EINS Evidence Base: A Semantic Catalogue for Internet Experimentation and Measurement

Xin Wang[1], Thanasis G. Papaioannou[2], Thanassis Tiropanis[1], and Federico Morando[3]

[1] Web and Internet Science Group
Electronics and Computer Science
University of Southampton, UK
http://www.ecs.soton.ac.uk/
xwang@soton.ac.uk
t.tiropanis@southampton.ac.uk
[2] Information Technologies Institute
Center for Research and Technology Hellas (CERTH), Greece
http://www.iti.gr/
thanasis.papaioannou@iti.gr
[3] Nexa Center for Internet & Society
Politecnico di Torino, Italy
http://nexa.polito.it/
federico.morando@polito.it

Abstract. To explore the socio-technical aspects of the Internet requires infrastructures to properly foster interdisciplinary work and the development of appropriate research methods. To this end we present a platform called EINS Evidence Base (EINS-EB) which is developed as part of the EINS project. The EINS-EB also aims to empower researchers, academics, organisations and society to engage with Internet Science research independent of background. Currently, it provides for the collection and discovery of data resources, of analytic and simulation tools, and, in the future, of the methodologies behind those tools an of relevant scholarly activity. We explore issues of data representation, dataset description, dataset catalogues and method catalogues for Internet Science. The evidence base adopts semantic technologies to provide an interoperable catalogue of online resources related to Internet science. We also present activities on making the evidence base interoperable with related e-Science activities by communities engaging in relevant interdisciplinary collaboration.

Keywords: evidence base, Schema.org, web observatory, semantic catalogue

1 Introduction

The Internet Science community has been actively engaged in intensive research with respect to networking parameters issues to performance, QoS, security and

availability guarantees. As this discipline evolves, Internet Science has become an "inter-discipline" that draws on disciplines such as computer science, economics, sociology and law, and requires to go beyond the networking techniques to the exploration of correlations between the evolution of the Internet infrastructure and of the societal and business sectors of activity that increasingly rely on it. To this end there is demand to set up e-Science infrastructures that are appropriate to foster research collaboration among disciplines, empowering the communities involved with access to methods, data resources, data collection tools, analytic tools, and simulation tools. Many of those tools are available but they scattered in different repositories maintained by different communities making it hard to discover and use them especially by members of different disciplines.

An appropriate representation of resources in the repository is essential in order to transform mere data into information about the structure, function and performance of the system. In addition, the ability to capture all (potentially) relevant aspects of the broader context, repeat an experiment, reuse the same data and combine various types of information for analytic purposes is vital to establishing and communicating the robustness, generalisability and implications of particular findings in line with accepted scientific methods. On the other hand, an expressive representation usually comes at the cost of complexity. A main challenge is to adopt and extend a representation that is expressive enough to capture relevant characteristics of resources without introducing too much complexity. We decided to use the Schema.org vocabulary which is a lightweight, flexible and extensible vocabulary that is supported by main search engines. We adopt and extend the Schema.org vocabulary and build a Internet Science resource repository called the EINS Evidence Base (EINS-EB). Furthermore we deploy semantic enrichment technologies to assist users to record rich metadata of published resources with minimum efforts.

The remain of this paper is organised as following: we provide an overview of the EINS-EB in Section 2; we discuss the design concerns and describe the techniques of the EINS-EB in Section 3; details of data hosting and the infrastructure of EINS-EB are given in Section 4; we discuss the relationships and differences between the evidence base and several related platforms in Section 5, and the conclusion and future plan in Section 6.

2 Overview of the EINS-EB

The aim of the evidence base is to foster interdisciplinary work by creating an online catalogue that will record and expose detailed metadata of existing selected datasets, methodologies and tools. Interoperability of the representation of catalogued resources is essential for the evidence base to be used by researchers form different disciplines, and a well designed community engagement mechanism is crucial for gathering as many as possible resources that are related to Internet science. With those requirements in mind, the EINS-EB is built to enable users to conveniently publish and share Internet science resources with rich semantics.

In the EINS-EB, there is emphasis on open datasets (as shown in Figure 1) to ensure engagement with the wider community around Internet science. Apart from datasets, the online resource will catalogue related tools that are required to collect, analyse, visualise data, and e-infrastructures which are necessary to carry out Internet science experiments in controlled environments.

