Undefined 1 (2012) 1–5
IOS Press

Eventseer: "Calls for Papers" as Linked Data

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

Thomas Brox Røst [a,*], Christophe Guéret [b], Amund Tveit [a] and Pablo Mendes [c]
[a] *Atbrox AS, Abels gate 5, 7030 Trondheim, Norway*
E-mail: {thomas,amund}@atbrox.com
[b] *Department of Artificial Intelligence, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands*
E-mail: c.d.m.gueret@vu.nl
[c] *Department of Information Systems, Freie Universität Berlin, Garystr. 21, 14195 Berlin, Germany*
E-mail: pablomendes@zedat.fu-berlin.de

Abstract. Finding relevant publication outlets is a necessity for all academics and researchers. The Eventseer web service was originally created to simplify this task by providing access to academic calls for papers in a semi-structured and searchable format. This paper describes the work being done to make the Eventseer data available as Linked Data, thereby further increasing its accessibility and usefulness to the scientific community. Details are given about the process of extracting necessary information such as event names and dates, deadlines and associated people, topics and organizations from the call for paper texts. The resulting mapping to a Linked Data RDF format and the modeling choices made are discussed. Examples of secondary use of Eventseer data are given; these include social network analysis of academic communities, altmetrics for measuring researcher impact, and automated modeling of topic hierarchies. Finally, a set of suggested improvements and known limitations are mentioned, along with plans for further improvement of the breadth and quality of the Linked Data.

Keywords: Linked Data, Eventseer, call for papers, academic event, information extraction

1. Introduction

The Linked Data Principles of Tim Berners-Lee [3] suggested a number of best practices for making structured data available on the World Wide Web. With Linked Data, Web URIs no longer just identify Web documents but can also be used to identify real-world entities and the links between them [23]. One of the major benefits of Linked Data is the potential for data discovery: the data description for a given data source may link to entities described by other data sources. This allows for the discovery of new data sources at runtime, and effectively sets the foundation for a "single global data space" rather than having isolated, topical data islands [4].

This paper gives an overview of the Eventseer [1] academic event data set and the on-going work of making it available as Linked Data. The data set originates from the Eventseer web service which was created to help academics and researchers find relevant scientific events and publication opportunities. The Linked Data, using a subset of the full database, is currently exposed as an alternative web service under a different namespace at `http://redux.eventseer.net`.

2. Eventseer Data

The primary function of the Eventseer web service is to allow for discovery and tracking of academic

*Corresponding author. E-mail: thomas@atbrox.com.

[1] `http://eventseer.net`

events and publication outlets such as conferences, workshops and journals. Since its inception in 1999, Eventseer has indexed academic calls for papers, or *CFPs*, and made them available through a search and discovery interface.

CFPs are proactively collected from mailing lists or received from users through a designated submission email address. Mailing lists such as Dbworld [1] and SEWORLD [2] are the traditional outlets for CFPs (at least within computer science) which increases the chances of finding relevant content. The choice of sources does give Eventseer a heavy bias towards CFPs related to computer science. Expanding coverage to other academic disciplines is on the future development roadmap.

2.1. Data Extraction and Modeling

The CFPs arrive as irregularly structured text and carry no additional markup information. We therefore apply various rule-based information extraction techniques to the CFP text so that the following entities are available as structured data:

- *Event name:* The name of an event or the name of a journal, book or similar publication.
- *Acronym:* Conferences and workshops will often have an acronym, so this is extracted if possible.
- *Event dates:* The start and end date for localized events, such as conferences and workshops.
- *Deadlines:* Deadlines have a descriptive text and a due date. There may be multiple deadlines in a CFP, such as deadlines for abstract submissions and participation registration.
- *Location:* The city, region and country of localized events.
- *People:* The names of known researchers.
- *Topics:* Science-related topic (e.g. "network security" and "human computer interaction").
- *Organizations:* Names of scientific organizations such as universities and research organizations.

All CFPs are associated with what we define as an *event*. A CFP can only be associated with a single event, while an event may have many CFPs. In the case of workshops and conferences, the event is the workshop or conference, while the CFPs are calls for abstracts, papers, participation and other relevant information for the given event. The event term is also used for publication outlets that are not strictly events, such as journals, competitions and summer schools. CFP types typically include calls for abstracts, papers, workshops, reviewers, participation and other related academic event communications.

