Proceedings of the first workshop on DownScaling the Semantic Web (DownScale2012)

Hosted at ESWC2012, May 28, Heraklion, Crete, Greece

Editors
Christophe Guéret
Stefan Schlobach
Florent Pigout

Program Committee

Anna Bon	VU University Amsterdam
Stephane Boyera	World Wide Web Foundation
Philippe Cudré-Mauroux	University of Fribourg
Mathieu D'Aquin	Knowledge Media Institute, the Open University
Jérôme David	INRIA Rhône-Alpes
Victor De Boer	VU Amsterdam
Christophe Guéret	Vrije Universiteit Amsterdam
Florent Pigout	OLPC France
Mike Powell	IKM Emergent
Louiqa Raschid	University of Maryland
Laurens Rietveld	Vrije Universiteit Amsterdam
Stefan Schlobach	Vrije Universiteit Amsterdam
Elena Simperl	KIT
Andreas Thor	Department of Computer Science, University of Leipzig
Sameer Verma	San Francisco State University
Maria Esther Vidal	Universidad Simon Bolivar, Dept. Computer Science

Table of Contents

Tackling a DISASTER using semantic technologies

Guillermo González-Moriyón[1], Emilio Rubiera[1], Marcos Sacristán[2], and Javier Collado[2]

[1] Fundación CTIC
Gijón, Asturias, Spain
{name.surname}@fundacionctic.org
http://www.fundacionctic.org/
[2] Treelogic
Llanera, Asturias, Spain
{name.surname}@treelogic.com
http://www.treelogic.com/

Abstract. In the event of a disaster, coordination of emergency responders is challenging due to the diversity of support systems in use. Work is now being done to tackle the problem leveraging semantic technologies. The chosen approach focuses on the interoperability between Emergency Management Systems (EMSs) via data mediation and a reference ontology. This paper introduces the DISASTER FP7 project and outlines its main activities planned for the upcoming years.

Keywords: emergency management systems, ontology, data mediation, ontology modularization, data sharing

1 Introduction

Near 100,000 people died in Europe in the first decade of this century due to natural and industrial disasters [6]. Transport accidents and terrorism also add casualities to this figure. Moreover, material losses ascend to 150 billion Euro from natural disasters alone in the same decade. Disasters such as L'Aquila, Eyjafjallajökull, Prestige and Chernobyl will remain carved in the collective European consciousness.

In the current era, developed countries have a number of bodies that respond to emergencies in order to minimize casualities and economic loss. First responders include fire brigades, police, hospitals and military forces. An effective response depends on their ability to quickly take decisions based on accurate and reliable data. Emergency Management Systems (EMS) are information tools supporting resolutive actions against the clock.

Nowadays Europe showcases a well-spread heterogeneity of EMSs depending on the stakeholder "color" (emergency jargon to identify first responders) and the administrative division (city, region, etc.) where they operate. Political boundaries and compartimentation of functional competences are a hindrance

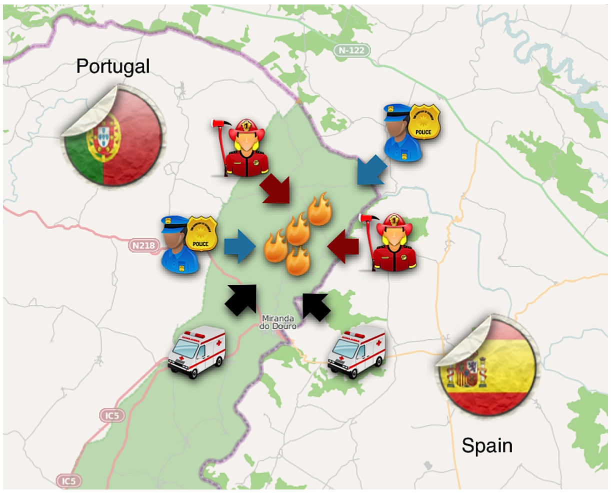


Fig. 1. First responders of different “colors” react to a forest fire on the border between two European countries.

for fluid communication and information exchange between EMSs. This fragmentation has evolved with time for cultural and historic reasons, and it is an unavoidable part of the European reality. The complexity of some emergency situations requires the participation of a variety of first responders who depend crucially on their ability to effectively exchange data (Figure 1).