For each resource catalogued in the EINS-EB, general information such as title, description, keywords etc. are recorded to enable prompt searching. Based on the textual description of each resource, a DBpedia[4] classification is automatically generated by the system (and can be further specified or corrected by the publisher). The evidence base does not facilitate direct access to listed resources, but it requires publishers to provide URLs from where resources are available, along with licensing information. All the above mentioned information is embedded as Microdata in the web pages using Schema.org vocabularies. Notice that pages are marked up using the simple Microdata syntax and the Schema.org vocabulary, as explained below; nevertheless, a mapping from Schema.org to RDF (expressed in RDF Schema) is available and applications can use services to obtain a Linked Data representation of the Microdata. This makes our approach simple and pragmatic, yet essentially compatible with the more ambitious efforts to build the Web of Data.

Name	Licensing	Keywords	DBpedia	Description
Stanford Large Network Dataset Collection	Open	social networks, ground-truth community networks, communication networks	Social Networks	This is a collection of datasets (collected by the...
Eurosys '09	Proprietary	Communications, Facebook, Social graph	Social_networks	Datasets containing Facebook social graph (friends...
WOSN '09	Proprietary	Social networking service, Facebook, Social network	Social_networks, Social_graph	Dataset containing Facebook social graph (friendsh...
MPI-SWS Datasets	Proprietary	Social network graph, Social interactions	Social_networks, Social_graph	Social network datasets (Flickr, LiveJournal, Orku...
M-Lab data	CC0	Net neutrality, traffic discrimination , network performance	Network performance	Measurement Lab (M-Lab) is an open, distributed se...

Fig. 1. The dataset view of the EINS evidence base.

It is possible for anyone to create an account and share resources on the evidence base. Listing a resource can be easily done by filling of form containing required information as stated above. To reduce user efforts, the evidence base provides a list of common licenses from which users can choose the appropriate ones. The evidence base also automatically analyses the descriptions and if possible the "About" page of published resources to provide suggestions of

[4] www.dbpedia.org/

DBpedia classification. More of this automatically classification is given later in Section 3.3. User account information is used to fill in publisher information of resources.

3 Cataloguing Resources with Rich Semantics

To provide a semantic-rich catalogue for Internet science related resources a key factor is to have the appropriate vocabulary to express the metadata of resources. Datasets take a significant proportion of these resources and therefore the choice of vocabulary is biased to capture as many as possible characteristics of datasets. At the same time, the chosen vocabulary also needs the ability to be extended to cover other resources such as tools and e-infrastructures.

The expressivity of a vocabulary usually comes at the cost of complexity. To reduce user efforts of providing accurate metadata, we employ a service called TellMeFirst which assists users with candidate classifications of resources by analysing their descriptions.

3.1 Microdata

An emerging approach supported by the dominant search providers is to use Microdata (http://schema.org) markup and vocabularies to describe Internet Science datasets available online. Many sites are generated from structured data, which is often stored in databases. When this data is formatted into HTML, it becomes very difficult to recover the original structured data. Schema.org is a collection of schemas that webmasters can use to markup HTML pages to describe the data structure in ways recognized by major search providers, and that can also be used for structured data interoperability (e.g. in JSON). Search engines that support Microdata including Bing, Google, Yahoo! and Yandex. On-page markup enables search engines to understand the semantics of the information on web pages and provide richer search results in order to make it easier for users to find relevant information on the web. Markup can also enable new tools and applications that make use of the structure.

Microdata is a simple semantic markup scheme that is an alternative to RDFa and it has been developed by WHATWG.The Microdata effort has two parts: markup and a set of vocabularies. The vocabularies are controlled and hosted at Schema.org. The markup is similar to RDFa in that it provides a way to identify subjects, types, properties and objects. The sanctioned vocabularies are found at Schema.org and include a small number of very useful ones: people, movies, etc. When a taxonomy for the description of the various properties of online resources is missing, then one can use DBpedia (http://en.wikipedia.org/wiki/Wikipedia:Quick_cat_index) and other widely adopted taxonomies.

