Journal of Universal Computer Science, vol. 18, no. 16 (2012), 2301-2318
submitted: 7/1/12, accepted: 25/8/12, appeared: 28/8/12 © J.UCS

Weaving Scholarly Legacy Data into Web of Data

Atif Latif
(ZBW - German National Library of Economics
Leibniz Information Center for Economics
Kiel, Germany
A.Latif@zbw.eu)

Muhammad Tanvir Afzal
(Centre for Distributed and Semantic Computing
Mohammad Ali Jinnah University
Islamabad, Pakistan
mafzal@jinnah.edu.pk)

Hermann Maurer
(Institute for Information Systems and Computer Media
Graz University of Technology
Graz, Austria
hmaurer@iicm.edu)

Abstract: The Linked Open Data project provides a new publishing paradigm for creating machine readable and structured data on the Web. Currently, the significant presence of data sets describing scholarly publications in the Linked Data cloud underpins the importance of Linked Data for the scientific community and for the open access movement. However, these semantically rich datasets need to be exploited and linked with real time applications. In the project we report on this. We have exploited numerous scholarly datasets and have created semantic links to papers in an online journal, particularly Journal of Universal Computer Science (J.UCS). The J. UCS plays an important part in the computer science publishing community and provides a number of innovative features and datasets to its web users. However, the legacy HTML format in which these features are made available makes it difficult for machines to understand and query. Keeping in mind the impressive benefits of the Linked Open Data project, this paper presents an approach to convert J.UCS legacy HTML data from its current form to machine understandable format (RDF). It also interlinks this data with other important Linked Data resources. The approach developed has successfully disambiguated and interlinked J.UCS authors and publications datasets with DBpedia, DBLP, CiteULike and faceted DBLP. Additionally, triplified and interlinked datasets are made available to the scientific and semantic web community for downloading and posing SPARQL queries. This semantically linked dataset can further be used by researchers and semantic agents to identify semantic associations, to build inferencing systems, and to extract useful knowledge.

Keywords: Linked Data Publishing, Linked Data, Semantic Web, Digital Journals
Categories: M.0, H.3.3, L.1.4

1 Introduction

The Linked Open Data project [Bizer et al., 2009] [Berners-Lee 2006] provides a new publishing paradigm for creating machine readable structured data on the Web. It also offers best practices for interlinking of relevant datasets which are otherwise isolated (or even guarede) by use of typed linking mechanism. The existence of various domain oriented datasets in the Linked Data cloud shows its success and acceptance in diverse communities. At present, structured datasets from the domains of government, geo-locations, medicine and social communities etc. are available in the Linked Data cloud. There are a number of big companies that have also explored these datasets to make interesting applications [Hausenblas 2009] such as: BBC[1], Freebase[2], MusicBrainz[3]. Additionally, the significant presence of data sets describing scientific publications (such as: DBLP, IEEE, ACM and Citeseer etc.) in the Linked Data cloud also underpins the importance of Linked Data in the scientific and scholarly community.

Recently, with the advancement in science of the Open Access Movement [Roberts et al 2001], the need and demand of open electronic publishing has increased rapidly. Different approaches have been employed to make this huge repository accessible to all scientific community [Marchionini and Maurer 1995]. This caused the emergence of many open access journals. The resources provided by open access journals are usually presented to users in the legacy HTML format. However, in context of Semantic Web, such data remained hard to locate since it was unlinked due to its unstructured nature. In fact, in the legacy HTML data presentation paradigm, more emphasis is put on metadata generation at the document level. However, the important concepts present in the form of data remain unfocused. The extraction of data contexts and their description is difficult due to un-typed linking within and outside the dataset. For example, to spot scientific papers published by an author in different venues indexed by various digital libraries is difficult: one has to rely on heuristics only. Furthermore, the integration and disambiguation is also difficult due to the inability of making semantic information available in HTML formats.

Linked Data provides a solution to these problems by not only focusing on metadata at document level but also treating conceptual data items in the document as important elements. Every concept is assigned a unique identity (URI) and the typed relationship (defined in Linked Data) makes the data more explicit. Data concepts are supported by hopefull enough semantic information. Moreover, the semantic concepts can be exploited to disambiguate and integrate the data automatically. Additionally, data is structured in an RDF graph model. Thus, a variety of complex graph queries becomes possible.

Taking in account the important features presented by Linked Data, we have RDFized the legacy HTML data of a digital journal. The journal selected for this study is the Journal of Universal Computer Science [J. UCS 2011]. The journal provides a number of innovative features and many important datasets such as: “Links

[1] http://www.bbc.co.uk/music

[2] http://www.freebase.com/

[3] http://musicbrainz.org/

into Future" [Krottmaier, 2003] [Afzal, 2009], "Expertise Recommendations" [Afzal et al., 2008], "Tags Recommendations" [Afzal, 2010], and "Author's Profiles Recommendations" [Latif et al., 2010a]. It is our vision that RDFizing these datasets may lead to semantic discoveries for the scientific community for knowledge creation and interlinking.

For the RDFization of J.UCS datasets we have employed an innovative modeling and interlinking strategy. To model the typed relationship within and outside datasets we used the well-known metad-data description standards such as: Dublin Core and SIOC (Semantically-Interlinked Online Communities) [Breslin et al., 2005]. For the interlinking process, we selected DBpedia [Auer et al., 2007] and DBLP [Michael, 2009]. In DBpedia we search for the author's URI and verify it using a set of heuristics. In DBLP we search for other publications of the author in different venues. After the RDFization process, the J.UCS dataset has been interlinked within the Linked Data cloud. This data has also been made available as RDF graph. For human use a web interface for navigation through J.UCS RDFized resources is provided. For machine use and querying purpose, a SPARQL endpoint is set up and made available.

