Semantic Web 0 (2012) 1
IOS Press

Exposing the Institute for Development Studies' Data using the API2LOD Linked Data Wrapper

Christophe Guéret [a], Victor de Boer [a], Duncan Edwards [b], Timothy G. Davies [c]
[a] *The Network Institute, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands*
E-mail: {c.d.m.gueret,v.de.boer}@vu.nl
[b] *Institute of Development Studies, Brighton, UK*
E-mail: d.edwards@ids.ac.uk
[c] *Web Science Doctoral Training Centre, University of Southampton, UK*
E-mail: tim.davies@soton.ac.uk

Abstract. This short paper provides a description of the wrapper (API2LOD) used to expose the data from the Institute for Development Studies' Knowledge Services as Linked Data. The wrapper provides Linked Data access to 35,000 research documents on development research as well as its metadata. Links are added from this metadata to a number of external sources: DBpedia, GeoNames and Lexvo. We expect that the IDS data will play a central role in the larger web of Linked Data for global development.

Keywords: international development, global change, knowledge sharing, data wrapper

1. Introduction

It has been recognized (*c.f.* [3]) that development knowledge is a global public good that belongs to everyone, and from which everyone should benefit. Knowledge sharing is therefore an important issue in the field of international development.

The Institute of Development Studies (IDS) is a leading global charity for international development research, teaching and communications. IDS was founded in 1966 and enjoys an international reputation based on the quality of its work and its commitment to applying academic skills to real world challenges. Its purpose is to understand and explain the world, and to try to change it - to influence as well as to inform.

IDS is a pioneer in development communications hosting a range of innovative and highly regarded knowledge services - including Eldis, BRIDGE, and the British Library for Development Studies[1]. These services seek to mobilise knowledge to support more informed decision-making by those in a position to influence change. This is based on the belief that decision-making is strengthened when it is underpinned by timely and relevant information that reflects a diversity of viewpoints. The Eldis and BRIDGE services put an emphasis on sourcing research from "Southern" organisations to attempt to balance the domination of research originating from "Northern" organisations to provide this diversity of perspectives on development issues.

This short paper provides a description of the wrapper used to expose the data from the IDS Knowledge Services as Linked Data [1] and the resulting data set. We first introduce the data provided by the Knowledge

[1]The portal for the IDS knowledge services can be found at http://www.ids.ac.uk/go/ids-knowledge-services

Services Open API and then describe how it is exposed as Linked Data. Then, we locate this data set within the bigger picture of Linked Data for development.

2. Data from the Knowledge Services Open API

The Knowledge Services Open API provides easy programmatic access to over 35,000 of thematically organised research documents that are freely available online, and over 8,400 metadata records about organisations working in development (*c.f.* Table 1 for an overview of the data served). This data is published under the Creative Commons BY Attribution licence and is therefore free to distribute, remix, and utilise to build applications and services.

The data is accessed through an API that offers a variety of search and identifier based queries. One can look for an entity in particular knowing its unique identifier or perform a topical search around one or more keywords. The Knowledge Services API is available at `http://api.ids.ac.uk/` and is free to use by anyone having registered on the portal.

To expose this API data as Linked Data, we decided to make use of the identifier based services. This allows for a straightforward mapping between an identifier based URI scheme and the relevant API call (see Section 3.1 for more information on the URI scheme we adopted). These API calls are issued by a wrapper which is a service developed separately from the Knowledge Services API and the associated browsing interface. This separation is a design choice allowing for more flexibility in the development and deployment of the Linked Data service.

In the following, we describe the technical details about this wrapper and the content of the exposed Linked Data.

3. The wrapper "API2LOD"

To expose the IDS data as Linked Data, we created a wrapper based on the Java restlet technology[2]. The wrapper is deployed on Google AppEngine at `http://api2lod.appspot.com`, and its code is freely available on GitHub at `https://github.com/cgueret/LinkedIDS`. It is meant to be used as is but can also be deployed locally on any kind of servlet container. It is also a generic wrapper that can be used to expose several data sets. The reasons for which we decided to design a wrapper are:

- The possibility to more easily push new linkers services by deploying them on the service side;
- Entities retrieved by the linker services are cached across the different data sets being exposed by the service;
- Data set publishers do not have to worry about the technical specificities of RDF publishing and can focus on the design of their API.

