COLINDA: Modeling, Representing and Using Scientific Events in the Web of Data

Selver Softic[1], Laurens De Vocht[2], Martin Ebner[1],
Erik Mannens[2], and Rik Van de Walle[2]

[1] Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria
{selver.softic,martin.ebner}@tugraz.at
[2] Ghent University, iMinds - Multimedialab,
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
{laurens.devocht,erik.mannens,rik.vandewalle}@ugent.be

Abstract. Conference Linked Data (COLINDA)[3], a recent addition to the LOD (Linked Open Data) Cloud[4], exposes information about scientific events (conferences and workshops) for the period from 2002 up to 2015. Beside title, description and time COLINDA includes venue information of scientific events which is interlinked with Linked Data sets of GeoNames[5], and DBPedia[6]. Additionally information about events is enhanced with links to corresponding proceedings from DBLP (L3S)[7] and Semantic Web Dog Food [8] repositories. The main sources of COLINDA are WikiCfP[9] and Eventseer[10]. The research questions addressed by this work in particular are: how scientific events can be extracted and summarized from the Web, how to model them in Semantic Web to be useful for mining and adapting of research related social media content in particular micro blogs, and finally how they can be interlinked with other scientific information from the Linked Data Cloud to be used as base for explorative search for researchers .

Keywords: Linked Data, Scientific Events, Linked Science, Research 2.0

1 Introduction and Motivation

COLINDA[11] contains information about scientific events worldwide (including location and proceedings references), published as Linked Data. The data contained in COLINDA is extracted and accumulated from the data dumps of WikiCfP , which are published yearly and freely available on request for research[12] purposes, and from data

[3] http://colinda.org

[4] http://lod-cloud.net/

[5] http://www.geonames.org/

[6] http://dbpedia.org

[7] http://dblp.l3s.de/d2r/

[8] http://data.semanticweb.org/

[9] http://www.wikicfp.com/

[10] http://eventseer.net/

[11] Available at: http://colinda.org/, see also http://datahub.io/dataset/colinda

[12] http://www.wikicfp.com/cfp/data.jsp

gathered via JSON interface from Eventseer. WikiCfP and Eventseer are two very popular online scientific event archives. WikiCfP contains calls for paper for about approximately 30.000 conferences and has approximately 100.000 registered users. Eventseer contains according the latest information[13] calls for around 21000 events and serves more then 1 million users. We also track the Twitter[14] feeds of both sites integrating on the fly arrival of upcoming scientific events using the Twitter API[15] to recieve the data from Twitter profiles of Wiki CfP and Eventseer. Currently COLINDA includes data about more than 15000 conferences. Event instances are enriched with information from Linked Data proceedings repositories DBLP (L3S)[16] and Semantic Web Dog Food[17] as well by location information from Geonames and DBPedia. Primary intention of COLINDA was to provide hashtag based identification system for scientific events in Twitter in the manner of the "5-star" quality Open Data[18]. Researchers are using very often hashtags, while they are discussing on Twitter. Specially during scientific events, they are using hashtags as abbreviated reference to the event they are attending [6]. E.g. ISWC (International Semantic Web Conference) 2012 is often referred as "iswc12" or "iswc2012". DBLP (L3S) Linked Dataset and Semantic Web Dog Food also use this kind of notation to reference the event of conference proceedings[19],[20]. The overall idea of COLINDA is to serve as mining reference for creation of semantically driven microblog data Mash Ups for Research 2.0 and as interlinking hub for other science relevant sources from the LOD cloud in order to enhance explorative search for researchers. Efforts made in this field using COLINDA will be introduced in detail in section 3.

2 Extraction, Modeling, Creation and Publishing of Linked Scientific Events

COLINDA data covers generally three domains: The first domain originates from WikiCfP and Eventseer and describes the **Conference** as basic scientific event with a start date, location, description, label and link to the event web page. Second domain is the **Location** of the event with geographic parameters resolved using the GeoNames and DBPedia data set in interlinking process. Each location contains reference to the city, country and coordinates of the location. Further as extension and third domain we have **Proceedings** of the conference represented by the links from DBLP (L3S) or Semantic Web Dog Food.