2 Requirements for effective operational data exchange

At the event of a disaster many entities are involved. Some may be affected by the disaster, others may need to take contigence actions, while others may be accidental watchers of the incident. Thus, each entity can provide a partial version of the total picture. The objective is to make the most complete picture available to decision makers by putting together the partial pictures coming from the different parts involved. Due to the decentralization of the information scenario a number of conditions must be fulfilled:

– It must be possible to gather and to manage heterogeneous information from many available sources. This information includes an appropriate description of the disaster event in terms of both quality and quantity. In addition, the description of the scenario must be enriched with contextual and environmental data, e.g., infrastructures, populated areas, weather forecast... This harvest process often faces challenges due to the incompleteness and incorrectness of the data.
– The description of the scenario must be shared and agreed by the stakeholders in real time. A common picture facilitates team coordination and

synchronization of response operations, for instance, to know in real time which areas have been already evacuated.
- Information overload must be prevented. Decisions are difficult to make when dealing with an overwhelming volume of heterogeneous data. Filtering information in terms of relevance is crucial. Moreover, the concept of relevance is subjective: each actor is interested in a different subset of the information.
- Information pieces must be referenced with respect to both geographical and temporal coordinates. To gain an insight of the situation, it helps to have map interfaces and events displayed in sequence.
- Message oriented communications among different responders must be effective regardless of cultural and technical differences. Due to the diversity at each end of the channel, many problems arise including divergences of communication protocols, data formats or information models. In addition to the core message, the sender may decide to include additional information that enriches the message and increases its usefulness to the receiver.
- When integrating information from various sources and communication breakdown occurs, previously exchanged information might still be valuable. Not all data has the same expiration date: topographic information will still be valid, whereas current positioning of units deployed is dynamic and potentially untrustable afterwards.

The aforementioned conditions match some of the research topics tackled by the Semantic Web, e.g., information quality, data and model sharing, or data enrichment. These matchings suggest that a potential solution to the EMS interoperability problem may reuse the findings from the Semantic Web community. Moreover, the work on Semantic Web has demonstrated the value of adopting open standards to reuse vocabularies and exchange data between decoupled systems in the absence of a central authority.

3 A novel approach to disasters

DISASTER (Data Interoperability Solutions at STakeholders Emergencies Reaction) is a collaborative European project including software vendors, research organizations and specialists in emergency response. The goal of DISASTER is to ease communication among existing EMSs considering the requirements aforementioned.

DISASTER project respects EMSs diversity in Europe. No “one size fits all” central EMS or unique exchange language is proposed. Instead, DISASTER relies on a shared reference ontology as well as on mediation techniques. As depicted on Figure 2, semantic technologies will make possible for any EMS to exchange information and to query external sources including Geographical Information Systems and the Linked Open Data cloud, and of course, other EMSs.

When it comes to mutual understanding, the first step is to agree on shared baseline knowledge. Under our approach, this knowledge is modelled with ontologies. Ontologies have been proved to solve common understanding problems [2].

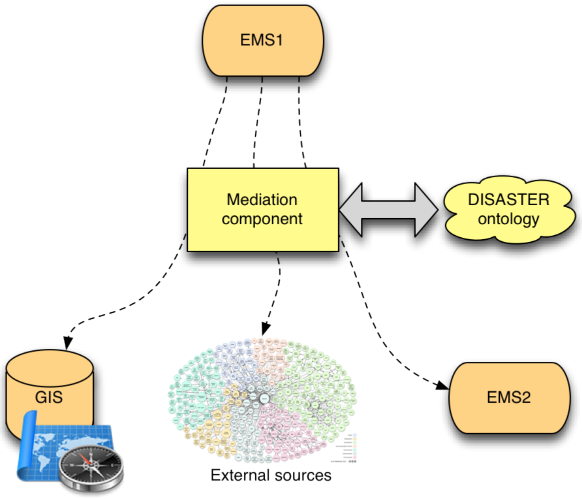


Fig. 2. DISASTER proposes data mediation based on a reference ontology to enable data exchange between existing EMSs and other systems.

Moreover, the use of ontologies to specify knowledge in the domain of emergencies have already been discussed in a number of studies [1, 3]. Despite this body of previous work, in 2011 the European Commission identified the need to develop an ontology shared by all stakeholders (FP7 research topic SEC-2011.5.3-2).

4 Prospective work

DISASTER has commenced in February 2012 and will run until 2015. The project has started by gathering requirements about different aspects: from first responders operative requirements to linguistic and cultural requirements.