Despite the genericity of the tool, we hereafter focus on describing its usage for the IDS data set.

The data available from the IDS API is divided into different types of entities grouped in collections. The wrapper defines URIs that directly map a resource to its collection, its type (see Table 1) and its identifier. The wrapper exposes these different entities on demand. Upon a HTTP call to a IDS URI, the wrapper fetches information from multiple sources:

1. First, the wrapper calls the IDS REST API to fetch all known information about the resource (document, organisation, theme, etc.). The JSON key-value pairs are translated into RDF predicates and objects that combined with the resource URI form the triples. The mapping between JSON keys to predicates is based on a hand-written RDF schema embedded with the wrapper code. This schema also lists the `rdf:range` of the predicate, allowing specific values to be converted to typed literals or resources.
2. Secondly, the wrapper enriches the translated data through a number of custom *linkers*. These linkers take a key-value pair and based on those establish links to internal or external resources. We describe the linkers used in Section 3.3.

In the following, we describe the URI scheme used and how it is mapped to the queries sent to the data API. We also explain the vocabulary used and how connections are established within the data set and with external data sets.

3.1. URI scheme

Every entity within the data set has a unique identifier assigned to it. The wrapper uses this identifier in a generic URI scheme based on the name of the collection, the type of the entity and the value for the identifier:

```
eldis-ids:resource/<type>/<id>
```

[2] `http://www.restlet.org/`

Type	Number	Description
Documents	>35,000	Describe research publications, including title, description, publisher information, thematic, region, and country categorisation, and a Url to the full text of the document
Organisations	>8,400	Describe development organisations including, name, Url, description, thematic, regional, and country categorisation - types of organisations include publishers, funding bodies, NGOs, INGOS, UN agencies
Themes	>1,050	Thematic categories - hierarchical category
Countries	244	Details of countries, including Name, ISO codes - used to describe research country of focus, country of publication, organisation location
Regions	10	Details of Regions - used to describe research region of focus

Table 1

List of entities documents with the IDS data base

In this URI pattern, "eldis-ids" is used for `http://api2lod.appspot.com/eldis-ids`. This address corresponds to the name of the server hosting the wrapper along with the name of the wrapped data set.

For instance, the following resource points to the description of "Gambia", which is a country within the collection Eldis:

```
eldis-ids:resource/Country/A1078
```

3.2. Vocabulary used

The data exposed by the data API is made of a collection of fields/values. By default, all the fields are turned into URIs which are accessible under the namespace `eldis-ids:vocabulary#`. We established mappings for some of them to common vocabularies, for example to Dublin Core for document metadata, FOAF for personal data and SKOS for thesaurus descriptions. We decided to keep the denomination from the data API to allow for a smoother transition for users currently using the API and willing to use the RDF data. These users still see the terms they are used to.

However, some fields are not used and are directly replaced by terms from popular vocabularies. For example, `eldis-ids:vocabulary#object_type` is replaced by `rdf:type`. Not doing so would have had a significant negative impact on the usability of the data set, as everyone expects resources to have an `rdf:type`.

3.3. Links to external resources

When building the RDF description of an entity, the wrapper calls a set of linkage services to find external and internal entities to re-use. There are currently linker services available for Lexvo, DBpedia and GeoNames.We hereafter describe them and give an indication of their precision and recall.

3.3.1. Lexvo

This linker allows to replace a language in full text by its resource on the Lexvo data set. For instance, this linker can be used to replace the string "English" by the resource `http://lexvo.org/id/iso639-3/eng`.

Implementation The search API provided by Lexvo is used to get a pointer to a page by looking up the language name in Lexvo. This page is de-referenced as RDF and the name of the entity to link to is search for within the description. In this way, a document description is enriched with Lexvo language information.

Efficiency Since the IDS API does not give us an opportunity to list all used Language strings, we evaluated the Lexvo mapper on a list of the 100 most used world languages [3]. On this list of languages (which does not contain any typos or misspellings) the linker performed without error, resulting in a recall and precision of 1.0.