2.1 Linked Scientific Events Creation Process

The data creation process comprises the following steps:

[13] http://eventseer.net/data/

[14] http://www.twitter.com/

[15] COLINDA

[16] http://datahub.io/dataset/l3s-dblp

[17] http://datahub.io/dataset/semantic-web-dog-food

[18] http://5stardata.info/

[19] e.g. for 'iswc2012' at DBLP(L3S): http://dblp.l3s.de/d2r/page/publications/conf/ISWC/2012

[20] e.g. for 'iswc2012' at SW Doog Food: http://data.semanticweb.org/conference/iswc/2012/

- Extraction - extraction and pre-processing of data sources (Subsection 2.2)
- Modeling of Events using SWRC Ontology - concept coverage (Subsection 2.3)
- Triplification - creating RDF data triples (Subsection 2.4)
- Interlinking - connection to other Linked Data sets (Subsection 2.5)

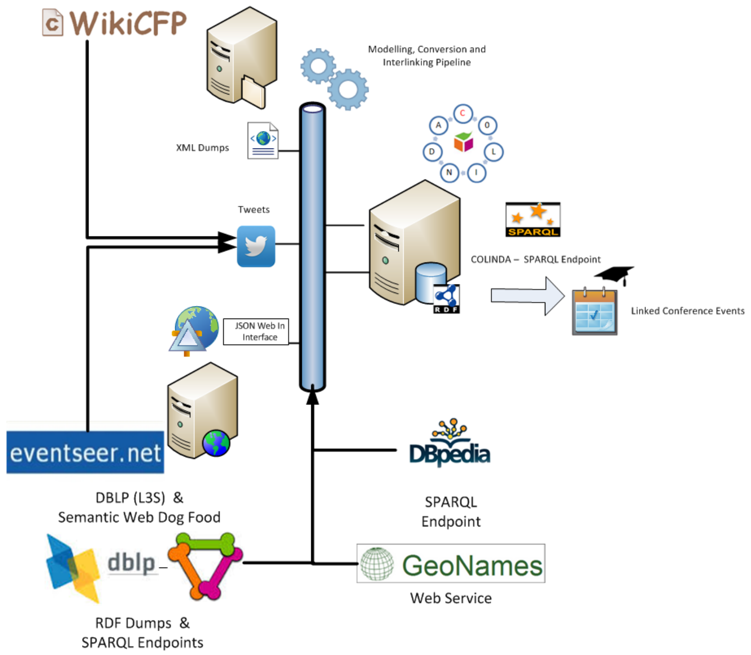


Fig. 1. Creation process of linked scientific events.

2.2 Data Extraction

COLINDA is constructed from variously structured sources. Therefore we defined a minimal set of properties that describe the **Conference** concept for a single RDF instance. During extraction, all properties from sources are being mapped to defined normalized set in order to harmonize the federated data. The **Location** and **Proceedings** concepts related to conference events as such are considered as optional enrichment which will be treated in the interlinking process. We made this decision having in mind 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. The Extraction process includes steps of either pre-processing of XML dumps from WikiCfP or JSON from Tweets and Eventseer into the temporary tables of values formatted as Comma

Separated Value (CSV). During the pre-processing cycle data fields like e.g. date or labels are being normalized to achieve uniform representation, and to provide easier processable input for triplification step which converts the extracted values from temporary tables into RDF formatted instances of Linked Data.

Table 1. Harmonized COLINDA - minimal properties set. Entries denoted with * are optional.

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

2.3 Modeling Scientific Events in the Web of Data

Basic representation of scientific events was well elaborated in previous research work about the SWRC ontology introduced by Sure et al [7]. This practice has been already approved and adapted by the implementation of Linked Data proceedings repositories DBLP (L3S) and Semantic Web Dog Food. We also followed the good practice of re-using existing vocabularies before we define our own. Minimal field set defined in table 1 for RDF instance generation matches well the range of SWRC concepts. Therefore, we have chosen the SWRC Ontology[21] and basic RDFS Schema[22] as established vocabularies to describe **Conference** instances. The same approach was applied for **Location** concept; needed set of geographical features to describe conference venues is well covered by elements from GeoNames[23] and Basic Geo (WGS84) Vocabulary[24]. Complete model with interlinked properties (proceeding and location) can be seen in figure 2, where a single complete and interlinked instance of a conference (ISWC2012) is depicted. Matchings between features and the vocabulary properties is shown in table 2.