The remainder of this paper is structured as follows: section 2 discusses the state of the art of the RDFization process. Section 3 elaborates the datasets which are used in this process of J.UCS RDFization. Section 4, 5, 6 explain the RDFization process and presents the design of the study for interlinking the publications in the Linked Data cloud. Section 7 illustrates the interface and sample SPARQL queries. Conclusions and potential future efforts are discussed in section 8.

2 Related Work

The publishing of data as Linked Data usually comprises two steps : (1) conversion of raw data into RDF and (2) interlinking of the RDFizied data with external Linked Data resources. Usually, the raw data is represented in legacy structured elements such as tables or spreadsheets. They give more flexibility and control over data handling and conversion when compared to mere textual representations. Currently, to facilitate data conversion and interlinking, many tools are available in the Linked Data community. However, these tools always require a human publisher who first selects the important data fields within the raw data and then provides a conceptual understanding of data in the form of ontology. Based on human input these tools convert the provided raw data and serve it over the web in an RDF graph as structured data. Afterwards, the interlinking tools connect the material with related structured external datasets. In a nutshell, these tools shield data publishers from much of the technical aspects and publish data according to the Linked Data principles. Some of the popular data conversion and interlinking tools are discussed in what follows.

2.1 Data Conversion Tools:

Data conversion tools usually come in to two categories. Category maps convert partial data aspects from the HTML webpage into structured data or they convert relation database content into Linked Data.

2.1.1 Triplify

Triplify [Auer et al., 2009] is a small plug-in, specially designed to create Linked Data from web applications. It works by exploiting SQL queries which are used to display important information of a web page. Triplify works with a configuration file containing important SQL queries, field names, and their equivalent semantic properties. When embedded on a web server, this configuration file starts to publish website data for machine processing in RDF and JSON format. Triplify is specially targeted towards mainstream Web applications such as CMS, Wikis, Blogs. It is best suited in situations where only partial data of webpage is to be served as structured data for semantic crawlers and agents. Overall it is easy to integrate it with various web applications but does not provide a SPARQL endpoint for querying.

2.1.2 SparqPlug

SparqPlug [Coetzee et al., 2008] also known as RDFizing service converts the legacy HTML documents into RDF. It works on the serialization of XHTML Document Object Model (DOM) and SPARQL query model for data conversion. Furthermore, it facilitates users for issuing SPARQL queries to get desired HTML content as RDF graph. However, its usage comes with complex routines and requires a fair degree of domain specific knowledge with SPARQL querying skills to produce RDF data. In a nutshell, SparqPlug approach is typically targeted towards more sophisticated Semantic Web practitioners.

2.1.3 D2R Server

The D2R server [Bizer and Cyganiak 2006] is a tool for publishing relational database as Linked data on the web. It presents a live view of Linked Data over the relational database and maintains data up-to-date. The D2R server provides a customizable declarative mapping file in which mapping of the relational table with targeted RDF vocabularies is defined. Based on this declarative mapping file, data publishers can create a view of Linked Data over relational database and they can create data dumps in RDF/XML or N3 format. The D2R server also provides an interface for Linked Data navigation and a description of every individual resource via the HTTP protocol. In addition, it provides a SPARQL endpoint which enables other applications to search and query the database using the SPARQL query language. The D2R server seems to be the best alternative in many cases, including ours, due to its good performance, scalability and the availability of SPARQL endpoint features.

2.2 Data interlinking Tools:

2.2.1 RDF-AI:

RDF-AI [Scharffe et al., 2009] is an interlinking tool for an already RDFized dataset. It offers modules for the matching and data fusions of RDF datasets depending on the provided targeted parameters of the external dataset. It uses string matching and taxonomical similarity measures for interlinking. It connects two datasets using the owl:sameAs relationship. It supports various outputs formats for user provided specifications i.e. RDF/XML, N3, JSON etc. The usage of this tool requires a basic

knowledge of Linked Data structures and the live availability of a Sparql endpoint for setting up RDF-AI parameters. This tool is best suited in situations where simple and straight matching of triples is required.

2.2.2 SILK Server

One of the popular similarity based interlinking tool is the SILK framework [Volz et al., 2009]. It provides a set of services to discover relationships of resources within different Linked datasets. It works with a "Link Specification Language" where data publisher can specify the conditions and type of RDF links needed to be present in Linked dataset. The SILK framework works on data sources that are interlinked and presented with the SPARQL specification. It is necessary to have an adequate knowledge about semantic structures and SPARQL querying to understand the underlying structured of targeted dataset and to assuring live availability of its SPARQL endpoint. The SILK framework provides many similarity based measures but fails in situations where in depth heuristics are needed for disambiguation purposes e.g. persons having the same name in the dataset and with not much semantic information available for them.

Keeping in mind the nature of the scholarly dataset to be used in this study we need a tool which can convert legacy relational database to RDF in real time and then serve this data over an SPARQL endpoint for querying purposes. Additionally, it must provide a flexibility to integrate our newly proposed interlinking technique results. In comparison, D2R Server nearly fulfills our needs concerning data conversion whereas the deep analysis of above mentioned interlinking tools gives us a guideline to create our own automatic quality interlinking technique.

3 Datasets

In this study, an open access digital journal is selected for the RDFization process: The Journal of Universal Computer Science (J. UCS). There are two reasons to select J.UCS: (1) its popularity and coverage, (2) the availability of important datasets. J.UCS is a high quality electronic journal that deals with all aspects of Computer Science [Calude et al., 1994] and is thus one of the oldest electronic journals. J.UCS has been appearing at least monthly since 1995 with uninterrupted publications. The statistics of hits, page views and download of papers from J.UCS website further validates that the Journal is a popular and trusted medium in dissemination of quality scientific information. The other reason to select J. UCS is the availability of unique features. The journal provides a number of innovative and important features to its web users including "Links into Future" [Afzal, 2009], "Expertise Recommendations" [Afzal et al., 2008], "Tags Recommendations" [Afzal, 2010], and "Author's Profiles Recommendations" [Latif et al., 2010a]. At present all of these features are available in legacy HTML format as illustrated in Figure 1. However, this information as is cannot be used by the semantic community and by semantic agents to discover and infer further useful knowledge. Therefore these important datasets are exploited and linked with relevant semantic datasets.

