Undefined 1 (0) 1–5
IOS Press

COLINDA - Conference Linked Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Selver Softic [a], Laurens De Vocht [b], Erik Mannens [b] and Rik Van de Walle [b]
[a] *Social Learning, Graz University of Technology, Münzgrabenstrasse 35A, 8010, Graz, Austria*
E-mail: selver.softic@tugraz.at
[b] *Multimedialab, Ghent University - iMinds , Sint-Pietersnieuwstraat 25, 9000, Ghent, Belgium*
E-mail: {laurens.devocht,erik.mannens,rik.vandewalle}@ugent.be

Abstract. We introduce a new LOD (Linked Open Data) Cloud member COLINDA (COnference LInked DAta) which exposes information about scientific events like conferences and workshops for the period from 2007 up to 2011. COLINDA includes also time and venue information of the scientific events which is interlinked to the GeoNames Linked Data set. The main sources of COLINDA are WikiCfP and Eventseer. COLINDA holds information about conferences from all over the world and contains information about 6000 scientific events generating around 140000 triples. More then 25000 new conferences are to come. This paper provides an introduction on the conference linked dataset,and demonstrates its applicability for adoption of web 2.0 into science also known as Research 2.0 or Science 2.0

Keywords: Linked Data, COLINDA, scientific events, Linked Science, Research 2.0

1. Introduction

We present our current work on the COLINDA[1] dataset which contains information about scientific events (conferences, workshops) from all over the world. Data published currently in COLINDA is extracted from the dumps of WikiCfP[2] and data has been extracted via JSON interface from Eventseer[3]. Data from Eventseer still waits to be published due to the still pending publication permission approval. Some parts of WikiCfP data are still in processing stage, which means that several thousands of additional conferences will be added consequently. COLINDA includes data of about 6000 conferences. Around 30000 additional conferences are planned to follow till the end of the year. In section 2 we describe the nature and origins of data and purpose of creation of the content included in COLINDA. Additionally we provide information about the availability and accessibility regarding this data. Section 3 is reserved for the details about the dataset creation process, while the Section 4 is aiming at outlining the usage of COLINDA. Finally in Section 5 we draw conclusions, discuss the limitations and report on our future work.

2. COLINDA Dataset

2.1. Data Source and Coverage

The main data sources of COLINDA are WikiCfP and Eventseer. Those are two very popular online Web 2.0 pages containing data about calls for papers, locations and topics of conferences as well as concrete call for papers description. Such pages can be considered as scientific event announcement pages editable by the users with archiving character. In order to add some event a user has to be registered and as soon he or she enters an event, a review process by internal editors is started in order to validate the entry qual-

[1] Available at: http://data.colinda.org/
[2] http://www.wikicfp.com
[3] http://eventseer.net

ity. This process is more strict by Eventseer then by WikiCfP. Data for these pages is provided by the scientific community users involved into organisation of such events. WikiCfP also provides annual data dumps for previous years in XML format. Currently WikiCfP contains data about around 30.000 conferences with around 100.000 registered researchers. Eventseer contains according the latest infromation[4] information about around 21000 events and serves more then 1 million users. Scientific events from both pages date from 2002 up to now also including the future events in the next year. Listing 1 shows a simple entry from an WikiCfP data dump that was used to create instances from COLINDA, while listing 2 represent the JSON source from Eventseer.

Listing 1: Sample entry from WikiCfP data dump.

```
<row>
 <field name="eventid">11426</field>
 <field name="createdate">2010−09−13 07:22:18</field>
 <field name="fullname">The 10th International Semantic Web Conference</field>
 <field name="handle">ISWC</field>
 <field name="year">2011</field>
 <field name="location">Koblenz, Germany</field>
 <field name="begindate">2011−10−22</field>
 <field name="finishdate">2011−10−27</field>
 <field name="presubdate">2011−06−16</field>
 <field name="submitdate">2011−06−23</field>
 <field name="notifydate">2011−08−08</field>
 <field name="cameradate">2011−08−28</field>
 <field name="weblink">http://iswc2011.semanticweb.org</field>
 <field name="info">cfp text ...</field>
</row>
```

