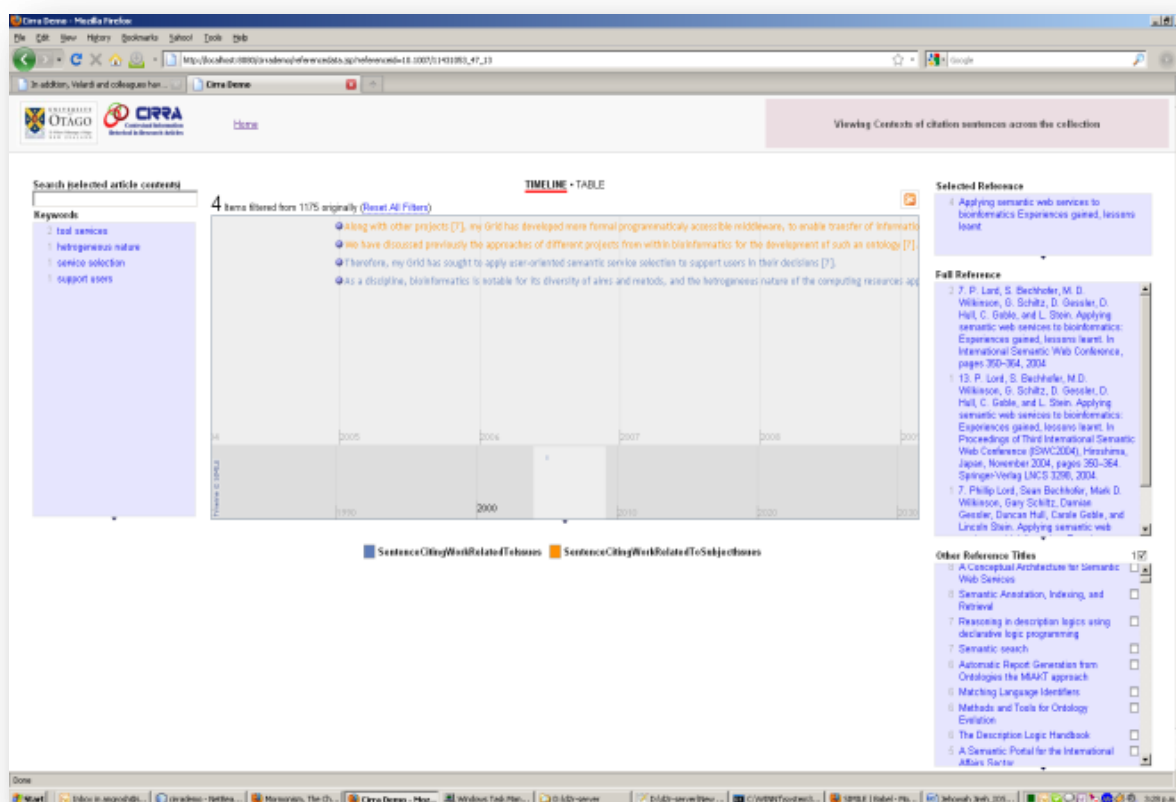


Figure 15: Screenshot of citations timeline

The facet on the left hand side provides a list of keywords extracted from the citation sentences. By selecting the keywords, users can filter the citation sentences and can opt to see only those citation sentences with selected keywords.

Each of the citation sentences displayed on the timeline are clickable and, when clicked, provide details of the citing article and associated sentences for the selected citation sentence. The citations timeline provides a unique feature in comparison to cur-

rent search engines by allowing users to see when and how the cited work is cited by other researchers.

### 8.1.2. Search Cited Articles in ESWC

The application also provides for searching cited documents in the published articles of ESWC collection. The interface employs SPARQL queries in order to display various details of cited documents. To start with, the title of the cited document for the searched term is obtained through the SPARQL query as shown in Listing 6:

**Listing 6**

```

SELECT
?title ?doctype ?citid ?articletype ?sourcedoc
WHERE {
  ?x rdf:type sentcon:Article .
  ?x sentcon:documentType ?doctype .
  ?x sentcon:articleType ?articletype .
  ?x sentcon:sourceDocument ?sourcedoc .
  ?x sentcon:citID ?citid .
  ?x dc:title ?title .
FILTER regex (?title, '${searchstring}', 'i')

```

The title of the cited article obtained in the above query is used to obtain the normalized title from the RDF data as shown in Listing 7.