The Microdata markup consists of three basic tags: itemscope, itemtype, itemprop. An itemscope attribute identifies a content subtree that is the subject about which we want to say something. The itemtype attribute specifies the

subjects type. An itemprop attribute gives a property of that type. As an example, observe the embedded tags in the HTML markup of Figure 2 describing a dataset collection.

```
<div itemscope itemtype="http://schema.org/Dataset>
  <a href="http://snap.stanford.edu/data/" itemprop="name">
    Stanford Large Network Dataset Collection</a>
  <meta itemprop="http://schema.org/url"
    content="http://snap.stanford.edu/data/">
</div>
```

Fig. 2. An example of Microdata using a Schema.org vocabulary. It describes a dataset with the name *Stanford Large Network Dataset Collection* and the URL *http://snap.stanford.edu/data/*.

3.2 The Schema of the Evidence Base

Regarding online resource description, we employed Schema.org, as it is a solution capable of offering high searchability and simplicity. We chose to use at least those properties and vocabularies from Schema.org corresponding to the ones specified in the standard Dublin Core (http://dublincore.org/documents/dces/). We briefly describe the main choices regarding our schemas for the online datasets, online tools and e-infrastructures below.

Dataset We partially adopted the type `Thing::CreativeWork::Dataset` from the Schema.org vocabularies for describing the various datasets available online. This type inherits some interesting attributes (named "Properties") from the `Thing` type that can be utilized for describing the various Internet Science datasets found, depicted in Table 1.

Tool and e-Infrastructure Schema We employed a subset of the type `Thing::CreativeWork::SoftwareApplication` from the Schema.org vocabularies for describing the various Internet tools and eInfrastructures available online.

We again employ here the same attributes inherited from the types `Thing` and `CreativeWork` that were employed for the description of an online dataset, described in Table 1. In Table 2, we briefly describe additional attributes from types `CreativeWork` and `SoftwareApplication` that are employed in the schema of the online tools and e-Infrastructures.

3.3 Description-Based Automated Classification

TellMeFirst[5] [2] is a tool for classifying and enriching textual documents via Linked Open Data. TellMeFirst leverages natural language processing and Se-

[5] http://tellmefirst.polito.it/

Attribute	**Description**
(*inherited from type* `Thing`)	
description	Text description of the dataset at the original site.
sameAs	URL link to the original site of the dataset or the wikipedia entry describing the datasets nature.
url	The URL of the dataset.
additionalType	Categorizes the type of the dataset using alternative vocabularies or taxonomies than Schema.org ones, e.g. typeof Dbpedia categories, etc.
(*inherited from type* `CreativeWork`)	
author	It may coincide with the "creator" attribute below, and refers to the creator of the dataset.
copyrightHolder	It refers to the license of the dataset.
copyrightYear	The year of the license.
datePublished	The date that the dataset became available online.
keywords	keywords describing the dataset. These are many times given in the web page of the dataset, but more "standards" classification keywords, e.g., from ACM taxonomy http://www.acm.org/about/class/2012, can be used.
audience	The scientific community of the dataset.
creator	It may coincide with the "author" attribute above.
dateCreated	The date of the dataset creation.
dateModified	The date of the dataset update.
version	This attribute can be used in case that there are multiple versions of the dataset available.
(*inherited from type* `Dataset`)	
catalog	A data catalog which contains a dataset.
distribution	A downloadable form of this dataset, at a specific location, in a specific format.
spatial	The range of spatial applicability of a dataset, e.g., for a dataset on EU demographics, EU.
temporal	The range of temporal applicability of a dataset.

Table 1. The Dataset schema.

Attribute	**Description**
(*attributes inherited from types* `Thing` *and* `CreativeWork` *listed in Table 1*)	
...	
(*additional attributes inherited from type* `CreativeWork`)	
audience	The intended audience of the item, i.e. the group for whom the item was created.
citation	A citation or reference to another creative work, such as another publication, web page, scholarly article, etc.
contributor	A secondary contributor to the CreativeWork.
provider	The organization or agency that is providing the service.
sourceOrganization	The Organization on whose behalf the creator was working.
version	The version of the CreativeWork embodied by a specified resource.
(*inherited from type* `SoftwareApplication`)	
applicationCategory	Type of software application, e.g. "Traffic Generator, Network Simulator".
downloadUrl	If the file can be downloaded, URL to download the binary.
featureList	Features or modules provided by this application (and possibly required by other applications).
fileFormat	MIME format of the binary (e.g. application/zip).
fileSize	Size of the application / package (e.g. 18MB). In the absence of a unit (MB, KB etc.), KB will be assumed.
installUrl	URL at which the app may be installed, if different from the URL of the item.
memoryRequirements	Minimum memory requirements.
operatingSystem	Operating systems supported (Windows 7, OSX 10.6, Android 1.6).
processorRequirements	Processor architecture required to run the application (e.g. IA64).
releaseNotes	Description of what changed in this version.
requirements	Component dependency requirements for the tool.
softwareVersion	Version of the software instance.