2.4 Triplification - Creation of RDF Instances of Scientific Events

The triplification[25] process uses as input temporary data tables in CSV like format generated in extraction and pre-processing step. Input generated in this way represents tab-

[21] http://ontoware.org/swrc/

[22] http://www.w3.org/TR/rdf-schema/

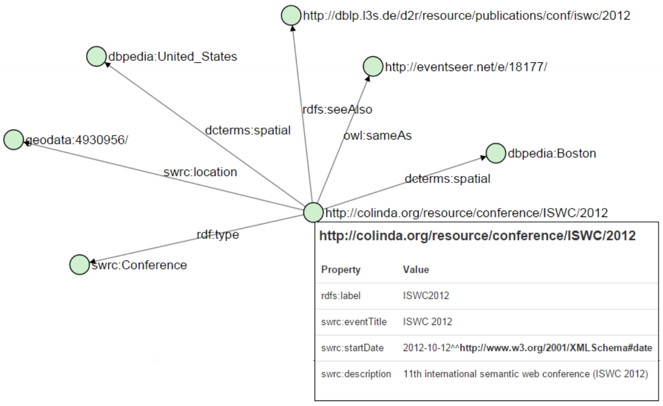
[23] http://www.geonames.org/ontology/

[24] http://www.w3.org/2003/01/geo/wgs84_pos#

[25] Under 'triplification' we understand 'triple-wise' creation of Linked Data instances as RDF graphs

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

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



http://colinda.org/resource/conference/ISWC/2012

Property	Value
rdfs:label	ISWC2012
swrc:eventTitle	ISWC 2012
swrc:startDate	2012-10-12^^**http://www.w3.org/2001/XMLSchema#date**
swrc:description	11th international semantic web conference (ISWC 2012)

Fig. 2. Sample interlinked **Conference** RDF instance of ISWC 2012 generated by Visual RDF.

ular set of values compatible with properties from table 1. This input is then parsed line by line and conference instance is generated as single RDF graph using the vocabulary properties defined in table 2. Each conference instance is accessible via REST (Representational State Transfer) call as described in subsection 2.6. To make them accessible by SPARQL endpoint, background batch process loads the conference instances into the ARC2[26] RDF triple store running on the server.

2.5 Interlinking to Other Interesting Sources

In order to provide 5-star data and led by the design issues described in [1], we used ***swrc:location*** as interlinking property in order to interlink the location data with GeoNames. The interlinking process uses GeoNames query service to resolve geographical

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

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 2 as well as online[27],[28]. Further we use dumps of DBPedia and Semantic Web Dog Food to enhance the instances with DBPedia location info using the ***dcterms:spatial*** property and for interlinking the proceedings from DBLP (L3S) ans Semantic Web Dog Food we match the conference's ***rdfs:label*** to the corresponding labels in those data sets via SPARQL queries. In matching case a link is established with correlating results using the ***rdfs:seeAlso*** property.

2.6 URI Design and Public Accessibility

Access to instances of COLINDA is possible via URIs with following pattern:

- http://colinda.org/resource/conference/{label}/{year}

All responses from COLINDA are formatted as RDF/XML fragment. Other supported formats are: HTML, Text, N3, NTRIPLES format[29]. Alternative access offers the SPARQL [30] endpoint. Current endpoint supports up to 250000 result triples per query and delivers results in different formats like: JSON, RDF/XML, XML, TSV etc. How to query the endpoint is shown by simple example in listings 1.1. Results from the query return the COLINDA link, city, country and the geo-location of ISWC 2012 conference. Recently, a dump of COLINDA was made available as Linked Data Fragments[31]. COLINDA RDF data dumps are also accessible via the CKAN Registry[32] of LOD Cloud.

Listing 1.1. Sample SPARQL query for retrieval of conference (geo) location.

```
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf−schema#>
SELECT DISTINCT ?x ?city ?country ?long ?lat
{
 ?x rdfs:label "ISWC2012";
    swrc:location ?loc.
    OPTIONAL
    {
              ?loc gn:name ?city;
                 gn:countryName ?country;
                 geo:lat ?lat;
                 geo:long ?long.
    }
}
```

[27] http://www.colinda.org/resource/conference/ISWC/2012?format=html

[28] http://graves.cl/visualRDF/?url=www.colinda.org/resource/conference/ISWC/2012