DISASTER proposes data mediation to tackle communication problems among EMSs. Efforts are planned to bring to practice the theoretical results of the research community on data mediation. Actual interoperability among EMSs poses a number of challenges regarding several dimensions of the problem: communication protocols, data models and data formats. A workable solution typically involves all these aspects.

The main result of the project will be the DISASTER ontology. The creation of this ontology will follow a modular approach: the core module will be based on upper level ontologies such as DOLCE [4] or SUMO [5]. This core will be complemented with transversal modules giving support for representation of temporal and spatial descriptions. These vertical modules are vital to express contextual information, and offer a chance to plug-in existing domain ontologies and taxonomies. Finally, vertical modules will extend the base functionality to different domains at the levels of both disaster description and stakeholder resources. Although a modular design brings a number of difficulties at design time, this methodology allows further extension to fit specific scenarios in the future. It is expected that at the end of the DISASTER project, a stable version of the ontology (as universal as possible) will be produced, leaving it feasible to store a local copy for each emergency system. By means of this ontology, mediation can take place in order to consume external data before any connection breakdown and under offline circumstances.

The ontology will serve to combine heterogeneous information coming from diverse sources. The project will study how to coherently assemble an agreed disaster scenario description. Moreover, map-based visualization paradigms will be explored to merge context information with operational data and user-oriented mechanisms will be defined to filter relevant subsets of information. Devices supporting on site operations should be taken into account. Responders are often equipped with mobile devices providing access to the information. However, connectivity issues must be addressed as connection cannot be ensured.

Finally, in order to ensure the DISASTER solution is grounded on reality, validation will be carried out. At this early stage of the project the details of the final project scenario are still a draft. However, one of the possibilities involves a (simulated) disaster at a large international European airport with a transnational component featuring stakeholders of different kinds and nationalities.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 285069.

References

1. Z. Fan and S. Zlatanova. Exploring ontology potential in emergency management. *Proceedings of the Gi4DM conference geomatics for disaster management 2010*, 179(13):e18, 2010.
2. D. Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, volume 48. Springer, 2003.
3. X. Li, G. Liu, A. Ling, J. Zhan, N. An, L. Li, and Y. Sha. Building a Practical Ontology for Emergency Response Systems. *International Conference on Computer Science and Software Engineering*, pages 222–225, 2008.
4. C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology, 2002.
5. I. Niles and A. Pease. Towards a standard upper ontology. *Proceedings of the international conference on Formal Ontology in Information Systems FOIS 01*, 2001(3):2–9, 2001.
6. E. E. A. Technical. Mapping the impacts of natural hazards and technological accidents in Europe An overview of the last decade. Technical Report 13, European Environmental Agency (EEA), 2010.

Linked data from your pocket

Jérôme David, Jérôme Euzenat, Maria-Elena Roşoiu

INRIA & Pierre-Mendès-France University
Grenoble, France
{Jerome.David,Jerome.Euzenat,Maria.Rosoiu}@inria.fr

Abstract. The paper describes a lightweight general purpose RDF framework for Android. It allows to deal uniformly with RDF, whether it comes from the web or from applications inside the device. It extends the Android content provider framework and introduces a transparent URI dereferencing scheme allowing for exposing device content as linked data.

1 Introduction

Smartphones are becoming the main personal information repositories. Unfortunately, this information is stored in independent silos managed by applications and thus, it is difficult to share data across them. Nowadays, mobile operating systems, such as Android, deliver solutions in order to overcome this, but they are limited by the application database schemas that must be known beforehand.

The difficulty to share phone data at the web scale can be seen as another drawback. One can synchronize application data, such as the contacts or the agenda using a Google account. However, they are not generic solutions and there it is no mean to give access to data straight from the phone.

Our goal is to provide applications with a generic layer for data delivery in RDF. Using this solution, applications can exploit device information in an uniform way without knowing from the beginning application schemas. This information can also be exposed to the web and web information can be considered in the same uniform manner. Moreover, we propose to do it along the linked data principles (provide RDF, describe in ontologies, use URIs, links to other sources).

For example, in the future, this application could be used as a personal assistant. When one would like to know which of his contacts will participate to a scientific conference, he can query the calendar of all his contacts in order to retrieve the answer. For sure, according to the security settings of the corresponding contact, he may be allowed or not to access the calendar.

The mobile device information can as well be accessed remotely, from any web browser, by any persons who has granted the access to it. In this case, acts like a web server.

