CEUR-WS-LOD: Conversion of CEUR-WS Workshops to Linked Data

Maxim Kolchin[1], Eugene Cherny[1,2], Fedor Kozlov[1], Alexander Shipilo[3], and Liubov Kovriguina[1(✉)]

[1] ITMO University, Saint-petersburg, Russia
{kolchinmax,eugene.cherny}@niuitmo.ru,
{kozlovfedor,alexandershipilo,lkovriguina}@gmail.com
[2] Åbo Akademi University, Turku, Finland
[3] Saint-Petersburg State University, Saint-petersburg, Russia

Abstract. CEUR-WS.org is a well-known place for publishing proceedings of workshops and very popular among Computer Science community. Because of that it's an interesting source for different kinds of analytics, e.g. measurement of workshop series popularity or person's contribution to the field by organizing workshops and etc. For realizing an insightful and effective analytics one needs to combine information from different places that can supplement each other. And this brings a lot of challenges which can be mitigated by using Semantic Web technologies.

Keywords: Information extraction · RDF · Semantic publishing · Linked open data · CEUR-WS

1 Introduction

"Semantic publishing refers to publishing information on the Web as documents accompanied by semantic markup"[1] using RDFa or Microformats, or by publishing information as data objects using Semantic Web technologies such as RDF and OWL. One of the areas where semantic publishing is actively used is scholarly publishing, where it helps bring improvements to scientific communication "by enabling linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers" [9].

We don't aim to survey the state-of-art of semantic publishing of scientific research in this paper, but we suggest to look at the existing works [1,5,6,9] and papers presented at the series of Workshops on Semantic Publishing[2] for more in-depth overview.

This paper presents a contribution to semantic publishing of scientific research by conversion of a well-known web-site for publishing proceedings

[1] Cf. http://en.wikipedia.org/wiki/Semantic_publishing.
[2] Cf. http://ceur-ws.org/Vol-1155/.

F. Gandon et al. (Eds.): SemWebEval 2015, CCIS 548, pp. 142–152, 2015.
DOI: 10.1007/978-3-319-25518-7_12

of workshops to Linked Data dataset. The work is carried out in framework of Semantic Publishing Challenge 2015[3], is based on the previous effort [3] extended by improving precision/recall of the information extraction and the ontology model.

Source Data. The source of data is CEUR-WS.org that publishes proceedings of workshops starting from 1995th year and is very popular among Computer Science community. At the time of writing, it contains information about 1346 proceedings and around 130 ones are added each year, over 19 000 papers and more than 33 000 people.

Challenges. As was described in the previous work [3], extraction of the needed information from the CEUR-WS's web pages faces several challenges, some of them:

- the web pages don't have uniform structured markup, therefore it's not feasible to relay on a single template for mapping data to RDF,
- 41.5 % of proceedings' web pages don't contain any markup, such as RDFa or Microformats. But even pages having the markup don't always follow its structure and semantics,
- a big part of the proceedings are jointly published by several workshops, e.g. http://ceur-ws.org/Vol-1244/ includes papers of ED2014 and GViP2014 workshops.

Table 1. Namespaces and prefixes used in the paper

Prefix	URL
swc	http://data.semanticweb.org/ns/swc/ontology#
bibo	http://purl.org/ontology/bibo/
swrc	http://swrc.ontoware.org/ontology#
owl	http://www.w3.org/2002/07/owl#
foaf	http://xmlns.com/foaf/0.1/
dcterms	http://purl.org/dc/terms/
dc	http://purl.org/dc/elements/1.1/
rdfs	http://www.w3.org/2000/01/rdf-schema#

Structure of the Paper. The structure of the paper is as follows. Section 2 presents our approach. Section 3 explains the ontology model and mappings to some well-known ontologies. Section 4 gives an overall view of the dataset and lists SPARQL query examples. Also Sect. 4 describes how the dataset is published and how users can access the data. The last section concludes the work and results. The prefixes used throughout the paper are defined in Table 1.

[3] Cf. http://github.com/ceurws/lod/wiki/SemPub2015.

2 System Description

In this work we apply *knowledge engineering approach* to the design of Information Extraction systems which requires expression of *rules* for the system are constructed by hand using knowledge of the application domain [2].