Listing 2: Sample entry from Eventseer JSON interface.

```
{
 "totalRecords": 1,
 "records":
 [
  {
  "Event":
  "<a href='/e/19285/'>
  13th IEEE/ACM international symposium on cluster,
  cloud and grid computing (CCGRID 2013)
  </a>",
  "City": "Delft",
  "Country": "Netherlands",
  "Date": "10 Nov 2012",
  "NextDeadline": "22 Nov 2012",
  "EndDate": "16 May 2013",
  "StartDate": "13 May 2013"
  }
 ]
}
```

Current exported data covers generally two domains. The first domain describes the **Conference** as basic scientific event with a start date, location, description, label and link to the event. The **Location** is then in interlinking process resolved using the GeoNames[5] data set. Each location contains reference to the city, country and coordinates of the location.

2.2. Purpose of Creation

The intention behind COLINDA was initially to provide tag based identification system for scientific events in the manner of the "5-star" quality Open Data[6]. Users in social microblogs like Twitter[7] are often using so called "hash tags" to describe an event they are attending. E.g. ISWC (International Semantic Web Conference) 2012 is often referred as "iswc12" or "iswc2012". Also the DBLP linked data set uses this kind of notation to reference the event where a publication belongs to[8]. More generally COLINDA is meant to be event driven connection data set for scientific LOD (Linked Open Data) Cloud datasets and to support in this way the efforts of Linked Science[9] initiative as well to be used as mining reference for creation of semantically driven microblog data Mesh Ups for Research 2.0 as it will be shown on simple example in section 4. Research 2.0 also known as Science 2.0 is a initiative in research community to adapt Web 2.0 for the needs of scientific community. Currently several efforts in this direction are running or has been done as e.g. DBLP data set containing data about publications or ResearchGate[10] a social network for scientists just to mention some of them. Events specific for the research community seem to be very intuitive base for connecting to the context since researches with the same interests seem to track and visit similar events. Further appliance of COLINDA would be also to connect information about the scientific events with other scientific data sets of relevance e.g. information about publications, citations, journal informations etc.

[4] http://eventseer.net/data/
[5] http://www.geonames.org/
[6] http://5stardata.info/
[7] http://www.twitter.com/

[8] for "iswc2012" : http://dblp.l3s.de/d2r/page/publications/conf/ISWC/2012
[9] http://linkedscience.org/
[10] http://www.researchgate.net/

2.3. Licensing and Availability

WikiCfP is a Semantic Wiki and supports "creative commons" Attribution-ShareAlike 3.0 License[11] which was also beside the page popularity decisive by choosing the data source for COLINDA. Eventseer however is run and developed by *Abiody AS* and *Thomas Brox Røst*, which implies an explicit request for permission to re-use the data we collected. Currently published data is available at http://data.colinda.org.

3. Linked Data creation process

The data creation process contains basically following steps:

- Extraction - preprocessing and harmonisation of data sources (Subsection 3.1)
- Ontology choice - concept coverage (Subsection 3.2)
- Triplification - creating RDF data triples (Subsection 3.3)
- Interlinking - connection to other Linked Data sets (Subsection 3.4)

Additionally information about data characteristics and URI design are described in subsections 3.5 and 3.6.

3.1. Extraction

Since COLINDA consumes data from different sources a minimal set of properties that describe a **Conference** concept for a single RDF instance has to be defined . All properties from source data sets has to be mapped to this normalized set in order to harmonize the federated data for import. **Location** concept related to conference event as such is considered as optional enrichment treated in the interlinking process. This decision was made because of the fact that all conference descriptions do not explicitly include the venue information. The quality of source data depends on the users that provide the information. Thus such data sources implicitly exclude assumption of completeness.Table 1 represents the minimal set of properties a **Conference** and **Location** instance should include.