Note that journal special issues are stored as unique events, rather than belonging to the main journal event. Also, satellite workshops associated with a main conference are stored as independent events if they have a given name and/or acronym. On the other hand, tracks and special sessions without specific names (e.g. "Special session on cloud computing" or "PhD student symposium") are usually stored under the main event. Each instance of periodically occurring events is typically categorized as a unique and independent event. As will be discussed later on, this simplified model was created for pragmatic reasons, primarily due to the limitations and complexity of automated data extraction and linking.

Event dates, if available, are given as a date range. Deadlines are given as a text and date tuple that indicate the type of submission and the due date. Typical deadline text types are "full paper", "camera ready paper", "posters", "participation", "workshop proposal", and so on. As will be discussed later, the deadline texts are neither normalized nor categorized into a deadline taxonomy; for now they are just plain strings extracted from the original CFP text.

Allowed locations are defined from a combination of city, region and country data from the GeoWorldMap [10] database and custom additions. A full city/region/country triple for the extracted location is generated where possible.

Named entities are found with the help of vetted gazetteer lists of topics, people and organizations. This ensures high precision at the cost of lower recall. These lists are updated both automatically and semi-automatically. For person names, sources such as the DBLP Computer Science Bibliography [18] are regularly queried. DBLP is also a convenient source for topics, although with some manual intervention. In addition, users can submit entity suggestions which are then manually approved.

New CFPs are automatically coupled to existing events via a text similarity measure, and a human administrator can approve or change the suggested connection.

The described relationships between core Eventseer resources are summarized in Figure 1. The decisions and trade-offs made when this model was originally created were at the time entirely bounded by the needs of the Eventseer application and the capabilities of the extraction technology. The primary goal was to make a web service that would simplify the act of finding

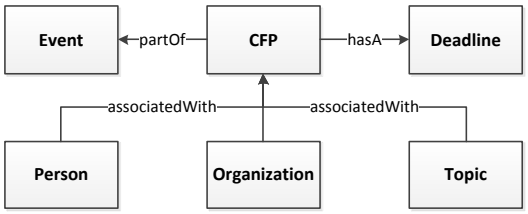

Fig. 1. Model entity relationships

Table 1
Data Set Statistics (date collected: May 21, 2012)

Resource	Number of items
Events	19,000
CFPs	41,700
People	1,033,000
Topics	5,200
Organizations	3,100

upcoming events. Although there were known limitations, such as the lack of granularity for handling collocated events, joint conferences and periodically occurring events, a choice was made to stick with a simple model. The primary motivation behind this choice was that populating a richer model would require more time spent and increased reliance on extraction technology and would as such come with the risk of reduced precision and data quality. There is a large potential for improving the core data model, for instance along the lines discussed in [25] and [28], and this is considered a major part of the future work for this project.

Table 1 shows some statistics about the data set size, derived from the data contained in the relational data base backend at the time of writing.

3. Eventseer as Linked Data

The Eventseer Linked Data is made available under the dedicated namespace `redux.eventseer.net`. The company Atbrox provides the funds for maintaining the data set and the web service and is responsible for the continuous availability of the Eventseer site. All the data is published as open data and is currently free for re-use in both commercial and non-commercial applications. Additional commercial data set access models are under consideration, this to generate revenue that will contribute to keeping the site running and the continuous growth and improvement of the data set.

Researchers from the Department of Artificial Intelligence at Vrije Universiteit Amsterdam have been responsible for the actual conversion to Linked Data and for making sure that best data publication practices have been followed.

Table 2
Linked Data URIs

Resource	URI structure
Events	`/event/<event_id>`
CFPs	`/event/<event_id>/cfp/<cfp_id>`
Deadlines	`/event/<event_id>/cfp/<cfp_id>` `/deadline/<deadline_id>`
People	`/person/<person_id>`
Topics	`/topic/<topic_id>`
Organizations	`/institution/<institution_id>`

3.1. URI scheme

The linked resources are made available through URI patterns that follow the same scheme as the one currently used to expose the corresponding HTML description. Table 2 gives an overview of the current URI structure.