Although this approach is laborious, the results of the previous challenge shown that it's performance much higher than the others [4]. The system submitted last year reached overall average precision/recall equal to 0.707/0.636 correspondingly while the next best result was 0.478/0.447.

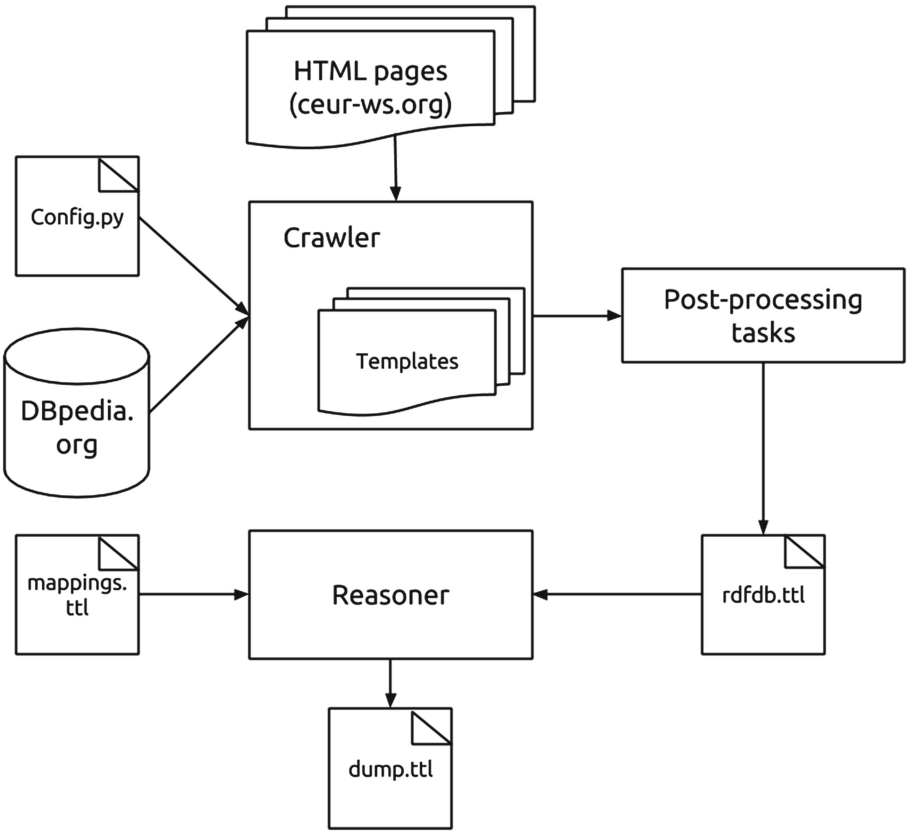

Fig. 1. Workflow of conversion CEUR-WS.org to linked data

The system developed to convert CEUR-WS.org to Linked Data dataset implements the workflow outlined in Fig. 1. The workflow consist of three major steps:

– crawling the web pages and serializing the extracted information to RDF,
– processing the resulted RDF dump to merge resources of persons with similar names, e.g. Dusan Kolář and Dusan Kolar is actually the same person, therefore he should be represented by a single resource,
– applying the mapping ontology to link the data to well-known ontologies.

The source code is open sourced and available at https://github.com/ailabitmo/ceur-ws-lod under the MIT License.

Crawling. In the system the *rules* are expressed using XPath expressions which constitute a *template* of an HTML block. The system has a separate template for each different HTML block presented on the web site's pages. These templates are run by the crawler implemented using Grab framework[4] that provides Python API for creating crawlers.

There are two *abstract templates* which aren't used by the crawler directly, but are used by the other templates as basis: *Parser* and *ListParser*. The difference between them is that *ListParser* is used for repeatable structures such as Table of Content of proceedings or list of proceedings on the index page.

The crawler groups the templates by the web site pages, such as *index*, *proceedings*, *publication*. There are 11 templates. In Table 2 all these *templates* with corresponding RegExp expressions that is used to categorize the pages are presented.