**Listing 7**

```

SELECT
?normalized_reference
WHERE
{?x rdf:type sentcon:article .
?x sentcon:normalizedReference ?normalized_reference .
?x dc:title ?title .
FILTER regex (?title, '${title}', 'i')

```

The normalized title obtained above is used to retrieve all citation sentences which have a reference to this normalized title. For example, the SPARQL query shown in Listing 8 is used to retrieve all citation sentences from the class of ‘SentenceCitingWorkRelatedToIssues’ which use the normalized title. Similarly, other classes of citation sentence are checked for the use of the normalized title. This facilitates in identifying the different contexts in which the cited work is used in the article, which helps in providing unique services to researchers.

**Listing 8**

```

SELECT
?ircwsentence
WHERE {
?x rdf:type sentcon:SentenceCitingWorkRelatedToIssues .
?x sentcon:sentence ?ircwsentence .
?x sentcon:normalizedReference ?norm_reference .
FILTER
regex (?norm_reference, ${normalized_reference}', 'i')

```

Figure 16 provides a screenshot of the retrieved list of cited documents in the ESWC collection.

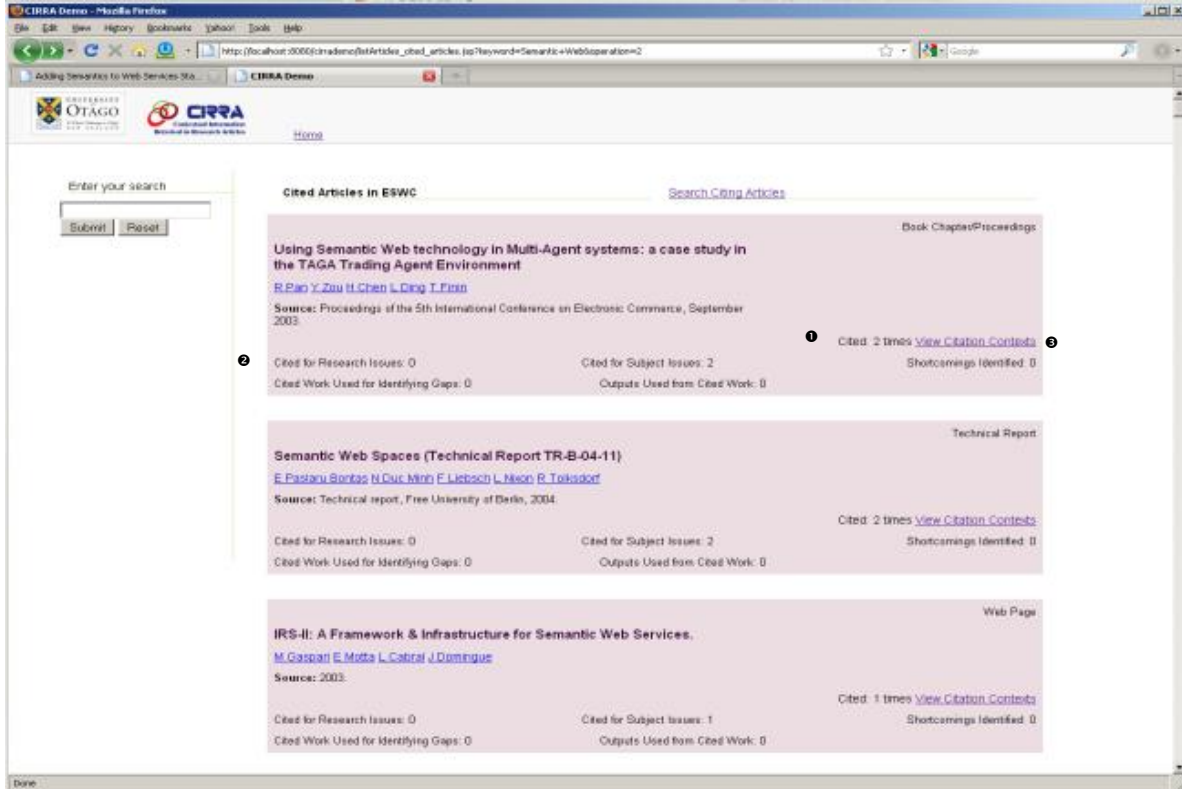


Figure 16: Screenshot of Retrieved articles from cited documents in the ESWC Collection

Besides providing bibliographic details of the articles, the interface provides the following useful metrics about how a specific work is cited in ESWC collection.