3.3.2. DBpedia

This linker makes use of the public SPARQL end point provided by DBpedia to find an entity of a particular type, having a indicated name. For instance, replace the indication of the theme "Food security" by its resource `http://dbpedia.org/resource/Food_security`.

Implementation The full name of the entity is sought and it's type, manually mapped to DBpedia equivalent, is indicated to better filter the results.

Efficiency Analysis of the DBpedia linker shows that for 307 out of the 1063 Eldis themes (29%), one or more links to DBPedia were found. On average, each linked concept was linked to 2.52 DBPedia resources.

[3] retrieved from `http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers`

We manually evaluated the links (or lack thereof) for a random sample (size 40) of Eldis themes. Out of the unlinked eldis, for 11% of the themes, we could find a matching DBPedia concept. Most of these matches were not found by our linker because of missing words (for example the Eldis theme "climate change africa" versus the DBPedia concept with the title "Climate Change in Africa". For the remaining 89%, no matching DBPedia concept could be found by hand. Most of these unmatchable themes are either broader or narrower than DBPedia concepts (for example the Eldis theme "International policy and aid financing" which combines two DBPedia concepts). More sophisticated mapping algorithms could be used to establish more (SKOS) links. Evaluation of the found links resulted in a precision of 0.72. Most incorrect matches occur when multiple links are found (for example, the Eldis theme "Corruption" is linked to the DBPedia concepts "Corruption", which is correct and "Corruption (Videogame)", which is incorrect). A filtering on concept type could further improve the precision. 92% of the linked concepts have at least one correct link.

3.3.3. GeoNames

The search API provided by GeoNames is used to fetch the identifier of the resource to link to. Using this linker, "Gambia" can be linked to `http://sws.geonames.org/2413451`.

Implementation For this query, the two letters code of the country found in the IDS data is used together with the full name. The result is turned into the matching RDF resource on GeoNames by using their URI scheme `http://sws.geonames.org/<id>`.

Efficiency Analysis of the GeoNames linker shows that for 187 out of the 244 Eldis countries, a GeoNames match to an Independent Political Entity is found, resulting in a recall of 0.77. Since we are only looking for countries that are recognized in GeoNames as independent political entities, the unlinked entities we miss are not of that type, but refer to countries with a different legal status (for example "Aruba"). Manual evaluation of a sample of the matches shows a precision of 100%.

3.3.4. Internal links

The data API provides relations between the documents/organisations and the themes/countries/regions they relate to. For instance, a given document will contain material about a particular region. We use the information provided by the API to generate internal links within the wrapped RDF data set.

3.4. Known limitations

There are some limitations directly related to the usage of a wrapper to expose the data.

3.4.1. No data export

The wrapper does not offer a dump of the data set it exposes, nor does it provide a SPARQL end point. Resource descriptions are accessible only through dereferencing their URI asking for RDF. In order to help finding a particular resource, it is possible to wrap the search service of the API, when available, to make it usable through the data set main page. The search interface for the IDS data is available on the homepage of the data set at `http://api2lod.appspot.com/eldis-ids`.

3.4.2. Warmup time

The linkage with external entities is done on demand and cached. Therefore, when a given entity is dereferenced for the first time in the wrapper, some seconds are spent querying the external services for links.

4. Example of usage

Use of linked data in the international development field is still nascent, but as [2] explores, knowledge intermediaries are starting to explore the opportunities to use linked data to connect across datasets, and demonstrator projects have proven vital in advancing understanding of how organisations can better collaborate through linked data. A number of hack days, including the 2012 Development Data Challenge in Helsinki, and January 2013 hackathon on agriculture and nutrition data at the iHub in Nairobi have drawn upon the IDS Linked Data Wrapper [4].

We are currently developing a client application that exploits the IDS Linked Data for explorative document browsing. In this application, a user can browse through the IDS documents and is shown the IDS metadata (theme, organisation, etc.). In addition to this, snippets and links to documents as well as images from external sources, that are linked to the document are presented to the user. These resources are for example linked via the theme of the object.