Table 2. Templates grouped by the web site's pages

Web page	RegExp	Template name
index	^http://ceurs-ws\.org/*$	ProceedingsRelations
		WorkshopSummary
		WorkshopAcronym
		WorkshopRelations
		ProceedingsSummary
proceedings	^http://ceur-ws\.org/Vol-\d+/*$	WorkshopPage
		EditorAffiliation
		EditorNameExpand
		JointWorkshopsEditors
		Publication
publication	^http://ceur-ws\.org/Vol-\d+/*\.pdf$	PublicationNumOfPages

Each such *template* is a Python class which extends *Parser* or *ListParser* classes and has one or more methods having *parse_template_* string as prefix in its name. The crawler executes these methods one by one while one of them matches the HTML block. After that the method extracts the information and passes it for the serialization.

[4] Cf. http://grablib.org/.

Name Disambiguation. At the post-processing step the system does the disambiguation of the peoples' names by fuzzy-matching sorted of tokenized name-string. The *fuzzywuzzy*[5] library was used for this task. For each pair of names in the dataset we have performed the following operations:

1 String normalization: convert to ASCII representation, make lowercase.
2 Split name string into tokens using whitespace separator and sort tokens in string.
3 Perform fuzzy string matching between token-sorted strings.

Entities that have similar names were interlinked with *owl:sameAs* property and exported as separate file[6].

We do not have tools to estimate correctness of the persons' interlinking, thus we only performed manual validation of the output file[7]. The results in general are good, except two moments. First, the algorithm has the $O(n^2)$ complexity and it took more than 12 hours to perform comparison of all names. Second, due to the nature of fuzzy string matching, the algorithm recognized a group of 32 persons with Asian names as one. This is due to the common names and surnames, such as "Li", and short lengths of the name-surname combination—the string matching algorithm often returns high similarity measure in such occasions.

Mapping to Well-Know Ontologies. The last step is to map the ontology used by the system to several well-known ontologies. To do it a parser based on Jena Inference API[8] was implemented which supports several RDFS and OWL constructs such as *rdfs:subClassOf*, *rdfs:subPropertyOf*, *rdf:type*, *owl:equivalentClass*, *owl:sameAs* and etc.

3 Ontology Model

We considered three ontologies for use as the basis of semantic representation of the crawled data:

– Semantic Web Conference Ontology (SWC) is an ontology for describing academic conferences,
– Semantic Web for Research Communities (SWRC) is an ontology for modeling entities of research communities such as persons, organisations, publications and their relationship,

[5] Cf. https://github.com/seatgeek/fuzzywuzzy.
[6] Cf. https://github.com/ailabitmo/ceur-ws-lod/releases/download/ceur-ws-crawler-v1.0.0/task-1-persons-sameas.ttl.
[7] Cf. https://github.com/ailabitmo/ceur-ws-lod/blob/master/ceur-ws-crawler/post processing/merged_persons.json.
[8] Cf. https://jena.apache.org/documentation/inference/index.html.

– Bibliographic ontology (BIBO) is an ontology providing main concepts and properties for describing citations and bibliographic references (i.e. quotes, books, articles, etc.).

Unfortunately, each of those ontologies alone are not sufficient to fully represent the structure of crawled information. For example, *BIBO* doesn't have an "event is part of bigger event" semantics and with *SWRC* we can't explicitly describe how many pages are in a publication, as *swrc:pages* could be used for describing page region, e.g. *255–259*; *SWC* reuses *SWRC*, thus they share the same limitations, and *SWC* does not introduce entities relevant for our work. Of course, this is not full list of all incompletenesses of those ontologies, but we think that detailed ontology comparison is out of scope of this paper, therefore we refer the reader to existing works [7,8]. Thus, based on subjective evaluation we decided to use *SWRC* as much as possible and add terms from other ontologies only if *SWRC* does not contain needed semantics. The structure of resulting ontology is represented on the Fig. 2.