[11] http://creativecommons.org/licenses/by-sa/3.0/

Table 1
Harmonised COLINDA instances minimal properties set.

Concept	***Property***
Conference	*label*
	title
	description
optional	*date*
	link
	location
Location	*placename*
	city
	country
	longitude
	latitude

The Extraction process includes steps of preprocessing XML and JSON inputs from WikiCfP and Eventseer into the series of values saved into Comma Separated Value (CSV) form as result. During the preprocessing cycle data fields like e.g. date are normalised to uniform representation in order to provide easier processable input for triplification step which converts the extracted values into RDF formated instances.

3.2. Ontology Choice

Representation of scientific events was already elaborated in previous research work [1]. Minimal field set defined in table 1 for RDF instance generation fitted very well to the existing ontologies. This is the reason why we have chosen the SWRC Ontology [1] and basic RDFS Schema[12] as established vocabularies to describe **Conference** instances. Same approach was applicable for **Location** concept. Geographical property set was easily mappable into combination of GeoNames[13] and Basic Geo (WGS84) Vocabulary[14]. How completely supported mapped model with interlinking looks like can be seen in figure 1, where a single complete and interlinked instance of Conference is depicted. How properties fit to the vocabulary properties can be seen in table 2.

3.3. Triplification

Triplification process uses as input in extraction, harmonisation and preprocessing step generated CSV

[12] http://www.w3.org/TR/rdf-schema/
[13] http://www.geonames.org/ontology/
[14] http://www.w3.org/2003/01/geo/wgs84_pos#

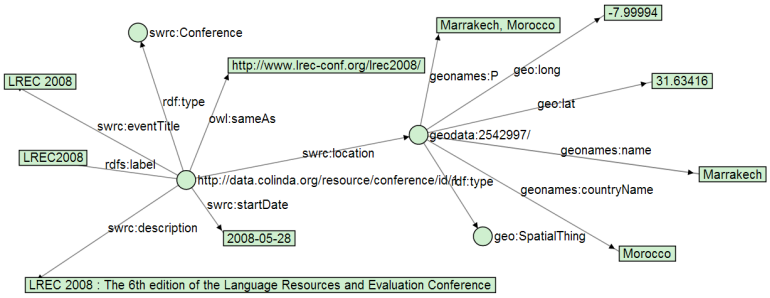

Fig. 1. Sample interlinked **Conference** RDF instance.

Table 2
COLINDA concept to ontology model mapping (note: geonames - GeoNames Ontology, geo - W3C GEO Vocabulary, swrc - SWRC Ontology).

Concept/Property	***RDF Class/Property***
Conference	swrc:Conference
label	rdfs:label
title	swrc:eventTitle
description	swrc:description
date	swrc:startDate
link	owl:sameAs
location	swrc:location
Location	geo:SpatialThing
placename	geonames:P
city	geonames:name
country	geonames:countryName
longitude	geo:long
latitude	geo:lat

output. Input generated in this way represents tabular set of values compatible with properties from table 1. This CSV file is then imported into appropriate MySQL database table that holds COLINDA data. From this place data is conference wise generated as single RDF instance using the vocabulary properties defined in table 2 and each RDF instance of a single conference svent is recallable via REST (Representational State Transfer) calls as described in subsection 3.5. In order to provide the whole data via SPARQL endpoint batch process synchronises the data from MySQL database table into the ARC2[15] RDF triple store running on the same database server.

3.4. Interlinking

In order to provide 5-star data and led by the design issues described in [2], we used ***swrc:location*** as interlinking property in order to interlink the location data with GeoNames. The interlinking process uses cURL[16] requests against GeoNames query service to resolve geographical information and retrieve coordinates. Although usually ***owl:sameAs*** is used to interlink to other data set we used this property to resolve the connection to the conference web page and since ***swrc:location*** seems regarding the GeoNames to be more appropriate choice. How this connection looks like can be seen in the sample depicted in figure 1. Due the fact that conference information entered by users does not always include location data, we generates some data statistics about how many of overall instances are linked to GeoNames as well some other interesting facts which are presented in subsection 3.6.