An example of such an enrichment path is a document on "Assessing Climate Change Vulnerability in East Africa" (with the URI `eldis-ids:resource/Document/A60737`) which is linked

[4] `http://bit.ly/X4YaKL`

to the term "Climate Change" (available at eldis-ids:resource/Theme/C308). The DBPedia linker establishes a relation to dbpedia:Climate_change, which in turn links to a number of people involved in this field (as well as to pictures of these people), publications written about the subject as well as other resources. The application will display a selection of these resources related by the DBPedia or IATI linker. At the same time, the information linked via the GeoNames linker will be used to -for example- plot IDS documents or organisations on a map.

Explicit links from IDS vocabularies to Linked Datasets make it easier for developers to build such mash-up applications. Furthermore, the fact that we reuse popular schemas such as DC or SKOS allow for applications that can browse and display these properties to use the IDS data easily. For example, generic SKOS Linked Data browsers can be used to browse the IDS thesaurus hierarchy.

5. The bigger picture

Mobilising knowledge is vital for effective international development. The IDS Eldis database is just one of a number of well curated publication repositories focussed on development, and increasingly the owners of these repositories are looking to expose their catalogues as linked data. For example, the Research for Development (R4D) catalogue of research funded by the UK Department for International Development (DFID) (http://www.dfid.co.uk/r4d/) has added an RDF export profile to it's Open Archive Initiative (OAI) interface, and is making a regular RDF dump available at http://dfid.gov.uk/r4d/rdf/R4DOutputsData.zip using a model based on the Food and Agriculture Organisation's (FAO) LODE-BD guidance (http://aims.fao.org/lode/bd). Even though most publication repositories use different taxonomies in their metadata, the common use of links to DBpedia when publishing linked data provides increased opportunities to provide search and retrieval across datasets, or to enhance searches on one dataset with information from another. For development information specifically relating to agriculture, the extensive FAO AGROVOC multilingual thesaurus, available as linked data (and already mapped to a range of other sources, including DBpedia) provides additional connections that can be exploited by applications to provide more advanced search and retrieval of relevant information.

Making knowledge accessible to development practitioners is not just about providing better search: it can also involve proactively integrating knowledge resources into practice environments. The International Aid Transparency Initiative (IATI)[5] has developed an XML standard for the publication of information on development activities and aid flows, and a linked data vocabulary, drawing on DC, SKOS and VCARD is currently in development, with proposed linkages to GeoNames, and the development of a project to map a number of key development project taxonomies together. A pilot project by Research for Development has linked R4D publication records to the aid project that funded them, exposing this in a web widget at http://r4d.herokuapp.com. Extending this concept, so that all research relevant to a particular project (particularly early stage projects - as IATI seeks publication of planned as well as in progress project information) based on geographical and thematic linkages, as well as funding, could be displayed has the potential to increase the use of past evidence in project planning and evaluation.

6. Sustainability and conclusions

In this paper we provided a description of a wrapper used to expose the data from the Institute for Development Studies Knowledge Services API. This service gives access to Linked Data descriptions of 35,000 research documents on development research. API2LOD currently links these RDF descriptions to a number of external sources including DBpedia, GeoNames and Lexvo.

This wrapper is a generic service that can be used to expose several data sets. Because of the use made of the wrapper within the institute and outside of it, IDS will contribute to support the costs for keeping the wrapper running and ensure the software is maintained.

We are currently working on exposing the data from the Openaid IATI Parser and API (OIPA) [6]. Linkers between the IDS and this OIPA dataset have already been developed and a first version of the OIPA Linked Data set is deployed at http://api2lod.appspot.com. We expect that the IDS data will play a central role in the larger web of Linked Data for global development. Future work include the creation of links

[5] http://www.aidtransparency.net/
[6] http://oipa.openaidsearch.org/api/v2/docs/

to more external sources, in particular other development related data sets. We are also busy with a number of client applications that are either planned or being actively developed at the moment.

References

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

[2] Timothy G. Davies and Duncan Edwards. Emerging implications of open and linked data for knowledge sharing in development. *The IDS Bulletin*, 43(5):117–127, September 2012.

[3] Worldbank. Sharing knowledge to achieve development goals. Précis report 234, Worldbank, 2003. http://bit.ly/JrUosl.