[29] e.g. http://www.colinda.org/resource/conference/ISWC/2012?format=html

[30] http://colinda.org/endpoint.php

[31] http://data.linkeddatafragments.org/colinda#dataset

[32] http://datahub.io/dataset/colinda

2.7 Actuality of Data

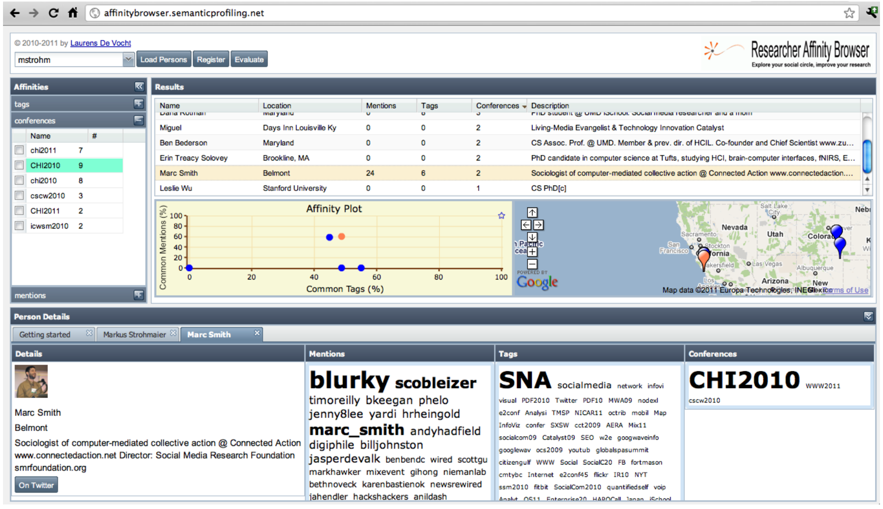
COLINDA is kept up-to-date by a daily cron job which grabs the newest event announcements over the Twitter API for accounts of WikiCfP and Eventseer. The cron job parses, creates, interlinks and synchs new events into the triple store. Each tweet also includes information about the call page link which allows retrieval of the extended information about events via web (WikiCfP) or available JSON (Eventseer) interface during the update task. Additionally to the automated job, also manual updates are ran as soon as the fresh dumps from both sites are available.

3 Applications and Use Cases

Both use cases introduced in following subsections address the challenges of Research 2.0. Research 2.0 as adaptation of the Web 2.0 for researchers defines researchers as main consumers of the information. The purpose of these research activities is to offer a set of tools and services which researcher can use to discover resources, such as publications or events they might be interested in, as well as to collaborate with each other via the web. These tools and services, according to the specifications of Research 2.0, are considered as Mash Ups, APIs, publishing feeds and specially designed interfaces based on social profiles [5, 8]. The role of COLINDA is addressed separately in application description.

3.1 Affinity Browser

The "Researcher Affinity browser" was developed as a tool to demonstrate semantically driven aggregation of microbolog data for Research 2.0. (use of Web 2.0 tools in scientific research). In this context, COLINDA was used as mining source for the faceted detection of similar scientist Twitter profiles based upon conferences they visited as special affinity criteria. This is done by matching the COLINDA tags with the hashtags of the Twitter user. Adequate demo video showing the "Researcher Affinity Browser" in action can be also viewed online[33]. The "Researcher Affinity Browser Application" [4] 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 purpose of the map is to get a better impression of where the affiliations of the found persons would lie. The affinity plot visualizes in a quick overview affinity correspondence between the analyzed profile and other profiles in the system.



affinitybrowser.semanticprofiling.net
© 2010-2011 by Laurens De Vocht
mstrohm
Load Persons
Register
Evaluate
Researcher Affinity Browser
Explore your social circle, improve your research
Affinities
tags
conferences
mentions
Results
Affinity Plot
Common Mentions (%)
100
80
60
40
20
0
0
20
40
60
80
100
Common Tags (%)
Nevada
Utah
Sacramento
San Francisco
Stockton
Bakersfield
Las Vegas
Albuquerque
Arizona
New
Salt Lake City
POWERED BY Google
Map data ©2011 Europa Technologies, INEGI
Terms of Use
Person Details
Getting started
Markus Strohmaier
Marc Smith
Details
Mentions
Tags
Conferences
On Twitter
blurky scobleizer timoreilly bkeegan phelo jenny8lee yardi hrheingold marc_smith andyhadfield digiphile billjohnston jasperdevalk benbendc wired scottgu markhawker mixevent gihong niemanlab bethnoveck karenbastienok newsrewired jahendler hackshackers anildash
SNA socialmedia network infovi visual PDF2010 Twitter PDF10 MWA09 nodexl e2conf Analysi TMSP NICAR11 octrib mobil Map InfoViz confer SXSW cct2009 AERA Mix11 socialcom09 Catalyst09 SEO w2e googwaveinfo googlewav ocs2009 youtub globalspasummit citizengulf WWW Social SocialC20 FB fortmason cmtybc Internet e2conf45 flickr IR10 NYT ssm2010 fitbit SocialCom2010 quantifiedself voip
CHI2010 WWW2011 cscw2010