3.5. URI Design and Data Set Publication

There are generally two ways to access instances of COLINDA. First way is using the direct access to single RDF instances via URIs designed as follows:

- http://data.colinda.org/conference?id={id}
- http://data.colinda.org/resource/conference/{id}

The {id} represent the COLINDA's internal id that starts by integer value of 1 up to the number of current instances which is around 6000. All responses are in RDF/XML notation. The REST like access is realised suing the REST PHP Framework Restler 2.0 version[17]. The second access possibility is via the SPARQL endpoint that is accessible at:
http://data.colinda.org/endpoint.php.
This endpoint supports up to 250000 result triples per query and retrieves results in various widely supported formats like JSON, RDF/XML, XML, TSV etc. How

[15] https://github.com/semsol/arc2/

[16] http://curl.haxx.se/

[17] http://luracast.com/products/restler/

Table 3
Conference counts by country in COLINDA.

Year	*Count*
2007	416
2008	2141
2009	2665
2010	768
2011	13

the endpoint can be queried is described by simple example in listings 3.

Listing 3: Sample SPARQL query for retrieval of conference meta fields.

```
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf−schema#>
SELECT ∗
{
 ?x rdfs:label "LREC2008"
 OPTIONAL
 {
   ?x swrc:description ?des;
      swrc:keywords ?key;
      swrc:startDate ?st;
      swrc:location ?loc.
 }
}
```

Further executable samples of SPARQL queries can be found at:
http://data.colinda.org/endpoint.html.
Additional information about COLINDA and the RDF data dumps is also accessible via the CKAN Registry of LOD Cloud:
http://datahub.io/dataset/colinda.

3.6. Data Characteristics

In order to offer a short overview about conferences we created a table of counts of conference instances per year as well an overview over conference occurence with respect to the venue presented in the tables 3 and 4. As it can be seen in table 3 currently most conferences date from 2008 and 2009 since those dumps from WIkiCfP has been imported completely yet. The top 5 conference count cover in sum only 1/6 part of the whole locations contained. Thus table 4 reveals us that the location dissemination of conferences tends to vary strongly inside COLINDA. A full resolution of interlinking is provided only when the location property ***swrc:location*** is included in instance making triples, which means only in this cases a connection to GeoNames is present for the conference instance. In order to approve the interlinking quality COLINDA was analysed upon this property. All **Conference** RDF instances which included the location property has been set to value 1 while others to degree 0. Out of this data a normed histogram over the whole range of instances in COLINDA has been generated. Histogram for this analysis can be seen in figure 2, showing very high degree of interlinked instances.

Table 4
Top 5 conference counts by country in COLINDA.

Country	*Count*
China	258
Germany	248
Italy	229
France	203
Spain	162
Japan	110

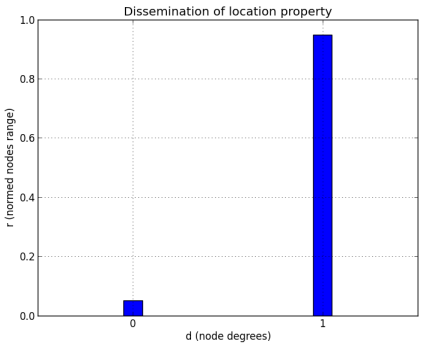

Fig. 2. Percentage of instances including the ***swrc:location*** property.

4. Usage

Based upon example of "Researcher Affinity Browser" [3] we want to demonstrate a possible appliance case for COLINDA linked data set."Researcher Affinity Browser" has been developed in the realm of our research as semantically driven microblog data Mesh Up for the needs of Research 2.0. COLINDA was used as mining source for the facetted distinction of scientist Twitter profiles based upon conferences they visited as special affinity criteria. Subsection 4.1 offers a short functional overview over the application as well the role that data from COLINDA plays within. Ad-

equate demo video showing the "Researcher Affinity Browser" in action can be also viewed online[18].