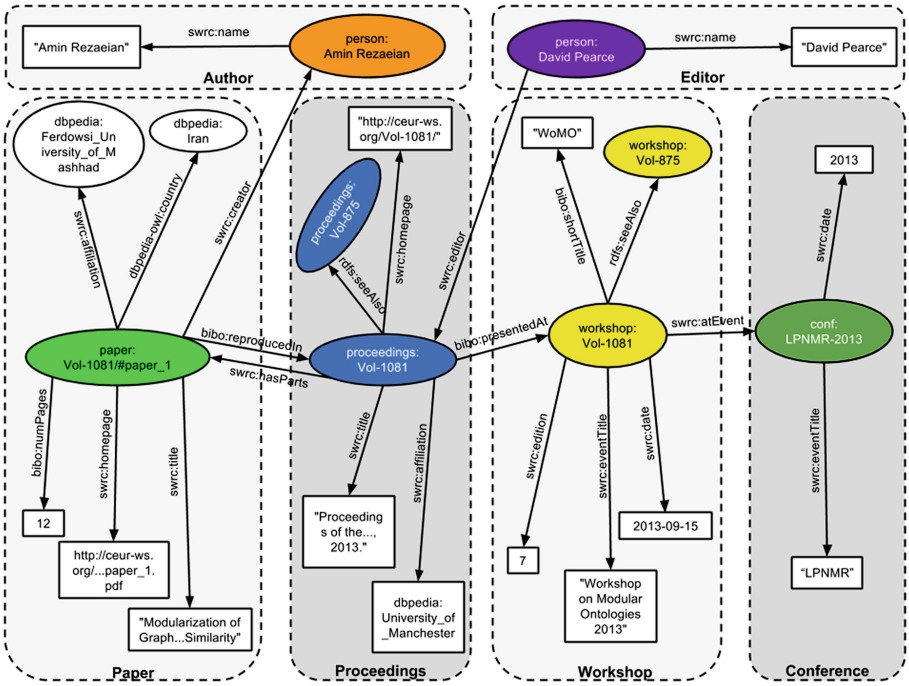

Fig. 2. Semantic representation of the crawled data

We used *SWC* ontology only once to mark a paper as invited one, making it an individual of class *swc:InvitedPaper*, because *SWC* is the poorest of those three ontologies in terms of semantic richness: the number of properties is much

lesser than in others, some classes have names like "Event-1" and "Role-1", a lot of them are deprecated, and, last but not least, official site of ontology is not accessible, so we were forced to download the ontology from the third-party site[9], which doesn't contain imports *SWC* depends on. All those factors suggest that development of this ontology was halted before reaching consistent usable state—this is why we tried to avoid using it in our work.

A concept "series of events" is not described in any of these ontologies, thus we choose to link workshops of the same series with the *rdfs:seeAlso* property. To keep things consistent we decided to use this approach to link "series of proceedings" and not to make additional *bibo:Series* class.

3.1 Mapping to Well-Know Ontologies

To compensate semantic inconsistencies in the resulting data set introduced by usage of properties and classes from different ontologies, we created the mappings between ontologies with *owl:eqivalentProperty* and *owl:eqivalentClass* properties. We interlinked only *BIBO* and *SWRC* ontologies, as *SWC* already has some dependencies on *SWRC*.

The full list of the mappings:

```
## Conference ##
swrc:Conference     owl:equivalentClass     bibo:Conference,
                                            swpo:Conference ;
                    rdfs:subClassOf         swc:OrganizedEvent .

## Workshop ##
swrc:Workshop       owl:equivalentClass     bibo:Workshop,
                                            swpo:Workshop ;
                    rdfs:subClassOf         swc:OrganizedEvent .
swrc:eventTitle     rdfs:subPropertyOf      rdfs:label, dcterms:title .
bibo:shortTitle     rdfs:subPropertyOf      rdfs:label, dcterms:title .
swc:isSubEventOf    owl:equivalentProperty  swrc:atEvent .
timeline:atDate     owl:equivalentProperty  swrc:date ;
                    rdfs:subPropertyOf      dcterms:date .

## Proceedings ##
swrc:Proceedings    owl:equivalentClass     bibo:Proceedings ;
                    rdfs:subClassOf         foaf:Document .
foaf:homepage       rdfs:subPropertyOf      foaf:page ;
                    owl:equivalentClass     swrc:homepage .
dcterms:issued      owl:equivalentProperty  bibo:created,
                                            swrc:creationDate ;
                    rdfs:subPropertyOf      dcterms:date .
swrc:editor         owl:equivalentProperty  bibo:editor ;
                    rdfs:subClassOf         foaf:maker,
```