We presented a first version of the RDF content provider in [2]. This layer, built on top of the Android content provider, allowed to share application data inside the phone. In this paper, we extend the previous version by adding capabilities to access external RDF data and to share application data as linked data on the web.

We first describe the context in which the Android Platform stores its data, and how it can be extended in order to integrate RDF. Then, we present two applications that sustain its feasibility: the first one is an RDF browser that acts like a linked data client and allows the navigation through the device information, and the second one is an RDF server which exposes its information to the outside world. We continue to present the challenges raised by such applications and solutions we implemented for them. Finally, we conclude presenting future improvements and challenges in this field.

2 Android Content Providers

Inside the Android system, each application runs in isolation from other applications. This Linux-based operating system assigns to each application a different and unique user. Only this user is granted access to the application files. This allows one to take advantage of a secure environment, but prevents the exchange of data across applications. To overcome this drawback, Android provides the content provider mechanism.

Content providers enable the transfer of structured data between device applications. They encapsulate the data and control the access to it through an interface. This interface empowers one to query the data or to modify it ([4], [3]).

A content provider is a subclass of `ContentProvider` and implements the following interface:

```
Cursor query( Uri id, String[] proj, String select, String[] selectArgs, String orderBy )
Uri insert( Uri id, ContentValues colValueList)
int update( Uri id, ContentValues colValueList, String select, String[] selectArgs )
int delete( Uri id, String select, String[] selectArgs )
String getType( Uri id ) .
```

With the content provider API, each data (table or individual) is identified by a URI having the following structure:

`content://authority/path/to/data`

The `content:` scheme is the cornerstone of each Content Provider URI, the `authority` identifies the provider, i.e., the dataset, and the `path/to/data` identifies a particular table or individual (row) in the dataset. For example, the URI `content://contacts/people` refers to all the people in the contact application, and the URI `content://contacts/people/33` identifies a specific instance of these, namely the instance having the id 33.

When an application wants to access a particular piece of data, it queries its URI. This is done through a call to the `ContentResolver` which is able to route the query to the right content provider.

From a semantic web point of view, using URIs to identify data is a strong point of the Android content providers. Still, there are several limitations if we would like to use them as a linked data interface.

Specifically, URIs used by content providers are local to each device, i.e., not dereferenceable on the web, and not unique. The content scheme used by providers is not a standard protocol. Furthermore, two distinct devices will use the same URI to identify different data. For example, by using `content://contacts/people` one would be able to access the contacts from both devices.

Another drawback is the SQL interface of the Android content providers. The queries are issued in an SQL manner and the results are presented to the user as a table.

3 The RDF Content Provider Framework

We designed the `RDFContentProvider` framework to give a semantic web flavour to Android and to overcome these problems. It is composed of the `RDFProvider` API and the `RDFContentResolver` application. The API must be included inside the applications that want to access RDF providers and inside the applications that want to define new RDF content providers. The `RDFContentResolver` application is the one that records all the RDF content providers installed on the device and routes queries to the relevant provider. Figure 1 gives an overview of the framework architecture.

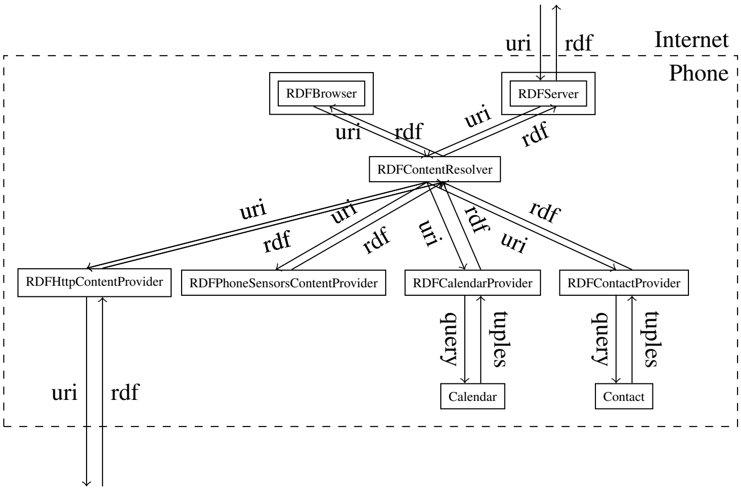


Fig. 1. The architecture components and the communication between them. Components with double square have a graphic user interface.