Table 2. The schema for online tools and eInfrastructures.

mantic Web technologies to extract main topics from texts in the form of DBpedia resources. Input texts may then be enhanced with new information and contents retrieved from the Web (images, videos, maps, news) concerning those topics.

3.4 Linking to Other Semantic Catalogues

Using TellMeFirst, the resources described in the evidence base may be tagged using a very broad vocabulary (consisting of the more than 4.6 million of entries of the English Wikipedia), which is also structured by the Wikipedia community, so that entities (i.e., pages) are nested within categories, which are in turn sitting within a three of higher level categories.

The link with Wikipedia/DBpedia is also a gateway to the Web of Data: in fact, DBpedia is at the centre of the Linked Open Data Cloud[6] and linking to this core resource indirectly generates semantic relations with many other resources.

Furthermore, the Schema.org vocabulary used by the EINS-EB is compatible with many other vocabularies, and it is straightforward to import metadata from other semantic catalogues that use one of the compatible vocabularies. For example, the Southampton University Web Observatory (SUWO) [5,4] also adopts Schema.org. EINS-EB and SUWO can crawl each other's pages and list each other's resources. CKAN[7] provides DCAT documents about datasets which are also compatible with the Schema.org vocabulary used in the EINS-EB, and also can be imported to the EINS-EB. This interoperability virtually leads to a global network of semantic catalogues and enables users to search with rich semantics resources from any catalogue in the network.

4 Data Hosting and Infrastructure

A complementary service provided by the evidence base is data hosting. The hosting service (as shown in Figure 3) is supported by various infrastructures such as Hadoop Distributed File System (HDFS), MongoDB, SQL database etc. Users can upload data into any of supported databases and provide a link on the EINS-EB. The hosting service provides extra flexibility for sharing and reusing data. For example, the Neubot[8] data consist of a large number of files which can only be analysed after downloading them. By utilising the hosting service the Neubot data are also served by MongoDB at the University of Southampton from where users can query the data online.

[6] http://lod-cloud.net/
[7] http://ckan.org/
[8] http://neubot.org/

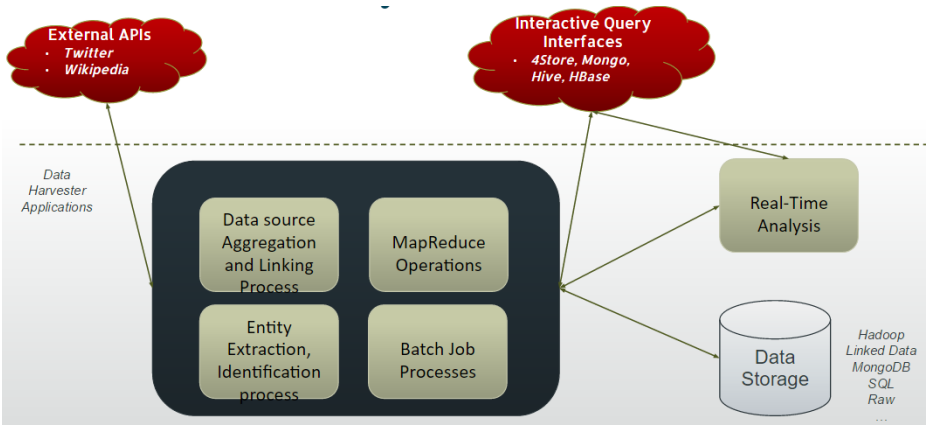

Fig. 3. Data hosting infrastructure of the EINS evidence base.