	Name	#
□	chi2011	7
□	CHI2010	9
□	chi2010	8
□	cscw2010	3
□	CHI2011	2
□	icwsm2010	2

Name	Location	Mentions	Tags	Conferences	Description
Miguel	Days Inn Louisville Ky	0	0	2	Living-Media Evangelist & Technology Innovation Catalyst
Ben Bederson	Maryland	0	0	2	CS Assoc. Prof. @ UMD. Member & prev. dir. of HCIL. Co-founder and Chief Scientist www.zu...
Erin Treacy Solovey	Brookline, MA	0	0	2	PhD candidate in computer science at Tufts, studying HCI, brain-computer interfaces, fNIRS, E...
Marc Smith	Belmont	24	6	2	Sociologist of computer-mediated collective action @ Connected Action www.connectedaction....
Leslie Wu	Stanford University	0	0	1	CS PhD[c]

Marc Smith

Belmont

Sociologist of computer-mediated collective action @ Connected Action www.connectedaction.net Director: Social Media Research Foundation smrfoundation.org

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

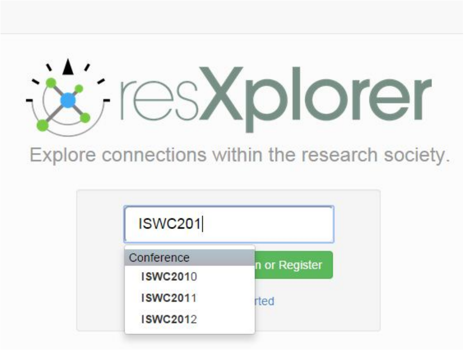


Fig. 4. Mapping of keywords

3.2 ResXplorer

"ResXplorer"[34] is an Research 2.0 [8] aggregated interface for search and exploration of the underlying Linked Data Knowledge Base. A demo video explaining the interface shown in figure 5 is available online[35]. Data from Linked Data Knowledge Base originates from: DBLP (L3S)[36] which is a bibliography of computer science conference proceedings, COLINDA[37] which is a main binding hub data set including informa-

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

[34] http://www.resxplorer.org

[35] https://www.youtube.com/watch?v=tZU97BQxE-0

[36] http://dblp.l3s.de/

[37] http://colinda.org

tion about scientific events and links to venue and proceedings, common Linked Data Knowledge Base DBPedia[38] and Open Linked Data repository with geographical information GeoNames[39]. The role of COLINDA is to act as a hub which connects all data sets in the knowledge base by pointing with links to other data sets. In this context it is used both for keyword matching together with other data sets and for enabling the algorithms in back-end to find better connections and paths between the terms visualized in the interface as well as for their expansion. Within ResXplorer interface a real-time keyword disambiguation guides researchers by expressing their needs. User selects the correct meaning from a typeahead drop down menu. Query expansion of terms happens in real-time. Figure 4 shows the typeahead expansion of "ResXplorer" in action. At the same time background modules also fetch neighbor links which match the selected suggestion. As result, selection of various resources is then presented to the researchers within radial interface. In case they have no idea which object or topic to investigate next, they get an overview of possible objects of interest (like points of interest on a street map e.g. figure 5). As shown in figure 5 features like color, shape and size of the items are used to enhance the guidance of the user during the search and exploration process [2]. Different shapes and colors represent different entities like conferences, persons, publications or proceedings. The explored items are marked black, and relations are marked red and clearly highlight the context and history of a search. Each presented resource is somehow related with some of other resources. This is expressed through lines and description of the relation which is a RDF property. The path distance in hops over links is expressed through the orbital layers.