For simplicity, we have categorized the important features and datasets of J.UCS into two broad categories such as: (1) features developed over internal datasets and

(2) features developed over external datasets. These two categories are discussed in detail in the following sections and explained in Figure 1. A screen similar to Figure 1 may be obtained by clicking on the button "Links into Future" while viewing the paper at http://www.jucs.org/jucs_11_6/fine_grained_transclusions_of.

3.1 Features Built on Internal Datasets

In this section details of features built around the internal datasets of J.UCS are discussed. More specifically, these features use the basic artifacts of digital journal like papers and authors as input datasets for the calculation and availability of new insights and results. The data of all published papers in J. UCS is maintained on the J.UCS server in volumes and issues. The metadata information of the papers such as: title, keywords, volume number, issue number, submission date, acceptance data, published date and categories is maintained in various data tables and are used as entry point for these features. Similarly, information about the paper's authors is also maintained separately on the J.UCS server. This information consists mainly of the author's name, email address, affiliation, together with the number of published papers in this journal and a record of who was serving as reviewer in this journal. The two features provided on these internal datasets are "Links into the Future" and "Expertise Recommendations".

3.1.1 Links into the Future

Links into the Future is an idea proposed by [Maurer, 2001], partially realized by [Krottmaier, 2003], and extended and implemented by [Afzal, 2009] to work on the complete archive of papers and authors of J.UCS. For each paper, the feature recommends semantically related papers published at future dates as compared to the date of the paper at hand. More details can be found in the relevant papers [Afzal, 2009] [Afzal et al 2009a]. An illustration of calculated "Links into the Future" is also shown in Figure 1.

3.1.2 Expertise Recommendations

The other feature built from internal datasets of J.UCS is "Expertise Recommendations". The "Expertise Recommendations" feature is developed to discover potential experts related to the topics of the paper at hand [Afzal et al., 2008]. This feature, on the basis of ACM topics of currently viewed paper, calculates and presents the potential experts within J. UCS. An illustration of "Expertise Recommendations" is also shown in Figure 1.

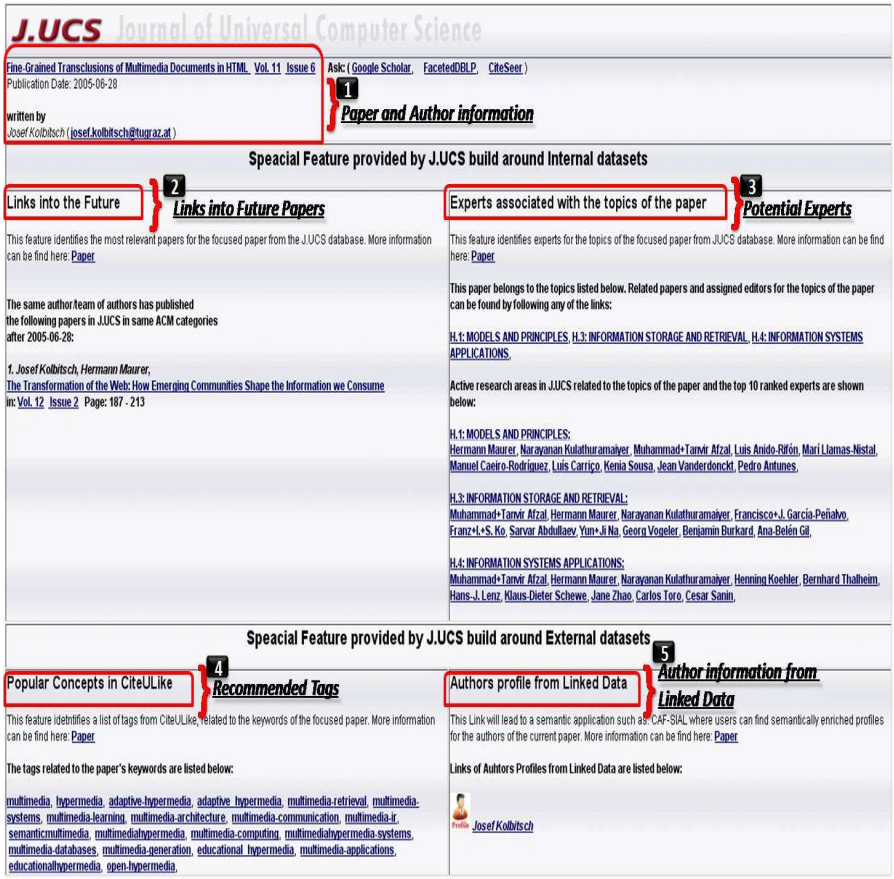

Figure 1: J.UCS offered Features

3.2 Features Built on External Datasets

The features mentioned in the previous section were built using J. UCS internal data. However, J.UCS also provides some features to its users which are built around external datasets. These features explore external datasets for the availability of related data to the journal. Currently provided features by J.UCS on external datasets are "Tags Recommendations" and "Authors' Profiles Recommendations". A summary of each feature is presented in the following sections.