3.2. RDF Examples and External Links

Listing 1 shows an example of a typical event entry, in this case for the 35th European Conference on Information Retrieval.

The event is defined as an academic event with an acronym via the Semantic Web Conference ontology (SWC) [19]. The event homepage is defined via the Friend of a Friend (FOAF) vocabulary. A start and end date is available, and we see that three related CFP resources are given. For this event a location is available through three GeoNames [11] links to the country/region/city triple of *Russia/Moskva/Moscow*.

Listing 2 shows one of the linked CFPs for this event. We observe that person and topic entities have been extracted for this CFP and that these are further defined as separate resources. Deadlines are also available.

For space reasons examples of deadline, person and topic resources are not given. It is worth noting that people and organization resources are currently defined in terms of the FOAF vocabulary, topics via SIOC Core Ontology types, and deadlines via the LODE [25] ontology. Work is underway to in-

Listing 1: Event RDF Example

```
<rdf:RDF>
  <rdf:Description rdf:about="http://redux.eventseer.net/event/19402">
    <rdf:type
     rdf:resource="http://data.semanticweb.org/ns/swc/ontology#AcademicEvent"/>
    <rdfs:label>35th european conference on information retrieval</rdfs:label>
    <swc:hasAcronym>ecir2013</swc:hasAcronym>
    <foaf:homepage>http://ecir2013.org/</foaf:homepage>
    <ical:dtstart rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
      2013-03-24</ical:dtstart>
    <ical:dtend rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
      2013-03-27</ical:dtend>
    <ical:location rdf:resource="http://sws.geonames.org/2017370/"/>
    <ical:location rdf:resource="http://sws.geonames.org/524901/"/>
    <ical:location rdf:resource="http://sws.geonames.org/524894/"/>
    <rdfs:seeAlso rdf:resource="http://redux.eventseer.net/event/19402/cfp/140919"/>
    <rdfs:seeAlso rdf:resource="http://redux.eventseer.net/event/19402/cfp/140789"/>
    <rdfs:seeAlso rdf:resource="http://redux.eventseer.net/event/19402/cfp/141097"/>
  </rdf:Description>
</rdf:RDF>
```

vestigate alternative and stronger vocabulary links, such as e.g. the Core Person Vocabulary [8]. The reader is encouraged to browse the RDF entries at `http://redux.eventseer.net` for further examples of event and CFP representations and the associated extracted data.

Work on establishing links with additional external resources is in progress. A first set of links has been created using the LATC (Linked Open Data Around-The-Clock) [17] platform[2]. The links there relate the entities exposed by Eventseer to their equivalent in DBLP through *sameAs*-relations. There is currently no RDF dump and no SPARQL endpoint (for running expressive queries against the data set) available although such extra services are planned.

4. Applications

Data from Eventseer has already been used in several research projects.

Klamma et al. used data from Eventseer and DBLP to create an academic event recommendation system based on social network analysis [15]. Their system was also used to visualize scientific communities with the purpose of analyzing how they developed over time. A fundamental assumption was that most people mentioned in Eventseer CFPs would be program committee members and that they were therefore representative of the community. In a similar application area, Stabeler et al. combined research interest data from the Academia.edu web service with event data from Eventseer in order to match research interests with academic events [26].

Jeong et al. used data from bioinformatics events to identify influential researchers in bioinformatics [14]. Their analysis suggested that elite-group membership in academic events can be used as a marker to measure the prominence of a scholar.

In a recent paper, Das et al. describe how Eventseer data was combined with data from Wordnet, Citeseer and Wikipedia to identify subtopics for a given topic, which greatly simplifies the task of creating domain-specific topic hierarchies [6]. In the long term, the work on exposing Eventseer data as Linked Data aims at making such data mashups easier to realize by doing the data integration up-front.

In general, there is currently a growing interest in alternative metrics for science (or *altmetrics*) [21] that also covers research activities that are currently not taken into account by classical metrics. The organization of a scientific event or participation in a program committee are two examples of such external activities that Eventseer makes it possible to harvest.