### Number of times a given work is cited

The application identifies the total number of times a given work is cited in the ESWC collection. The data indicated by the label ❶ in Figure 16 shows the total number of times the document is cited in the ESWC collection.

### Identify contexts of cited work

Besides providing the total number of times a document is cited, the application also provides details about the context in which the cited work is used. For example, if the document is cited six times, it could be cited for issues twice, and for subject issues twice and the author could have identified shortcomings in the cited work two times. The data indicated by ❷ in Figure 16 shows these details.

### Viewing citation contexts in the timeline

The interface also provides for viewing the contexts of citation sentences in the timeline view. The link with the caption ‘View Citation Contexts’ (indicated by the label ❸ in Figure 16) facilitates in navigating the user to the citation sentences timeline for viewing the full contexts in which the work is cited.

Figure 17 provides the screenshot of the citations timeline where the user can view the full context of the citation sentence. The citations timeline was explained in detail earlier in this section.

This timeline would provide in a single view details about how a document is cited by other articles in the ESWC collection. Users can click on the citation sentence to see associated sentences with the citation sentence. The facets on the right hand side facilitate in selecting other cited works.

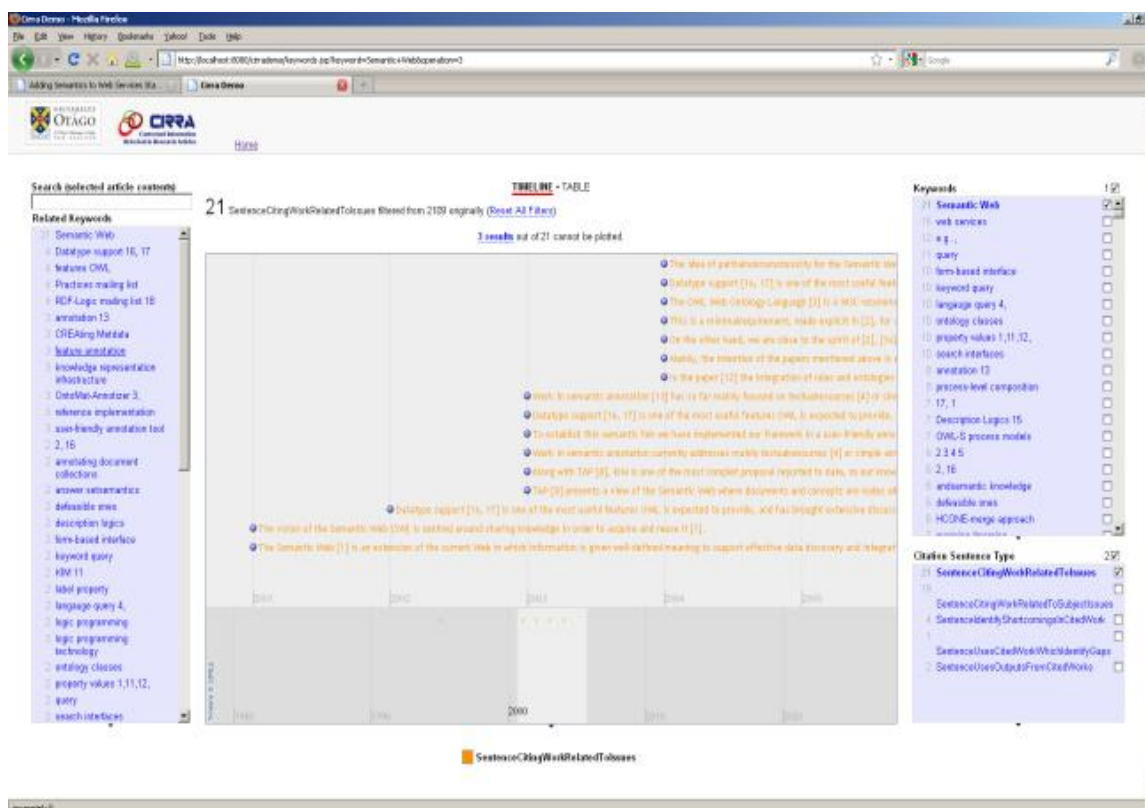


Figure 17: Viewing contexts of citation sentences in citation timeline

### 8.1.3. Search Citation Sentences with Specific Keywords

The application also facilitates in searching citation sentences for specific keywords across the ESWC collection. The keywords timeline displays

all citation sentences for the selected keyword on the timeline.

Figure 18 provides a screenshot where all citation sentences across the ESWC collection are displayed for the keyword ‘Semantic Web’. The following are the key features of the keywords timeline.