3.2.1 Tags Recommendation

This feature exploits authors' keywords of scientific publications to link these resources with tags and papers available in CiteULike. CiteULike is a social bookmarking and tagging application which provides useful annotations to scientific papers. The paper's keywords are compared with the corresponding tags in CiteULike. Subsequently, a user can follow the link to find all papers in CiteULike

annotated with that particular tag. An illustration of the "Tag Recommendations" is shown in Figure 1.

3.2.2 Authors' profiles Recommendations

Another feature provided by J.UCS is the linking of journal's authors with their profiles extracted from Linked Data. This feature is solely built around external datasets of Linked Data such as: DBpeda, DBLP and Linked Data application named as CAF-SIAL [Latif et al., 2010]. An illustration is shown in Figure 1. Following is a detailed description of external datasets which are used for the development of this feature.

3.2.2.1 DBpedia

DBpedia, a semantic version of Wikipedia, is one of the biggest examples of Social Semantic Webs. DBpedia is considered one of the most promising knowledge bases, having a complete ontology along with Yago classification [Suchanek et al., 2007]. It currently describes more than 3.64 million things, including at least 416,000 persons, 526,000 places etc. [Auer et al., 2007]. The openly available RDF dumps make DBpedia an interesting subject of study. There has been valuable work done on studying the reliability of Wikipedia URI's [Hepp et al., 2007] that is a backbone of the DBpedia. This study suggests that the meaning of a URI stays stable approximately 93% of the time. Its heavy interlinking within the LOD cloud makes it a perfect resource to search URIs. For current experiments, we concentrated on the part of DBpedia that encompasses data about persons.

Two RDF dumps about personal information (*Persondata* and *Links to DBLP*) were selected to find relevant information of J.UCS authors. These datasets are freely available in RDF dumps[4]. The above mentioned datasets are exploited to discover authors' profiles. J. UCS authors are linked with profiles discovered this way.

3.2.2.2 Links to DBLP

Links between computer scientists in DBpedia and their contributions in the DBLP database are enlisted in a *same:as* relationship in this dataset. To follow the DBLP links, The D2R Server, a *semantified* version of DBLP bibliography was accessed from Berlin and Hanover SPARQL endpoints[5]. This D2R Server is based on the XML dump of the DBLP database. The DBLP database provides bibliographic information of all major computer science journals and conference proceedings. The database contains more than 800.000 articles and 400.000 authors [Michael, 2009]. This dataset was used to find contributions of J. UCS authors published outside J. UCS.

All of these discussed internal and external features and datasets are very important parts of J.UCS. The datasets may play a pivotal role in building semantic-aware and inference-based systems. Furthermore, the semantic associations may be discovered to find interesting insights. Therefore, the dataset need to be RDFzied for

[4] http://wiki.dbpedia.org/Downloads34

[5] http://dblp.l3s.de/d2r/snorql/

machine processing and to allow sharing information with the scientific community for knowledge discovery and new knowledge creation.

4 Weaving Approach to Web of Data

For the weaving of J.UCS datasets into the Web of Data, two approaches are used: (1) modeling for RDFization and (2) interlinking.

4.1 Modeling for RDFization

A first step towards RDFization of J.UCS Data is to look around for available ontologies for modeling the characteristics of J.UCS datasets i.e. persons and papers. By sticking to Linked Data community norms, we selected FOAF[6] [Brickley and Miller, 2004], SIOC[7], Dublin Core[8], SWC[9], SWRC[10], VCard[11] and SKOS[12] ontologies for this process. To model person related properties, we used FOAF and SIOC ontologies while Dublin Core, ISWC, VCard and SKOS ontologies were used to define publication related properties. In Figure 2, the overview of the modeled properties is presented for papers and persons. Paper- and person classes are connected to each other by paper-authorship properties.

The next step in this process is to use a tool for the conversion of this legacy data into RDF. A D2R server is preferred in this study due to its performance, scalability and the availability of SPARQL endpoint feature. D2R Server uses a customizable D2RQ mapping file which further allows the translation of raw data into RDF data. This mapping file was created manually on top of J.UCS data tables very carefully on the basis of properties modeled in our first step. In writing the mapping file, a rich description of properties was emphasized to maintain interoperability between ontologies and to enhance knowledge discovery. For example, we used various indicators like swc:Paper, dcmi:Text, sioc:Item, swrc:Article, foaf:Document to describe the paper. A similar approach was used to describe the publication of an author. A D2R mapping file was written for J.UCS dataset conversion. In the end, a D2R server mapped application was deployed as a servlet which allows the RDF data to be browsed, searched and queried in real time.

4.2 Interlinking

For the interlinking of J.UCS data within the Linked Data cloud, a set of services was developed. These services search for the relevant information from Linked Data. This approach uses SINDICE API [Oren et al., 2008] to query for the intended resource

[6] http://xmlns.com/foaf/0.1/

[7] http://rdfs.org/sioc/ns#

[8] http://purl.org/dc/elements/1.1/

[9] http://data.semanticweb.org/ns/swc/ontology#

[10] http://swrc.ontoware.org/ontology#

[11] http://www.w3.org/2001/vcard-rdf/3.0#

[12] http://www.w3.org/2004/02/skos/core#

URI's. Next, it filters DBLP and DBpedia URI's out of the returned results. Subsequently, a validation service was developed to verify the collected URIs.