[2] See items starting with "eventseer" in the collection at *https://github.com/LATC/24-7-platform/tree/master/link-specifications*.

Listing 2: CFP RDF Example

```
<rdf:RDF>
  <rdf:Description rdf:about="http://redux.eventseer.net/event/19402/cfp/140789">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
    <rdfs:seeAlso
     rdf:resource="http://redux.eventseer.net/event/19402/cfp/140789/deadline/151937"/>
    <rdfs:seeAlso
     rdf:resource="http://redux.eventseer.net/event/19402/cfp/140789/deadline/151936"/>
    <rdfs:label>[Dbworld] ECIR 2013 Call for tutorials</rdfs:label>
    <lode:involvedAgent rdf:resource="http://redux.eventseer.net/person/49875"/>
    <dct:subject rdf:resource="http://redux.eventseer.net/topic/295"/>
    <dct:subject rdf:resource="http://redux.eventseer.net/topic/5260"/>
    <dct:subject rdf:resource="http://redux.eventseer.net/topic/2221"/>
    <dct:subject rdf:resource="http://redux.eventseer.net/topic/5261"/>
    <ical:dtstamp rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">
      2012-07-05</ical:dtstamp>
  </rdf:Description>
</rdf:RDF>
```

5. Related Work

The Event Ontology (EO) [22] is recognized as the most commonly used Linked Data event ontology [25]. Other ontologies of note include the CIDOC CRM [7], the ABC Ontology [16], EventsML-G2, DOLCE+DnS Ultralite (DUL) [9] and the Event-Model-F [24]. Shaw et al. gives an overview of various approaches towards event modeling and also suggests mappings between existing event ontologies to increase interoperability [25].

The mentioned ontologies primarily concern themselves with events in the broader sense: They represent general events that have happened or will take place, that occur during a time interval, and that may have participants and localities. The Semantic Web Conference Ontology [20,19] is an example of an ontology that is geared towards academic conferences, encompassing classes for event types, documents, roles and places. The Call Vocabulary [12] provides an ontology for calls for papers that also caters for deadlines. In a recent paper, Tomberg et al. [28] proposes another call for paper ontology, and also discusses relevant work in parsing and representing CFPs.

In addition, the Semantic Web Dog Food Corpus [27,20] contains some information that is also found in Eventseer.

6. Future Work and Conclusions

There are a number of improvements that can be made to the data set which will be dealt with in future research and development.

For practical use (i.e. finding upcoming publication opportunities) the deadlines should be categorized according to a deadline type definition or taxonomy. As an example, a camera-ready or revision deadline is only relevant for an author who has already been accepted for publication. Moreover, for events with many CFPs there may be many and often conflicting deadlines (e.g. when a submission deadline is extended). For the validity of the data set, more work into deadline disambiguation and normalization is needed.

Given that the data set has been used for social network analysis, altmetrics and automated topic hierarchy building, improving the quality of the named entities extracted from the CFPs is a priority. The data would be more useful if the context of the entities could be established, e.g. by recognizing a person as a program committee member or as being associated with a specific organization.

The current event database has a strong bias towards computer science. Adding more academic disciplines is on the long-term roadmap.

An often requested feature is to have the ability to follow periodically occurring events, such as annual conferences. As of now, each event is considered independent and any linking to previous events is the responsibility of the data user.

Finally, the data set originated from outside the Semantic Web and Linked Data communities and was created to solve a particular problem (i.e. publication opportunity search) without thinking about the bigger picture. This led to some ad-hoc modeling choices and clearly more work is needed in terms of both naming, intra-model relations and making it easier to adhere to best practices for Linked Data publication. A related area for future work is to link to more external data sources so that Eventseer data is more tightly connected to the Web of Linked Data.

References

[1] ACM SIGMOD. DBWorld. http://research.cs.wisc.edu/dbworld/.

[2] ACM SIGSOFT. SEWORLD. http://www.sigsoft.org/seworld/.

[3] Tim Berners-Lee. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[4] Christian Bizer. The Emerging Web of Linked Data. *IEEE Intelligent Systems*, 24(5):87–92, 2009.

[5] Sergey Brin. Extracting Patterns and Relations from the World Wide Web. In *WebDB '98 Selected papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, 1998.