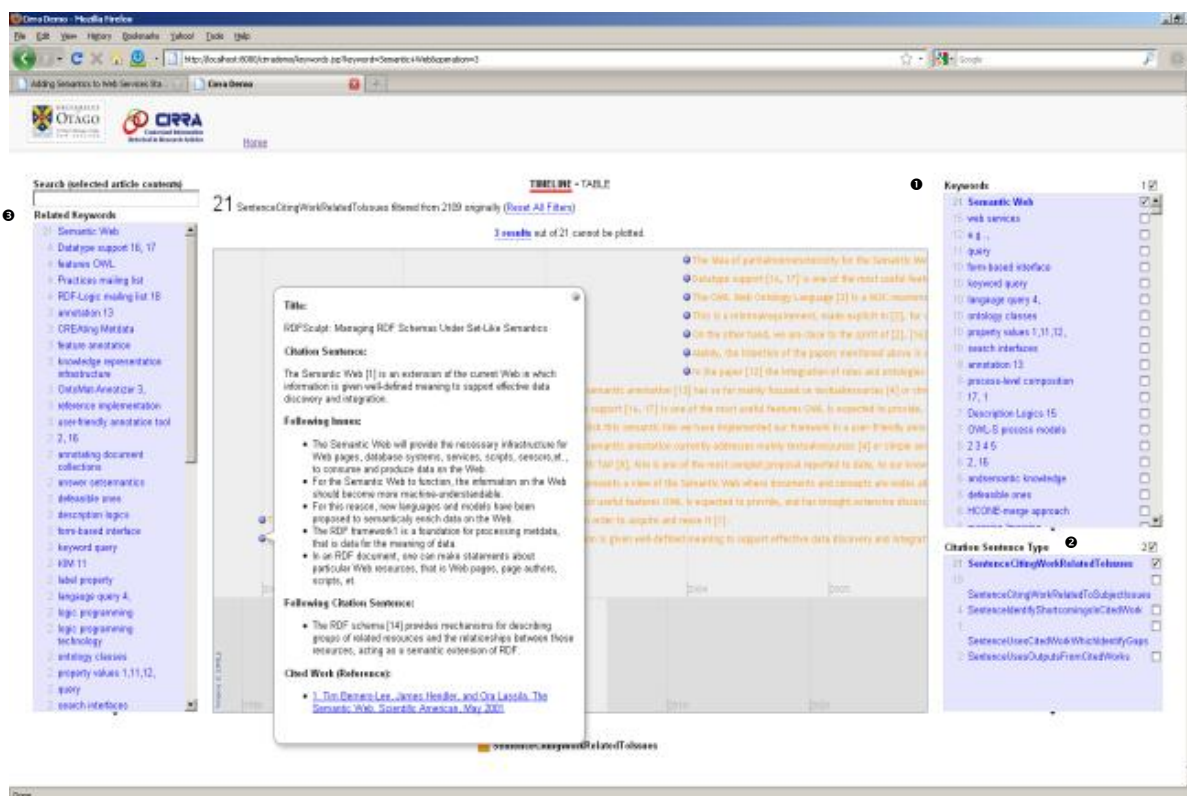


Figure18: Screenshot displaying citation sentences for a specific keyword on the keyword timeline

### Search citation sentences based on keywords

The application facilitates in searching citation sentences based on keyword search. In order to achieve this functionality, the application uses *topia.termextract*, a Python extraction library for extracting the keywords. *Topia.termextract* uses Parts-Of-Speech (POS) and simple statistical analysis for determining the terms and their strengths [42]. The first facet on the RHS with the caption ‘Keywords’ (indicated by the label ❶ in Figure 18) displays all keywords extracted from citation sentences. Users can select the required keyword in order to see re-

lated citation sentences for the selected keyword across the collection.

### Filter retrieved citation sentences by type

The application also facilitates in selecting a specific type of citation sentence for a given keyword. The second facet on the right hand side under the caption ‘Select Citation Sentence Type’ (indicated by the label ❷ in Figure 18) provides this functionality. If, for example, a user is interested in viewing all citation sentences that identify shortcomings in cited works for the keyword ‘Semantic Web’, he can

choose accordingly and can view only these kinds of citation sentences.