[9] Cf. http://lov.okfn.org/dataset/lov/vocabs/swc.

```
                                    dcterms:creator .
swrc:title          owl:equivalentProperty  foaf:title ;
                    rdfs:subPropertyOf      rdfs:label, dcterms:title,
                                            dc:title .

## Paper ##
swrc:InProceedings  owl:equivalentClass     bibo:Article ;
                    rdfs:subClassOf         foaf:Document .
swrc:creator        owl:equivalentProperty  foaf:maker,
                                            dcterms:creator .

## Person ##
swrc:Person         owl:equivalentClass     bibo:Person, foaf:Person .
foaf:Person         rdfs:subClassOf         foaf:Agent .
foaf:Agent          owl:equivalentClass     dcterms:Agent .
swrc:name           owl:equivalentProperty  foaf:name ;
                    rdfs:subPropertyOf      rdfs:label .
```

4 Overview of Dataset

Publishing. The data is published using a Linked Data Fragments [10] server and available at http://data.isst.ifmo.ru. The users can use a Linked Data Fragments client for querying the data using SPARQL language. Or the data is also available as an HDT[10] dump in the GitHub repository[11].

Statistics. The dataset includes 402 648 triples and 55 893 subjects. The distribution of resource types are depicted on Fig. 3.

In absolute numbers the dataset includes information about 1 344 proceedings, 1 360 workshops, 18 875 regular and 203 invited papers, 252 conferences, 33 859 persons with 2 657 editors.

4.1 Example Queries

In this section several SPARQL queries are presented which provide some interesting insights.

Query 1. Top-10 persons how was an editor of the highest number of workshop series:

```
SELECT ?editor (COUNT(DISTINCT ?workshop) AS ?count) {
  {
    SELECT DISTINCT ?workshop {
```

[10] Cf. http://www.rdfhdt.org/.
[11] Cf. http://github.com/ailabitmo/ceur-ws-lod.

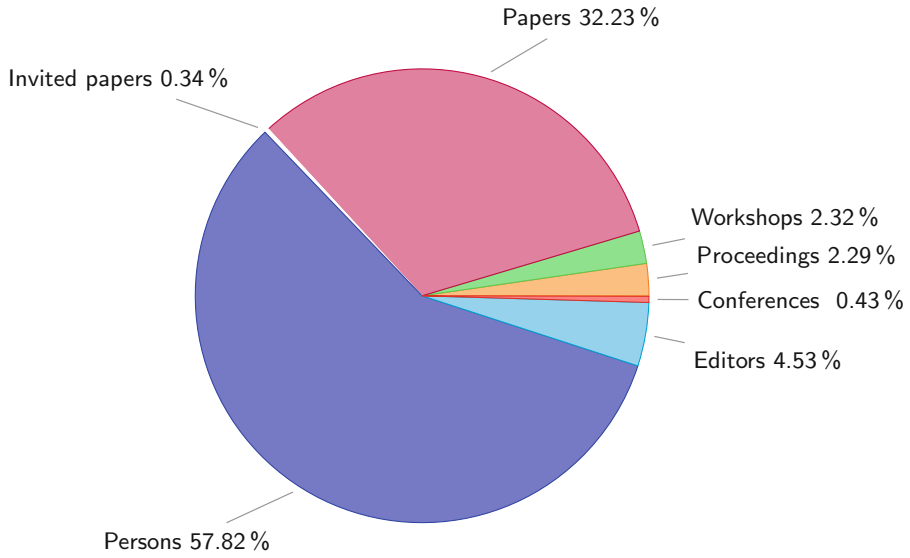

Fig. 3. Distribution of resource types in the dataset (# of triples – 402 648)

```
      ?workshop a bibo:Workshop ;
        rdfs:seeAlso ?inseries .
      FILTER NOT EXISTS { [] rdfs:seeAlso ?workshop }
    }
  }
  {
    ?proceedings a swrc:Proceedings ;
      swrc:editor ?editor .
    { ?proceedings bibo:presentedAt ?workshop }
    UNION
    { ?proceedings bibo:presentedAt ?inseries .
      ?workshop rdfs:seeAlso ?inseries .
    }
  }
}
GROUP BY ?editor
ORDER BY DESC(?count)
LIMIT 10
```