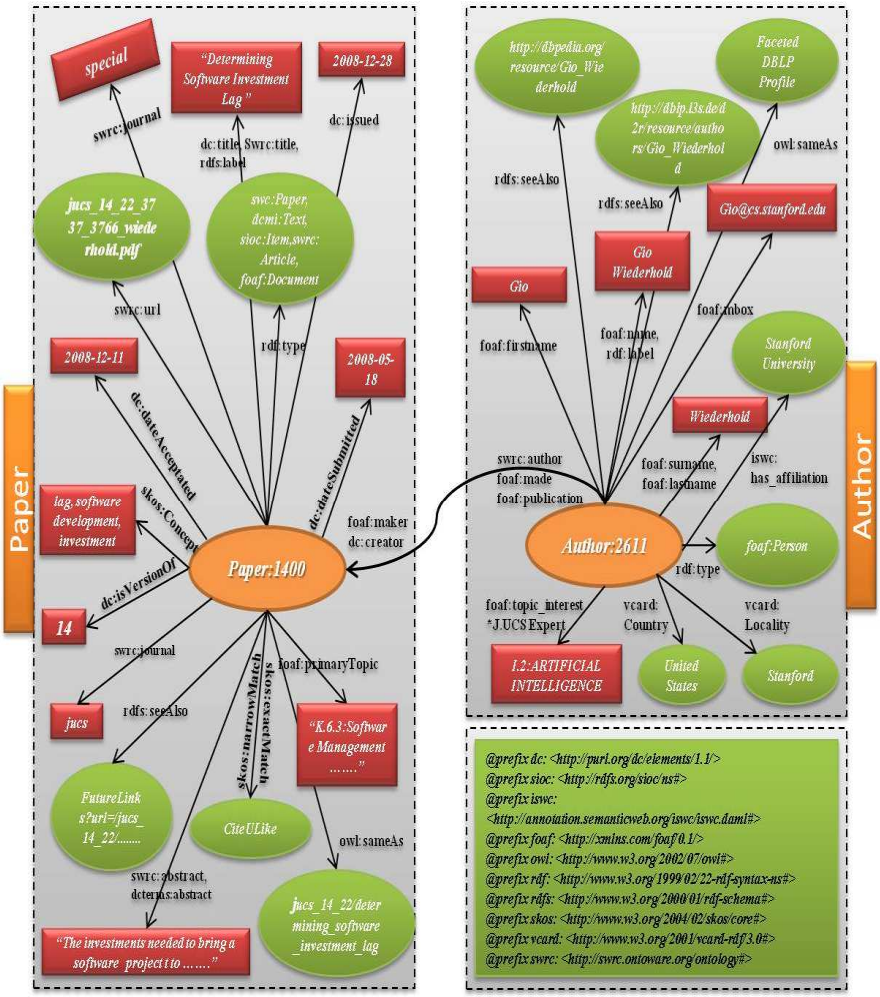

Figure 2: Modeling of Properties

This validation service works with a set of heuristics. Based on the output of verification module, discovered URI's were linked with owl:sameAS property. We also have included CiteULike tag and Faceted DBLP links respectively to bring in the conventional web flavour. An overview of the weaving approach in steps is illustrated in Figure 3.

The detail of the process is presented in the following sections.

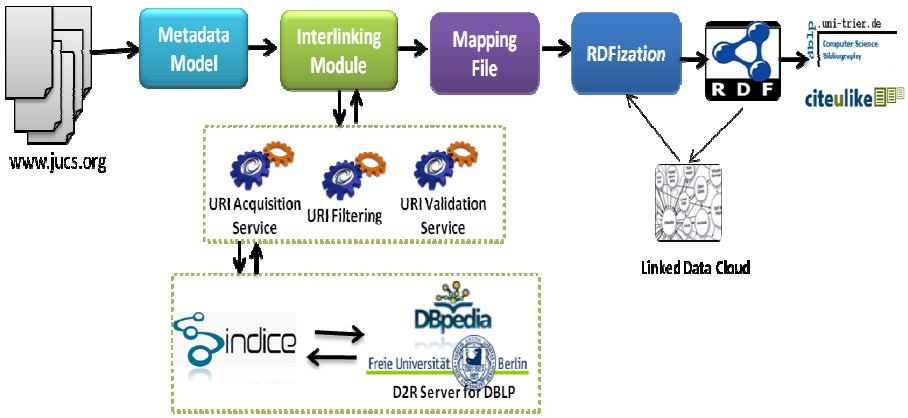

Figure 3: Framework for RDFization

4.2.1 URI Acquisition

A layered approach was employed to acquire the intended URI from local triple store and from Linked Open Data cloud developed previously [Latif et al., 2009c]. Four processes were involved in this approach. The output of each process resulted in an input for the subsequent layer. This ended up with the retrieval of the desired resources. The details of each process in presented below:

- J.UCS dataset Pre-Processing
- Direct matching of J.UCS authors with DBpedia Person-data dataset
- Direct matching of J.UCS authors with Links to DBLP dataset
- Querying and Filtering of URI from Sindice
- URI Validation

4.2.2 J.UCS dataset Pre-Processing

Sometimes, the authors' names contain “Umlaute” and other character anomalies (irregularity in names not common in English) which need to be processed before matching. An automated script was written to remove such inconsistencies. Subsequently first, middle, and last names were concatenated to construct a full name for the matching process.

4.2.3 Direct Matching of J.UCS Authors with DBpedia Person-data dataset

A complete J.UCS author name, acquired from the previous step, was considered for matching in the *DBpedia Persondata triple store*. However, this operation gave only a low success rate.

4.2.4 Direct Matching of J.UCS Authors with Links to DBLP dataset

In this process, authors were matched with DBLP local triple store. The result of this matching was also not satisfactory.

To improve the discovery of URIs, different online semantic search engines were analyzed like Falcon [Cheng et al., 2008], Swoogle [Ding et al., 2004] and SINDICE. Based on the up-to-date, large indexing corpus, and easy API access, SINDICE was selected for further matching of URIs.

4.2.5 Querying and Filtering of URI from SINDICE

A web service was written to call the API of SINDICE with the formulated query. The web service was called recursively for every J.UCS author not found. In response, SINDICE provided the list of the URI's which were further filtered on the basis of DBpedia provenance. Furthermore, the direct matching of author's full name was performed with the DBpedia to select the exact and trusted URI.