[6] Sujatha Das, Prasenjit Mitra, and C Lee Giles. Phrase Pair Classification for Identifying Subtopics. *Advances in Information Retrieval*, 7224:489–493, 2012.

[7] Martin Doerr. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24(3):75–92, 2003.

[8] European Union. Core Person Vocabulary. http://joinup.ec.europa.eu/asset/core_person/home.

[9] Aldo Gangemi and Peter Mika. Understanding the Semantic Web through Descriptions and Situations. In *Proceedings of ODBASE03 Conference*, pages 689–706, 2003.

[10] Geobytes. GeoWorldMap. http://www.geobytes.com/freeservices.htm.

[11] GeoNames. GeoNames Ontology. http://www.geonames.org/ontology/documentation.html.

[12] Andreas Harth and Michael Kassoff. The Call Ontology. http://data.semanticweb.org/ns/swc/swc_2009-05-09.html, 2005.

[13] Marti A Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[14] Senator Jeong, Sungin Lee, and Hong-gee Kim. Are You an Invited Speaker? A Bibliometric Analysis of Elite Groups for Scholarly Events in Bioinformatics. *Journal of the American Society for Information Science and Technology*, 60(6):1118–1131, 2009.

[15] Ralf Klamma, Pham Manh Cuong, and Yiwei Cao. You Never Walk Alone: Recommending Academic Events Based on Social Network Analysis. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 4(1):657–670, 2009.

[16] Carl Lagoze and Jane Hunter. The ABC Ontology and Model. In *DCMI '01 Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 160–176, 2001.

[17] LATC. Linked Open Data Around-The-Clock. http://latc-project.eu/.

[18] Michael Ley. The DBLP Computer Science Bibliography: Evolution , Research Issues , Perspectives. In Alberto Laender and Arlindo Oliveira, editors, *String Processing and Information Retrieval*, volume 2476 of *Lecture Notes in Computer Science*, pages 481–486. Springer Berlin / Heidelberg, 2002.

[19] Knud Möller, Sean Bechhofer, and Tom Heath. The Semantic Web Conference Ontology. http://data.semanticweb.org/ns/swc/swc_2009-05-09.html, 2009.

[20] Knud Möller, Tom Heath, Siegfried Handschuh, and John Domingue. Recipes for Semantic Web Dog Food - The ESWC and ISWC Metadata Projects. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, pages 802–815. Springer-Verlag, 2007.

[21] Jason Priem, Dario Taraborelli, Paul Groth, and Cameron Neylon. Altmetrics: A Manifesto. http://altmetrics.org/manifesto/, 2010.

[22] Yves Raimond, Samer Abdallah, Mark Sandler, and Frederick Giasson. The Music Ontology. In *8th International Conference on Music Information Retrieval (ISMIR'07)*, Vienna, Austria, 2007.

[23] Leo Sauermann, Richard Cyganiak, and Max Völkel. Cool URIs for the Semantic Web. Technical Report 681, DFKI GmbH, 2007.

[24] Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. F-A Model of Events based on the Foundational Ontology DOLCE + DnS Ultralite. In *Proceedings of the fifth international conference on Knowledge capture*, pages 137–144, 2009.

[25] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. LODE: Linking Open Descriptions of Events. In *ASWC '09 Proceedings of the 4th Asian Conference on The Semantic Web*, pages 153–167, 2009.

[26] Matthew Stabeler, Graeme Stevenson, Simon Dobson, and Paddy Nixon. Basadaeir: Harvesting user profiles to bootstrap pervasive applications. In *7th International Conference on Pervasive Computing - Late Breaking Results*, 2009.

[27] SWSA. Semantic Web Dog Food. http://data.semanticweb.org/.

[28] Vladimir Tomberg, David Lamas, Mart Laanpere, Wolfgang Reinhardt, and Jelena Jovanovic. Towards a comprehensive call ontology for Research 2.0. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, 2011.

[29] Niklaus Wirth. What Can We Do about the Unnecessary Diversity of Notation for Syntactic Definitions? A Note On Reflection-Free Permutation Enumeration. *Communications of the ACM*, 20(11):822–823, 1977.