### Filter retrieved results with additional keywords

The application also facilitates in filtering the retrieved citation sentences by using related keywords. The facet on the left hand side under the caption ‘Related Keywords’ (indicated by the label **Ⓚ** in Figure 18) helps in achieving this functionality. The facet lists all keywords extracted from the retrieved citation sentences. Thus, the user can further refine

his search by selecting the required keyword from this facet.

### View Contexts of Citation Sentences

Besides listing all citation sentences for a given keyword, the application also facilitates in viewing the full context of each of these citation sentences. Figure 19 provides a screenshot where a citation sentence is selected in the keywords timeline.

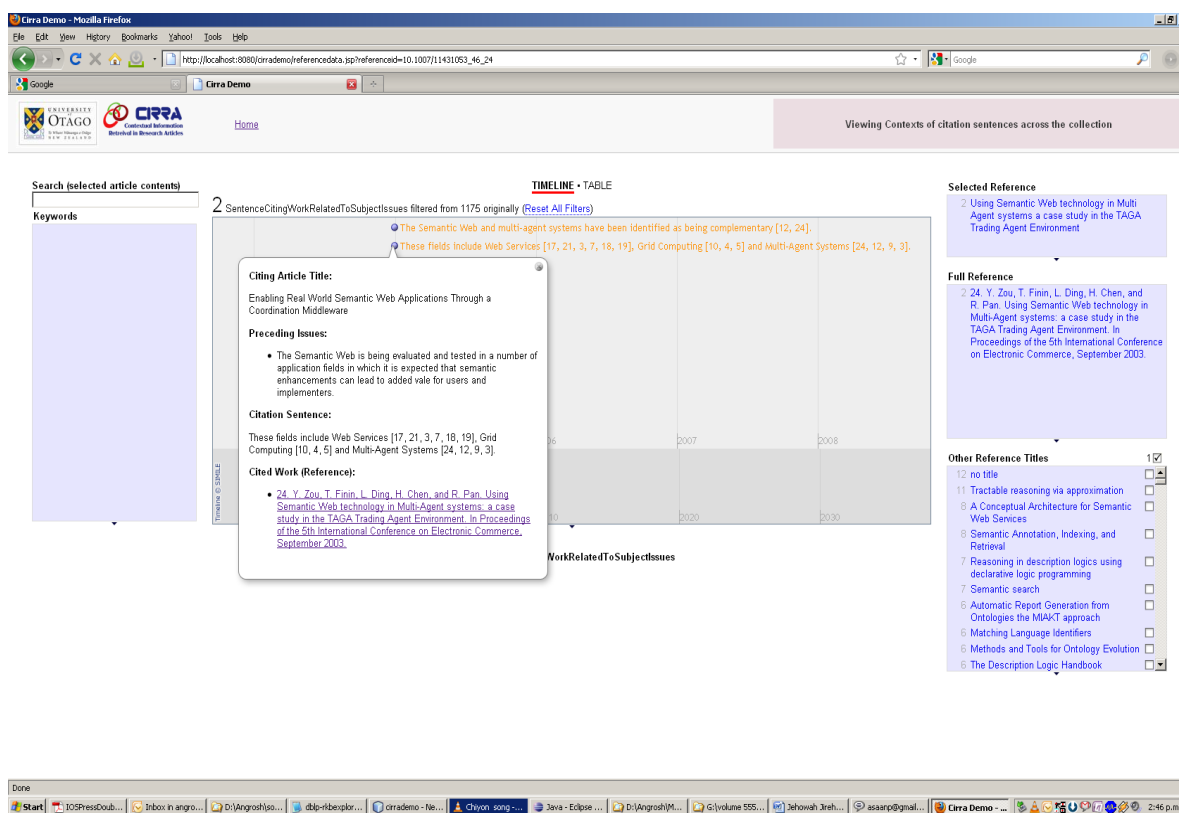


Figure 19: Viewing details of citation sentence in the keywords timeline

## 9. Discussion and Conclusion

We presented in this paper our work carried out for identifying contexts associated with sentences in research articles and using this information for providing value-added information services. The key focus of this paper has been to develop a linked data application using contextual information extracted

from papers published in the ESWC collection. In order to achieve this objective, the following steps were followed:

Step 1 – Deduced a conceptual framework for defining various contexts associated with citation and non-citation sentences in research articles (as described in Section 4).

Step 2 – Developed Sentence Context Ontology for modelling these contexts and derive machine-understandable data (as described in Section 5).

Step 3 – Carried out supervised learning experiments using conditional probabilistic models for achieving automatic classification (as described in Section 6).