4.2.6 URI Validation

Our past research of CAF-SIAL system [Latif et al., 2009c] helped us in developing a set of heuristics to validate the acquired resource URI. The validation and disambiguation of URI is an important part of this application. By manual de-referencing and inspection of the acquired URI's, we discovered some inconsistencies:

- URI of the respective author exists (wrongly indexed by SINDICE) but with no information making it useless.
- Many ambiguous URI's which matched with exact name of the intended J.UCS author leading to wrong persons.

We discovered that the inconsistencies and wrong results were not the fault of our algorithm; rather, semantic indexers introduced this. To further disambiguate authors, a set of heuristics was written. After inspection, it was noted that there were certain kind of properties for a person type which would be exploited to disambiguate individuals. These properties are dbpedia:Abstract / dbpedia:Comment and SKOS categories. For example, SKOS categories and keywords, being used to represent the persons belonging to education profession are: "computer science, computer scientist, professor, informatics, researcher ....." etc. All of these important properties represent a person belonging to the scientific community. Thus the persons having same names and belonging to different professions can easily be filtered out. An automated script was written to check the keywords in the abstract property and SKOS categories of the URI. This resulted in correcting the problems introduced by semantic indexers. In a nutshell, our proposed interlinking approach provides an extra edge over other interlinking tools by first searching and then filtering data with specifically created disambiguation heuristics. This interlinking approach helped us to interlink with quality external person's data from DBpedia.

5 Design of URIs

URI is required in order to identify entities, types, and their relationship. According to Linked Data principles, URI's of the resources must be persistent and unique. The URI's used for the J.UCS dataset followed the pattern of:

http://www.jucs.org:8181/d2rq/(resource or page)/(type)/(id or title)

Where "*(type)"* represents the type of resource such as: authors or papers. The "*(id or title)"* corresponds to the unique identification of resource maintained in the dataset. The "*(resource or page)"* is subjected to a content negotiation process i.e. in HTML description (if supported by Web-Browser), the "*(page)" is used.* Whereas the "*(resource)*" literal is used for direct RDF description / dereferencing. We used the name space of the J.UCS website as the part of URI to preserve information about the origin and owner of this datasets. For example, the following URI: *http://jucs.org:8181/d2rq/resource/Authors/1396* identifies author "*Klaus Tochtermann"* while the URL: *http://www.jucs.org:8181/d2rq/resource/Papers/788* describes the J. UCS paper with title "*The Dortmund Family of Hypermedia Models - Concepts and their Application*".

6 Data Access

The conversion of legacy HTML data into semantically-typed RDF graphs was achieved as described in the previous sections. In this section, we demonstrate the usefulness of this semantically rich dataset for the scientific community. It allows information to be visualized as a single information space like a graph. This graph can be queried and de-referenced to get further insights using a variety of ideas such as: semantic associations, semantic-aware systems, inference-based systems. By following the design principles of Linked Data, it was possible to provide URIs, which are deference-able. This will allow the discovery of additional meaningful information.

The semantic representation of an author "Pascal Costanza" and a paper titled "Software Is More Than Code" is presented in Figure 4 and Figure 5, respectively, in a semantic web server environment. Both of these figures illustrate the semantic properties, which are further dereference-able and are interlinked with other external resources by sameAs and seeAlso properties. A SPARQL endpoint of RDFized J.UCS data is provided for querying at this link: http://jucs.org:8181/d2rq/snorql/.

Pascal Costanza

Resource URI: http://jucs.org:8181/d2rq/resource/Authors/2911

Home | All Authors

Property	Value
vcard:Country	<http://dbpedia.org/resource/Belgium>
vcard:Locality	<http://dbpedia.org/resource/Brussels>
swrc:author	<http://jucs.org:8181/d2rq/resource/Papers/1548>
swrc:author	<http://jucs.org:8181/d2rq/resource/Papers/1550>
is dc:creator of	<http://jucs.org:8181/d2rq/resource/Papers/1548>
is dc:creator of	<http://jucs.org:8181/d2rq/resource/Papers/1550>
foaf:firstname	Pascal
iswc:has_affiliation	Vrije Universiteit Brussel
rdfs:label	Pascal Costanza
is foaf:maker of	<http://jucs.org:8181/d2rq/resource/Papers/1548>
is foaf:maker of	<http://jucs.org:8181/d2rq/resource/Papers/1550>
foaf:mbox	<mailto:pascal.costanza@vub.ac.be>
foaf:name	Pascal Costanza
rdfs:seeAlso	<http://dblp.l3s.de/?q=Pascal+Costanza&search_opt=authorsOnly&newQuery=yes&resTableName=query_resultW9SkYi&resultsPerPage=100&synt_query_exp=none>
rdfs:seeAlso	<http://dbpedia.org/resource/Pascal_Costanza>
rdfs:seeAlso	<http://www4.wiwiss.fu-berlin.de/dblp/resource/person/272714>
foaf:surname	Costanza
foaf:topic_interest	H.3.3:Information Search and Retrieval
rdf:type	foaf:person