4.1. Researcher Affinity Browser

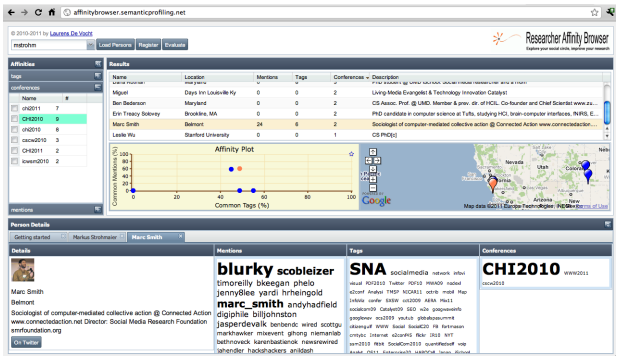

Fig. 3. "Researcher Affinity Browser Application" snapshot.

The "Researcher Affinity Browser Application" [3] is depicted in figure 3. At the beginning it retrieves a list of relevant users. Those results represent a current snapshot which means that every time users produce new tweets on Twitter, the analysis result evolves with it. The relevance is measured according to the number of common conceptual affinities.Different affinity facets are displayed on the left. Users can explore three types of affinities: conferences, tags and mentions. Activation of a certain affinity filters the list of matching persons. There is the result table that displays detailed information about each person and how many affinities are shared. Further there is a map view and an affinity plot synchronized with the result table. The affinity plot visualizes in a quick overview affinity correspondence between the analyzed profile and other profiles in the system.

5. Discussion and Future Work

As outlined in subsection 4.1 COLINDA has been already used as mining source for generation and enhancement of *Researcher Affinity Browser* [3] a semantically driven microblog data Mesh Up interface for Research 2.0. However a shortcoming of current state of the COLINDA is still low number of triplified conference instances that counts currently about 6000 which produces around 140000 triples. We are aiming at publishing the rest of the conference data (additional 25000 up to 30000 conferences) we extracted as soon we get the permissions and finish the preprocessing. The interlinking process with cURL requests is very time consuming, therefore we will use GeoNames dump instead. Further we want to interlink via conference labels our conference entries to the DBLP (Digital Bibliography and Library Project)[19] Linked Data Set. DBLP also provides dumps which will make the process more easiers as it was in the case of GeoNames. Extending the REST interface to allow retrieval of triples of single conferences in the manner:
http://url/resource/conference/label/year
will be included in the next steps. We also want to implement an Lookup[20] and Spotlight[21] service comparable to the DBPedia and a facetted search interface like the one provided by the DBLP. In order to approve the quality of COLINDA an evaluation against Linked Data Integration Benchmark (LODIB)[22] will be done.

[18] http://www.youtube.com/watch?v=A25DrP3Mv8w

Acknowledgement

The research activities that have been described in this paper were funded by Ghent University, the Social Learning Department at Graz University of Technology, iMinds (an independent research institute founded by the Flemish government to stimulate ICT innovation), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

References

[1] Y. Sure,S. Bloehdorn, P. Haase, J. Hartmann and D. Oberle,*The SWRC ontology - Semantic Web for research communities*, in: Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005), 2005, pp. 218–231.

[2] T. Berners-Lee, *Linked Data - Design Issues*, *http://www.w3.org/DesignIssues/LinkedData.htm*, W3C, 2006.

[3] L. De Vocht, S. Softic, M. Ebner, and H. Mühlburger, *Semantically driven social data aggregation interfaces for Research 2.0*, in: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, New York, NY, USA, ACM, 2011, pp.43:1–43:9.

[19] http://dblp.l3s.de/
[20] http://lookup.dbpedia.org
[21] https://github.com/dbpedia-spotlight/
[22] http://lodib.wbsg.de/