Step 4 – Using these principles and techniques, we developed a linked data application, which uses contextual information extracted from papers published in the ESWC collection (as described in Section 7 and 8).

The linked data application facilitated in providing the following unique features and services for the research community:

#### *9.1. View Contexts of Citation Sentences*

The citation sentence timeline interface which provides for searching and browsing contexts of citation sentences in research articles (described in Section 8.1.1) provides a unique feature for the research community. The interface allows users to select different types of citation sentences and view their related sentences. This information helps in obtaining a better understanding of the use of cited works in the article and facilitates in adjudging the importance of cited works in a given paper. The navigation links provided to author and citations timeline helps the user in learning more about the authors and the cited works respectively.

#### *9.2. Understanding Works of Authors*

The author timeline interface described in Section 8.1.1 helps in viewing in a single view, different works of a given author. This includes both the published works as well as cited works. The timeline displays the contexts in which the author's work is cited and helps in understanding how an author's work is cited over a period.

#### *9.3. Understanding Contexts of Cited*

The citations timeline interface described in Section 8.1.1 allows users to view the contexts in which a given work is cited by different articles. This helps in understanding how a given work is cited over a period

#### *9.4. Search and Browse Cited Articles Data*

The interface for searching cited articles in ESWC described in Section 8.1.2 provides an important search feature for the research community by providing all details about how a given document is cited by other articles in a single view. This is in contrast to current search facilities, which simply provide the number of documents citing a given document.

The application, instead of just identifying the number of documents citing a given document, identifies the number of times a document is cited in each article and results in the total number of times a document is cited in different articles across the collection. It also identifies all contexts in which the cited work is used and displays them in a single view. This overcomes the time-consuming task of referring to each article to learn how an article is cited and helps in learning quickly the use of the cited work. The metrics provided about the contexts of cited works offer a new way of looking at the notion of citation analysis and indexing.

Thus, instead of just counting the number of times a given document is counted, the application also identifies the reasons and contexts in which the work is cited. This facilitates in a better understanding of the cited work and can lead to a better evaluation of the cited work.

#### *9.5. Sketch Intellectual Lineage for a given Idea*

The keywords timeline described in Section 8.1.3 facilitates in tracing the intellectual lineage for a given idea. This interface facilitates in viewing in a single view, the views of different authors and the use of different cited works for a specific keyword. Citation sentences are placed according to published year of the cited work and helps in understanding how different works are cited over a period for a given keyword. This facilitates in sketching the intellectual lineage for a given idea.

To sum up, this paper presented our research work carried out for identifying contexts associated with sentences in research articles and employs this information for developing intelligent information services. The paper presented the linked data application developed for the ESWC collection and explained different value-added information services offered by the application. Our future work involves

carrying out an inter-rater reliability study for establishing the choice of labels for sentences as earlier defined in Section 4. Further, we also intend to develop a larger training dataset using the ESWC data with a focus on achieving a higher accuracy, particularly for classes, which currently suffer from poor F-Scores. Finally, the overall objective would be to develop a robust linked data application for the research community based on the semantic publishing models presented in this paper.