Figure 4: Semantic Representation of an Author

Software Is More Than Code

Resource URI: http://jucs.org:8181/d2rq/resource/Papers/1067

Home | All Papers

Property	Value
skos:Concept	formal methods,software engineering
dc:SizeOrDuration	602-606
dcterms:abstract	This paper reviews the current practice of software engineering and outlines someprospects for developing a more holistic and formally grounded approach.
is swrc:author of	<http://jucs.org:8181/d2rq/resource/Authors/1843>
dc:creator	<http://jucs.org:8181/d2rq/resource/Authors/1843>
dc:dateAcceptated	2007-05-25 (xsd:date)
dc:dateSubmitted	2007-05-07 (xsd:date)
dc:identifier	1067 (xsd:integer)
dc:isPartOf	5 (xsd:integer)
dc:isVersionOf	special
dc:issued	2007-05-28 (xsd:date)
swrc:journal	jucs
rdfs:label	Software Is More Than Code
foaf:maker	<http://jucs.org:8181/d2rq/resource/Authors/1843>
foaf:primaryTopic	D.2.4:Software/Program Verification
foaf:primaryTopic	F.3.1:Specifying and Verifying and Reasoning about Programs
owl:sameAs	<http://www.jucs.org/jucs_13_5/software_is_more_than>
owl:sameAs	<http://www.jucs.org/jucs_13_5/software_is_more_than/jucs_13_5_0593_0601_rajamani.pdf>
rdfs:seeAlso	<http://www.jucs.org:8181/mashup/servlet/FutureLinks?url=/jucs_13_5/software_is_more_than>
dc:title	Software Is More Than Code
rdf:type	swc:Paper
rdf:type	dcmi:Text
rdf:type	sioc:Item
rdf:type	swrc:Article
rdf:type	foaf:Document
swrc:volume	13 (xsd:integer)

Figure 5: Semantic Representation of a Paper

7 Query Example

As mentioned earlier the conversion of J.UCS Data into a single information space gives us the leverage to ask complex question in the form of SPARQL queries. For example, we want to find an author who fulfills all of the following conditions: 1) the author has published a paper in J.UCS after the year 2005, 2) the author is a ranked expert in the area of "Information Search and Retrieval", and 3) the author has a DBpedia page. However, to perform such kind of query over unstructured legacy HTML representation is impossible and unsupported. For instance, a searcher has to first divide the tasks into sub steps and further has to perform various filtering mechanism, requiring lots of manual effort to get the results. This is where the importance and utility of structured data comes in, making it possible to issue such a complex queries and obtaining the result in no time. By issuing this query over SPARQL endpoint, we discovered that "Pascal Costanza" affiliated with "Vrije Universiteit Brussel" fulfilled our query criteria. The formalization of this query is given in Listing1.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix sioc: <http://rdfs.org/sioc/ns#>.
@prefix iswc: <http://annotation.semanticweb.org/iswc/iswc.
    daml#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

SELECT * WHERE
{
?Publication sioc:title ?Title;
dc:issued ?PublicationDate.
?Author foaf:maker ?Publication;
foaf:topic_interest "H.3.3:Information Search and Retrieval";
foaf:name ?Name;
foaf:mbox ?Email;
iswc:has_affiliation ?Affiliation;
rdfs:seeAlso ?ExternalLinks.