As additional feature in ResXplorer is that researchers, when they sign in with their Twitter account, they can either use the mentions and hashtags automatically for search setup instead of typing keywords or to check visually the status of their network. This happens through visualization of recent collaboration and interactions based upon data from their Twitter accounts [3] (link to video on this procedure [40]). Figure 6 depicts the network of a researcher. The size of the scholar is in the middle between the minimum and maximum size of a node. As much as possible users are placed around the focused researcher. The more publications someone coauthored with the scholar, the bigger the node. Several visual aspects aid the user in focusing and exploring the current state of their network:

- *Spatial*: the number of *co-authorships* determines distance to center, a higher number results in a closer distance.
- *Size*: a higher frequency of being *mentioned* together on social media (i.e. Twitter) increases the size.
- *Color*: green, already in their Twitter network; red, not in their Twitter network.
- *Tooltip*: displays facts about the collaborations (e.g. co-authorships and mentions), i.e. the number of mentions for a specific conference (where conference identification is done over user hashtags which are automatically matched with COLINDA conference labels) and the the number of co-publications. The co-autorships are re-

[38] http://dbpedia.org

[39] http://geonames.org

[40] `http://youtu.be/QopnPvWIFzw`

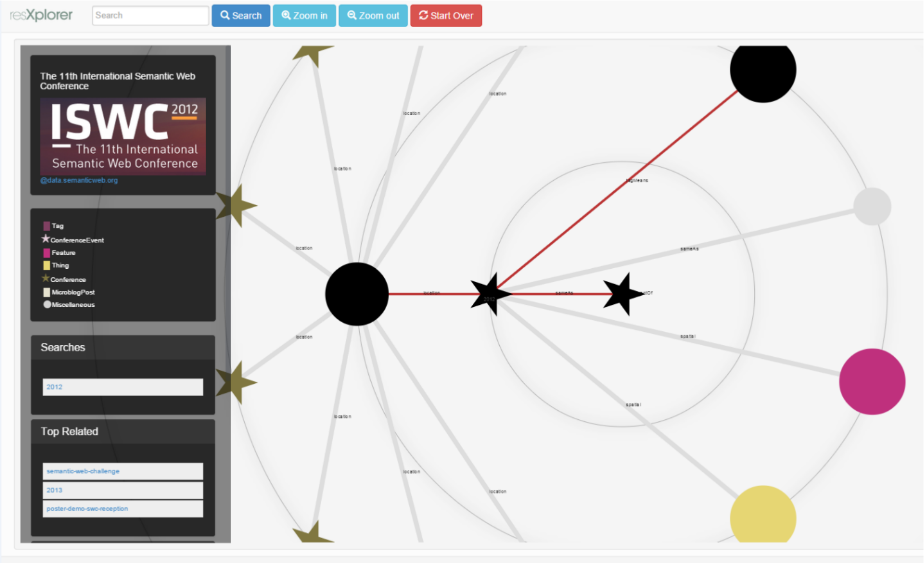
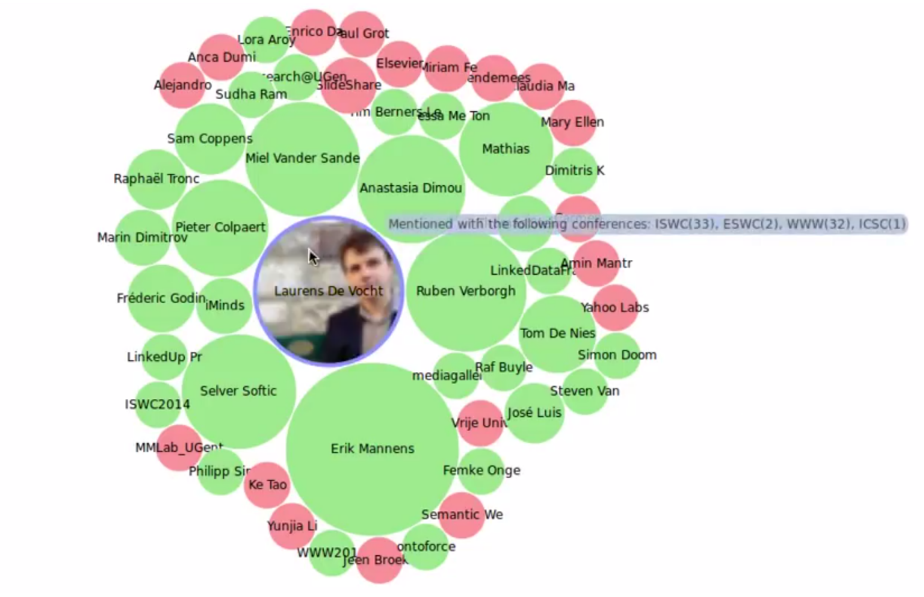


Fig. 5. ResXplorer - discovering scholar artifacts like conferences (represented as stars), miscellaneous related resources such as locations or microblog posts (represented as dots in different colors) etc. The distance to central node represents the intensity of relation.