Query 2. Top-10 workshops with the highest number of authors:

```
SELECT ?workshop (COUNT(DISTINCT ?author) as ?num_authors) {
  ?paper a swrc:InProceedings ;
    swrc:creator ?author ;
```

```
    dcterms:partOf ?proceedings .
  ?proceedings bibo:presentedAt ?workshop .
}
GROUP BY ?workshop
ORDER BY DESC(?num_authors)
LIMIT 10
```

Query 3. Latest workshops of top-10 workshop series with the longest history.

```
SELECT ?workshop (COUNT (?related) + 1 AS ?count) WHERE {
  ?workshop a bibo:Workshop ;
    rdfs:seeAlso ?related .
  FILTER NOT EXISTS { [] rdfs:seeAlso ?workshop .}
}
GROUP BY ?workshop
ORDER BY DESC(?count)
```

5 Conclusion

In this paper we described a system that converts a well-known web-site for publishing proceedings of academic events, called CEUR-WS.org, to Linked Data dataset. Also we described semantic representations (ontologies) that are used to create the dataset. The system is based on *knowledge engineering approach* to design Information Extraction systems.

To overview the resulted dataset we introduced some statistical information, such as amount of papers and proceedings. Also we presented example SPARQL queries which provide some interesting insights from the extracted information.

The presented system is developed in the framework of Semantic Publishing Challenge 2015[5] and based on the previous work [3] which was extended with richer semantic representations and was improved in terms of precision and recall.

Acknowledgments. This work has been partially financially supported by the Government of Russian Federation, Grant #074-U01.

References

1. Auer, S., Lange, C., Ermilov, T.: Towards facilitating scientific publishing and knowledge exchange through linked data. In: Bolikowski, Ł., Casarosa, V., Goodale, P., Houssos, N., Manghi, P., Schirrwagen, J. (eds.) TPDL 2013. CCIS, vol. 416, pp. 10–15. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-319-08425-1_2
2. Eikvil, L.: Information extraction from world wide web - a survey. Technical report, July 1999. http://user.phil-fak.uni-duesseldorf.de/rumpf/SS2003/Informationsextraktion/Pub/Eik99.pdf

3. Kolchin, M., Kozlov, F.: A template-based information extraction from web sites with unstable markup. In: Presutti, V., et al. (eds.) SemWebEval 2014. CCIS, vol. 475, pp. 89–94. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-319-12024-9_11
4. Lange, C., Di Iorio, A.: Semantic publishing challenge – assessing the quality of scientific output. In: Presutti, V., et al. (eds.) SemWebEval 2014. CCIS, vol. 475, pp. 61–76. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-319-12024-9_8
5. Nevzorova, O., Zhiltsov, N., Zaikin, D., Zhibrik, O., Kirillovich, A., Nevzorov, V., Birialtsev, E.: Bringing math to LOD: a semantic publishing platform prototype for scientific collections in mathematics. In: Alani, H., et al. (eds.) ISWC 2013, Part I. LNCS, vol. 8218, pp. 379–394. Springer, Heidelberg (2013). http://dx.doi.org/10.1007/978-3-642-41335-3_24
6. Peroni, S.: Semantic Publishing: issues, solutions and new trends in scholarly publishing within the Semantic Web era. Ph.D. thesis, Universit di Bologna (2012). http://dx.doi.org/10.6092/unibo/amsdottorato/4766
7. Peroni, S., Shotton, D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Web Semant. Sci. Serv. Agents World Wide Web **17**, 33–43 (2012). http://www.sciencedirect.com/science/article/pii/S1570826812000790
8. Ruiz Iniesta, A., Corcho, O.: A review of ontologies for describing scholarly and scientific documents. http://ceur-ws.org/Vol-1155#paper-07
9. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learn. Publ. **22**(2), 85–94 (2009). http://www.ingentaconnect.com/content/alpsp/lp/2009/00000022/00000002/art00002
10. Verborgh, R., et al.: Querying datasets on the web with high availability. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 180–196. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-319-11964-9_12