FILTER(?PublicationDate > "2005-12-31"^^xsd:date).
FILTER regex(?ExternalLinks, "http://dbpedia.org" )
}
```

Listing1. Sample SPARQL Query

The RDFization of J.UCS now provides journal users and administration with plenty of options to analyze data in different contexts with SPARQL queries. For example, a journal user may want to locate an author who has maximum publications in JUCS for his expertise to seek guidance. In another case, J.UCS administration may want to know an ACM category of J.UCS which has a paper that has got

maximum tags in CiteuLike social bookmarking network. In addition to these, varieties of other complex queries are also possible now. For example:

- Finding an author who has the biggest co-author network within J.UCS dataset.
- Searching for an author who is currently an editor and has contributed to J.UCS with more than five publications,
- Finding the name of authors who have publications in J.UCS special issues.
- Locating a J.UCS category which has received minimum amount of tags in CiteULike.
- Finding the authors who have single-authored publications in the topic "Computer Graphics" in the year 2010.

8 Conclusion and Future Work

In this paper, we presented an approach to convert J.UCS legacy HTML data into machine readable format. Furthermore, we discussed the interlinking of authors and papers with the following datasets such as: DBpedia, DBLP, Faceted DBLP and CiteULike datasets. Subsequently, we highlighted the benefits of J. UCS structured data in retrieving meaningful insights from complex queries. This paper makes the following contributions.

1. It converted the legacy HTML journal Data into an RDF graph. Furthermore, the data has been made available to the Linked Data cloud for the scientific community to reuse and interlink.
2. It interlinked journal Data with semantic datasets (such as: DBpedia, DBLP, and Faceted DBLP) and conventional Web resources (such as: CiteULike).
3. It developed, implemented, and evaluated a set of heuristics to acquire intended URI's for a resource.
4. Author's disambiguation was achieved using a set of heuristics at different layers.
5. It provided an HTML interface on top of converted semantic data for navigation between interlinked connected resources.

The generated semantic data can be accessed at http://www.jucs.org:8181/d2rq/. In future, we are planning to find semantic associations based on the J.UCS RDF Graph. Furthermore, the J.UCS dataset may be interweaved with other Linked Data publications of scientific libraries like IEEE, CiteSeer and ACM.

References

[Afzal et al., 2008] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2008). Expertise Finding for an Electronic Journal, In: Proceedings of International Conference on Knowledge Management and Knowledge Technologies, pp. 436-440, Graz, Austria, 3-5, Sep. 2008.

[Afzal, 2009] Afzal, M. T.: Information Supply of Related Papers from the Web for Scholarly e-Community, Lecture Notes in Business Information Processing, vol. 45, pp. 61-72 (2010).

[Afzal et al., 2009] Afzal, M. T., Latif, A., Ussaeed, A., Sturm, P., Aslam, S., Andrews, K., Tochtermann, K., Maurer, H.: Discovery and Visualization of Expertise in a Scientific Community, In Proc. International Conference of Frontiers of Information Technology, Islamabad, Pakistan, 16-18, Dec. 2009.

[Afzal et al 2009a] Afzal, M. T., Balke, W., T., Kulathuramaiyer, N., Maurer, H.: Rule based Autonomous Citation Mining with TIERL, Journal of Digital Information Management, 8 (3), 196-204 (June 2010)

[Afzal , 2010] Afzal, M. T.: Context Aware Information Discovery for Scholarly e-Community, PhD thesis, Graz University of Technology, Austria, 2010.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R.and Ives, Z.: DBpeddia: A Nucleus for a Web of Open Data, In Proc. 6th International Semantic Web Conference, Springer, Busan, Korea, 11- 15, Nov. 2007.

[Auer et al. 2009] Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., Aumueller, D.: Triplify - Lightweight Linked Data Publication from Relational Databases. In Proceedings of the International World Wide Web Conference (WWW 09), Madrid, Spain, pp. 621-630, 2009.

[Berners-Lee 2006] Berners-Lee, T.: Linked Data Design Issues. 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[Bizer and Cyganiak 2006] Bizer, C., Cyganiak, R.: D2R Server – Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference (ISWC), Nov. 2006.

[Bizer et al., 2009] Bizer, C., Heath, T., Berners-Lee, T., : Linked data – the story so far; International Journal on Semantic Web and Information Systems (IJSWIS), 2009.

[Breslin et al., 2005] Breslin, J. G., Harth, A., Bojars, U., Decker, S.: Towards Semantically-Interlinked Online Communities. In Proceedings of the Second European Semantic Web Conference, ESWC 2005, May 29- June 1, 2005, Heraklion, Crete, Greece, 2005.

[Brickley and Miller, 2004] Brickley, D., Miller, L.: FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/.

[Calude et al., 1994] Calude, C., Maure, H., Salomaa, A.: Journal of Universal Computer Science, In Journal of Universal Computer Science 0 (0), pp. 109-116, 1994.

[Candela et al., 2009] Candela, L., Castelli, D., Fuhr, N., Ioannidis, Y., Klas, C.-P., Pagano, P., Ross, S., Saidis, C., Schek, H.-J., Schuldt, H., Springmann, M.: Current Digital Library Systems: User Requirements vs Provided Functionality, Deliverable D1.4.1, Mar. 2006.

[Cheng et al., 2008] Cheng, G., Ge, W., Qu, Y.: Falcons: Searching and Browsing Entities on the Semantic Web. In: Proceedings of 17th International World Wide Web Conference, pp. 1101-1102, Beijing, China, 21-25 Apr. 2008.

[Coetzee et al. 2008] Coetzee, P., Heath, T., Motta, E.: Sparqplug: Generating linked data from legacy html, sparql and the DOM. In Proceeding of the CEUR-WS Vol-369 of Linked Data on the Web (LDOW2008), Beijing, China, 2008.

[Ding et al., 2004] Ding, L., Finin, T., Joshi, A., Pan, R., S. Cost, R., Peng, Y., Reddivari, P., C. Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. In: Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, pp. 652 - 659, Washington, D.C., USA, 8-13, Nov. 2004.

[Hausenblas 2009] Hausenblas, M.: Linked Data Applications. Technical Report, DERI, 2009.

[Hepp et al., 2006] Hepp, M., Siorpaes, K., Bachlechner, D.: Harvesting Wiki Consensus Using Wikipedia Entries as Vocabulary for Knowledge Management, IEEE Internet Computing. 11(5), pp. 54-65. 2007.

[J.UCS 2011] Journal of Universal Computer Science. 2011. http://www.jucs.org

[Krottmaier, 2003] Krottmaier, H.: Links to the Future, Journal of Digital Information Management, In Journal of Universal Computer Science 1 (1), pp. 3-8, 2003.

[Latif et al., 2009] Latif, A., Tanvir, M.T., Hoefler, P., UsSaeed, A., Tochtermann, K.: Translating Keywords into URIS, In proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 24-26 Nov. 2009.

[Latif et al., 2010] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K.: Harvesting Pertinent Resources from Linked Data, In Journal of Digital Information Management (JDIM) 8 (3), pp. 205-212, June 2010.

[Latif et al., 2010a] Latif, A. Afzal, M. T., Helic, D., Tochtermann, K., Maurer, H.: Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal)", LDOW at World Wide Web conference 2010, April 24-30, 2010, Raleigh, North Carolina.

[Marchionini and Maurer 1995] Marchionini, G., Maurer, H.: The roles of digital libraries in teaching and learning, Communication of the ACM, vol. 38, No. 4, pp. 67-75, 1995.

[Maurer, 2001] Maurer, H.: Beyond Digital Libraries. Global Digtial Library Development in the New Millenium, In: Proceedings of NIT Conference, pp.165-173, Beijing, China, 2001.

[Lay, 2009] Michael, L.: DBLP - Some Lessons Learned. PVLDB 2(2), pp. 1493-1500, 2009.

[Oren et al., 2008] Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies, 3(1), pp. 37–52, 2008.

[Roberts et al., 2001] Roberts, R.J., Varmus, H.E., Ashburner, M., Brown, P.O., Eisen, M.B., Khosla, C., Kirschner, M., Nusse, R., Scott, M., Wold, B.: Building A GenBank of the Published Literature. Science, 291 (5512), 2318-2319.

[Scharffe et al., 2009] Scharffe, F., Liu, Y., Zhou, C.: RDF-AI: an Architecture for RDF Datasets Matching, Fusion and Interlink. In Proceedings of the IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena, CA US, 2009.

[Suchanek et al., 2007] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge, In Proc. 16th international World Wide Web conference, ACM Press, Banf, Alberta, Canada , 8-12, May, 2007.

[Volz et al., 2009] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk – A Link Discovery Framework for the Web of Data. In Proceeding of the 2nd Workshop about Linked Data on the Web. 2009.