Mentioned with the following conferences: ISWC(33), ESWC(2), WWW(32), ICSC(1)

Fig. 6. The scholar is centered in the middle and the network is visualized in nodes around the central (blue) node.

solved by bibliographic records from DBLP which are matched pair-wise between the users.

Users who whether have no co-autorships or common mentions and conference hashtags with central user profile are not included in visualization.

4 Conclusion and Outlook

In this work we described how we extract, model and create scientific events as Linked Data from known conference portals. We showed also how those events can be enhanced with additional relevant information and applied as as mining source for generation and enhancement of *Researcher Affinity Browser* as well as main interlinking hub for discovery of research related artifacts for the *ResXplorer*. This potential has been also recognized by the LinkedUp Challenge at the ISWC 2014[41] and upcoming Semantic Publishing Challenge 2015[42] at ESWC 2015 where COLINDA is nominated as reference Linked Data set for scientific events. As one of the future efforts we also want to implement a DBPedia Lookup[43] and Spotlight[44] like service for detection and identification of scientific events with COLINDA. We also want to link the instances to WorldCat URIs of the published proceeding volumes and to he Crossref DOIs of the published conference articles to make it more useful for the library linked data community. Finally, to verify the quality of COLINDA we will run in the future an evaluation against Linked Data Integration Benchmark (LODIB)[45].

Acknowledgments.

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. Berners-Lee, T.: Linked data (2006), `http://www.w3.org/DesignIssues/LinkedData.html`
2. De Vocht, L., Mannens, E., Van de Walle, R., Softic, S., Ebner, M.: A search interface for researchers to explore affinities in a linked data knowledge base. In: Proceedings of the 12th International Semantic Web Conference Posters & Demonstrations Track. pp. 21–24. CEUR-WS (2013)

[41] http://data.linkededucation.org/linkedup/catalog/browse/

[42] https://github.com/ceurws/lod/wiki/SemPub2015

[43] http://lookup.dbpedia.org

[44] http://dbpedia-spotlight.github.io/demo/

[45] http://lodib.wbsg.de/

3. De Vocht, L., Softic, S., Dimou, A., Verborgh, R., Ebner, M., Mannens, E., Van de Walle, R.: Visualizing collaborations and online social interactions at scientific conferences for scholarly networking. In: Proceedings of the Workshop on Semantics, Analytics, Visualisation: Enhancing Scholarly Data; 24th International World Wide Web Conference (May 2015)
4. De Vocht, L., Softic, S., Ebner, M., Mühlburger, H.: Semantically driven social data aggregation interfaces for research 2.0. In: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies. pp. 43:1–43:9. i-KNOW '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2024288.2024339
5. Parra Chico, G., Duval, E.: Filling the gaps to know More! about a researcher. In: Proceedings of the 2nd International Workshop on Research 2.0. At the 5th European Conference on Technology Enhanced Learning: Sustaining TEL,. pp. 18–22. CEUR-WS (Sep 2010)
6. Reinhardt, W., Ebner, M., Beham, G., Costa, C.: How people are using Twitter during conferences. In: Hornung-Prähauser, V., Luckmann, M.(Hg.): 5th EduMedia conference, Salzburg. pp. 145–156 (2009)
7. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D.: The SWRC ontology - Semantic Web for Research Communities. Progress in Artificial Intelligence pp. 218–231 (2005), http://dx.doi.org/10.1007/11595014_22
8. Ullmann, T.D., Wild, F., Scott, P., Duval, E., Vandeputte, B., Parra Chico, G.A., Reinhardt, W., Heinze, N., Kraker, P., Fessl, A., Lindstaedt, S., Nagel, T., Gillet, D.: Components of a research 2.0 infrastructure. In: Lecture Notes in Computer Science,. pp. 590–595. Springer (2010)