## 10. References

- [1] Research4Life, Research Output in Developing Countries Reveals 194% Increase in Five Years, 2009., [http://www.research4life.org/Documents/Increase\\_in\\_developing\\_country\\_research\\_output.pdf](http://www.research4life.org/Documents/Increase_in_developing_country_research_output.pdf).
- [2] J. Gaillard, "The Characteristics of R & D in Developing Countries by, 2008". Measuring R & D in Developing Countries, the UNESCO Institute of Statistics (UIS), Final Paper, March 2008., [http://www.uis.unesco.org/template/pdf/S&T/Gaillard\\_final\\_report.pdf](http://www.uis.unesco.org/template/pdf/S&T/Gaillard_final_report.pdf)
- [3] S.J.B. Shum, V. Uren, and G. Li, "Modelling Naturalistic Argumentation in Research Literatures : Representation and Interaction Design Issues," *International Journal of Intelligent Systems*, vol. 22, 2007, pp. 17-47.
- [4] M.A. Angrosh, S. Cranefield, and N. Stanger, "Context Identification of Sentences in Related Work Sections using a Conditional Random Field : Towards Intelligent Digital Libraries," *Joint Conference on Digital Libraries*, June 21-25, 2010, Gold Coast, Australia, ACM Press, 2010, pp. 293-302.
- [5] M.A. Angrosh, S. Cranefield, and N. Stanger, "Ontology-based Modelling of Related Work Sections in Re-search Articles : Using CRFs for Developing Semantic Data based In-formation Retrieval Systems," *I-Semantics 2010*, September 1-3, 2010, Graz, Austria, ACM, 2010.
- [6] M.A. Angrosh, S. Cranefield, and N. Stanger, "Context-based Information Retrieval System for Research Articles : Using Machine Learning and Semantic Web Technologies for Modelling Scientific Discourse," Submitted to *International Journal of Semantic Web and Information Systems (Special Issue on Induction on Semantic Web)*, 2010.
- [7] M.A. Angrosh, S. Cranefield, and N. Stanger, "Modelling Argumentation in Research Papers: Towards Developing Intelligent Research Tools," Submitted to *International Conference on Digital Information Management*, 14-16 September, 2011, Melbourne, Australia, 2011.
- [8] E. Garfield, "Can Citation Indexing Be Automated?," *Statistical Association Methods for Mechanized Documentation*, L.B.H. Mary Elizabeth Stevens, Vincent E. Giuliano, ed., National Bureau of Standards Miscellaneous Publication, Washington, 1965, pp. 189-192.
- [9] M.J. Moravcsik and P. Murugesan, "Some Results on the Function and Quality of Citations," *Social Studies of Science*, vol. 5, Feb. 1975, pp. 86-92.
- [10] H. Nanba and M. Okumura, "Towards Multi-paper Summarization Using Reference Information," *Proceedings of IJCAI*, 1999, pp. 926-931.
- [11] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Association for Computational Linguistics, 2006, pp. 103-110.
- [12] H. Nanba, N. Kando, and M. Okumura, "Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation," *American Society for Information Science SIG Classification Research Workshop: Classification for User Support and Learning*, 2000, pp. 117-134.
- [13] M. Garzone and R.E. Mercer, "Towards an Automated Citation Classifier," *Canadian AI 2000*, H. Hamilton and Q. Yang, eds., Springer-Verlag Berlin Heidelberg, 2000, pp. 337-346.
- [14] S.B. Pham and A. Hoffmann, "A New Approach for Scientific Citation," *Artificial Intelligence 2003*, T.D. Gedeon and L.C.C. Fung, eds.,

Springer-Verlag Berlin Heidelberg, 2003, pp. 759-771.

[15] R.E. Mercer and C.D. Marco, "The Importance of Fine-Grained Cue Phrases in Scientific Citations," *Artificial Intelligence 2003*, LNAI 2671, Springer-Verlag Berlin Heidelberg, 2003, pp. 550-556.

[16] D. Kaplan, R. Iida, and T. Tokunaga, "Automatic Extraction of Citation Contexts for Research Paper Summarization: A Co-reference-chain based Approach," *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009*, Suntec, Singapore, 2009, pp. 88-95.

[17] A. Ritchie, S. Teufel, and S. Robertson, "Using Terms from Citations for IR: Some First Results," *ECIR 2008*, Springer-Verlag Berlin Heidelberg, 2008, pp. 211-221.

[18] M. Hoang Le, T.-bao Ho, and Y. Nakamori, "Detecting Citation Types Using Finite-State," *PAKDD 2006*, LNAI 3918, Springer-Verlag Berlin Heidelberg, 2006, pp. 265-274.

[19] F. Peng and a Mccallum, "Information extraction from research papers using conditional random fields☆," *Information Processing & Management*, vol. 42, Jul. 2006, pp. 963-979.

[20] K. Hirohata, N. Okazaki, S. Anania-dou, and M. Ishizuka, "Identifying Sections in Scientific Abstracts using Conditional Random Fields," *Proceedings of the Third International Joint Conference on Natural Language Processing, Association for Computational Linguistics*, 2008, pp. 381-388.

[21] L. French, S. Lane, L. Xu, and P. Pavlidis, "Automated recognition of brain region mentions in neuroscience literature.," *Frontiers in Neuroinformatics*, vol. 3, Jan. 2009, pp. 1-7.

[22] J. Zou, D. Le, and G.R. Thoma, "Locating and parsing bibliographic references in HTML medical articles.," *International journal on document analysis and recognition*, vol. 13, Jun. 2010, pp. 107-119.

[23] L. Gao, Z. Tang, and X. Lin, "CEB-BIP: A Parser of Bibliographic Information in Chinese Electronic Books," *Joint Conference on Digital Libraries*, 2009, pp. 73-76.