5 Related Work

The EINS-EB is aiming at harmonisation with a number of similar infrastructures and tools in the areas of measurement collection, analytics and data repositories. This section outlines those related efforts and how the EINS-EB aims to interoperate with them.

Measurement Lab (M-Lab)[9] is the largest repository of open Internet performance data and has been backed many Internet measurement publications (e.g. [3,1]). It hosts a large collection of open source Internet measurement tools, data collected by those tools and visualisations based on the data. Comparing to M-Lab, the EISN evidence base focuses on providing rich metadata (via Schema.org vocabularies and TellMeFirst service) for registered resources and improving interoperability to other disciplines. M-Lab resources can be listed on the evidence base to gain better discoverability.

Southampton University Web Observatory (SUWO) [5,4] is a portal for datasets and analystics. It provides metadata of registered resources as well as facilitates access to those resources in a secure way. Resources listed on SUWO can be private and permission has to be required before accessing them. SUWO also gives Schema.org Microdata, and thus the EINS-EB can import resources from SUWO, and add extra classification information using TellMeFirst.

DatCat[10] is an online catalogue of datasets regarding Internet measurements. They employ a proprietary list of object types to describe the data collections, having each object comprising multiple fields, and maintain indices that allow advanced data querying through their web site. However, their datasets are not searchable through search engines, as opposed to our approach due to Schema.org

[9] http://www.measurementlab.net/about
[10] http://imdc.datcat.org

Microdata descriptions. Moreover, their vocabulary does not adhere or link to any standard for describing data semantics.

CKAN[11] is an open data catalogue platform widely used by several data hubs. CKAN focuses on publishing metadata of datasets to increase discoverability, while the EINS-EB fosters not only datasets but also other resources that are related to Internet science, such as tools and e-infrastructures. CKAN provides a Linked Data presentation of the metadata it catalogued while the evidence base utilises Schema.org Microdata, which is better recognised by major web search engines such as Google, Bing, Yahoo!, and Yandex. Furthermore, the EINS-EB is assisted by the TellMeFirst service to automatically classify registered resources.

6 Conclusions and Future Plan

The EINS Evidence Base (EINS-EB) is fostering interdisciplinary research for the Internet Science community by providing, at first instance, a way to catalogue and discoverer related data resources and analysis or simulation tools available online. Requirements for harmonisation with existing repositories have led us to technological choices for resource description (Schema.org) and a number of utilities based on linked data and content analysis.

Leveraging the fact that Schema.org Microdata can be easily represented as Linked Data expressed with the RDF formalism (as mentioned above), we could couple the existing Microdata syntax with a formal and explicit RDF representation. This can easily be done within the HTML of the evidence base pages, using the RDFa serialisation, but with some additional effort dereferenceable IRIs can also be provided, as well as a SPARQL end-point. In this way, our initial pragmatic (and general purpose search-engine oriented) approach can be maintained, while the interoperability with Linked Open Data catalogues, including CKAN, may be increased. Note that TellMeFirst related data are already expressed as DBpedia resources, so connecting these to the Linked Open Data cloud would be trivial.

References

1. Basso, S., Meo, M., Martin, J.C.D.: Strengthening Measurements from the Edges: Application-level Packet Loss Rate Estimation. In: ACM SIGCOMM Computer Communication Review 2013 (2013)
2. Futia, G., Cairo, F., Morando, F., Leschiutta, L.: Exploiting Linked Data and Natural Language Processing for the Classification of Political Speech. In: International Conference for E-Democracy and Open Government 2014 (2014)
3. Masala, E., Servetti, A., Basso, S., Martin, J.C.D.: Challenges and Issues on Collecting and Analyzing Large Volumes of Network Data Measurements. In: New Trends in Databases and Information Systems 2014 (2014)
4. Tiropanis, T., Hall, W., Hendler, J., de Larrinaga, C.: The Web Observatory: A Middle Layer for Broad Data. In: Big Data. pp. 129–133 (2014)

[11] http://ckan.org/

5. Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., Hendler, J.: The Web Science Observatory. IEEE Intelligent Systems 28(2), 100–104 (Mar 2013), http://eprints.soton.ac.uk/354604/1/TheWebScienceObservatory-postprint.pdfhttp://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6547975