[24] P. Lopez, "Automatic Extraction and Resolution of Bibliographical References in Patent Documents," *IRFC 2010*, H. Cunningham, A. Hanbury, and S. Ruger, eds., Springer-Verlag Berlin Heidelberg, 2010, pp. 120-135.

[25] I.G. Councill, C.L. Giles, and M.-yen Kan, "ParsCit: An open-source CRF reference string parsing package," *Proceedings of LREC, European Language Resources Association*, 2008, pp. 661-667.

[26] Q. Zhang, Y.-G. Cao, and H. Yu, "Parsing citations in biomedical articles using conditional random fields.," *Computers in biology and medicine*, vol. 41, Apr. 2011, pp. 190-194.

[27] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology.," *Journal of biomedical informatics*, vol. 41, Oct. 2008, pp. 739-51.

[28] P. Ciccarese, "Swan Citations Ontology Specification," 2008., <http://swan.mindinformatics.org/spec/1.2/citation.s.html>.

[29] F. Giasson and B. Darcus, "The Bibliographic Ontology," 2011., <http://bibliontology.com/>.

[30] D. Shotton, "CiTO, the Citation Typing Ontology.," *Journal of biomedical semantics*, vol. 1 Suppl 1, Jan. 2010, p. 1-18.

[31] T. Groza, S. Handschuh, K. Moller, and S. Decker, "SALT - Semantically Annotated L TEX for Scientific Publications," *ESCW 07- Proceedings of the 4th European Conference on the Semantic Web: Research and Applications*, Springer-Verlag Berlin Heidelberg, 2007, pp. 518-532.

[32] H.D. White, "Citation Analysis and Discourse Analysis Revisited," *Applied Linguistics*, vol. 25, Mar. 2004, pp. 89-116.



- [33] Springer, “Springerlink.com.”, <http://www.springerlink.com/>.
- [34] M. Pistore, P. Roberti, and P. Traverso, “Process-Level Composition of Executable Web Services : ‘ On-the-fly ’ Versus ‘ Once-for-all ’ Composition,” European Semantic Web Conference, 2005.
- [35] “OWL Web Ontology Language Reference.”, <http://www.w3.org/TR/owl-ref/>.
- [36] “The Protégé Ontology Editor and Knowledge Acquisition System,” 2011., <http://protege.stanford.edu/>
- [37] A. McCallum, D. Freitag, and F. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation,” Proceedings of the International Conference on Machine Learning, 2000, pp. 591-598.
- [38] J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields : Probabilistic Models for Segmenting and Labelling Sequence Data,” Proceedings of International Conference on Machine Learning, 2001, pp. 282-289.
- [39] A.K. McCallum, “MALLET: A Machine Learning for Language Tool-kit,” 2002., <http://mallet.cs.umass.edu/>.
- [40] T. Berners-Lee, “Linked Data: Design Issues,” 2009., <http://www.w3.org/DesignIssues/LinkedData.html>.
- [41] C. Bizer, R. Cyganiak, and T. Heath, “How to publish Linked Data on the Web,” 2007., <http://www4.wiwiw.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.
- [42] “topia.termextract 1.1.0.”, <http://pypi.python.org/pypi/topia.termextract/>
- [43] NLTK, “Natural Language Toolkit.”, <http://www.nltk.org/>
- [44] S. Richter, “lxml - Processing XML and HTML with Python,” 2011., <http://lxml.de/>.
- [45] “D2R Server – Publishing Relational Databases on the Semantic Web,” 2010., <http://www4.wiwiw.fu-berlin.de/bizer/d2r-server/>
- [46] “Sewese: JSP/Servlet Environment to Build Semantic Web Applications.”, <http://www-sop.inria.fr/teams/edelweiss/wiki/wakka.php?wiki=Sewese>.
- [47] MIT, “Exhibit: Publishing Framework for Data-Rich Interactive Web Pages,” 2009., <http://www.simile-widgets.org/exhibit/>.
- [48] “Exhibit JSON.”, [http://simile.mit.edu/wiki/Exhibit/Creating,\\_Importing,\\_and\\_Managing\\_Data](http://simile.mit.edu/wiki/Exhibit/Creating,_Importing,_and_Managing_Data)
- [49] “Babel.”, <http://service.simile-widgets.org/babel/>.
- [50] MIT, “Timeline: Web Widget for Visualizing Temporal Data,” 2009., <http://www.simile-widgets.org/timeline/>.
- [51] “Jena Semantic Web Framework.”, <http://jena.sourceforge.net/>.