3.1 The RDF Provider API

The `RDFProvider` API delivers the following classes and interfaces:

- `RdfContentProvider`: An abstract class that should be extended if one wants to create an RDF content provider. In fact, it subclasses the `ContentProvider` class belonging to the Android framework;
- `RdfContentResolverProxy`: A proxy used by applications to send queries to the `RDFContentResolver` application;
- `Statement`: A class used for representing an RDF statement;
- `RdfCursor`: An iterator on a set of RDF statements;

- `RdfContentProviderWrapper`: A subclass of `RdfContentProvider` which allows for adding RDF content provider capabilities to an existing classical content provider.

`RDFContentProvider` follows primarily the same kind of interface as `ContentProvider`. The minimal interface to implement linked data applications is:

- `RDFCursor getRdf( Uri id )`

The Cursor iterates on a table of subject-object-predicate (or object-predicate) which are the triples involving the object given as a URI. If one wants to offer a more elaborate semantic web interface, i.e., a minimal SPARQL endpoint, the following methods have to be also implemented:

- `Uri[] getTypes( Uri id )`: returns the RDF types of a local URI;
- `Uri[] getOntologies()`: ontologies used by the provider;
- `Uri[] getQueryEntities()`: classes and relation that the provider can deliver;
- `Cursor query( SparqlQuery query )`: returns results tuple;
- `Cursor getQueries()`: triple patterns that the provider can answer.

The RDF providers that we have developed so far are implementing only the first three primitives.

3.2 The RDF Content Resolver Service

The `RDFContentResolver` service has the same goal as the `ContentResolver` belonging to the Android framework. It maintains the list of all the installed RDF content providers, and forwards the queries it receives to the corresponding one. This application is never visible to the user, therefore we have implemented it as an Android service.

When an RDF Content Provider is instantiated by the system, this provider automatically registers to the `RDFContentResolver`. A principle similar to the one from the Android Content Provider framework is used.

The `RDFContentResolver` can route both the local (`content:`) and external (`http:`) URI-based queries. In case of a local URI, i.e., starting with the `content` scheme, the resolver decides to which provider it must redirect the query. In case of an external URI, i.e., starting with the `http` scheme, the provider automatically routes the query to the `RDFHttpContentProvider` (see Figure 1).

The `RDFHttpContentProvider` allows one to retrieve RDF data from the Web. It parses RDF documents and presents them as `RDFCursor`s. So far, only the minimal interface has been implemented, i.e., the `getRdf( Uri id )` method.

3.3 RDF Providers for Address Book, Calendar and the Phone Sensors

The RDF Content Resolver application is also bundled with several RDF content providers encapsulating the access to Android predefined providers. The Android framework has applications that can manage the address book and the agenda. These two applications store their data inside their own content provider.

In order to expose this data as RDF, we developed the `RDFContactProvider` and the `RDFCalendarProvider`. These providers are wrapper classes for the ContactProvider and the CalendarProvider residing inside the Android framework.

`RDFContactProvider` exposes contact data using the FOAF ontology. It provides data about a person's name (display name, given name, family name), about its phone number, email address, instant messenger identifiers, homepage and notes.

`RDFCalendarProvider` provides access to the Android calendar using the RDF Calendar ontology[1]. The data supplied by this provider contains information about events, their location, their date (starting date, ending date, duration, and event time zone), the organizer of the event and a short description.

`RDFPhoneSensorsContentProvider` aims to expose sensor data from the sensors embedded inside the mobile device. Contrary to the others, they are not offered as Content Providers. At the present time, it only delivers the geographical position (retrieved using the Android LocationManager service). In order to express this information in RDF, we use the geo location vocabulary[2], the one that provides a namespace for representing lat(itude) and long(itude).

4 RDF Browser

The RDF Browser acts like a linked data client. Given a URI, the browser makes an HTTP URI request in order to retrieve the information from the specified location. If the data contains other URIs, the user can click on them and the browser will issue a new query with this URI.

An example can be found in Figure 2. In this case, the user uses the `RDFBrowser` to get the information about the contact having the id 4. When the browser receives the request, it sends it further to the `RDFContentResolver`. Since the URI starts with the `content://` scheme and has the `com.android.contacts` authority, the resolver routes the query to the `RDFContactProvider`. This provider retrieves the set of triples describing the contact and sends it to the calling application which displays it to the user. Thereupon, the user decides that he wants to continue browsing and selects the contact's homepage. In this case, since the URI starts with the `http://` scheme, the resolver routes the query to the `RDFHttpContentProvider`. The same process repeats and the user can see the remote requested file, i.e., Tim Berners-Lee FOAF file.

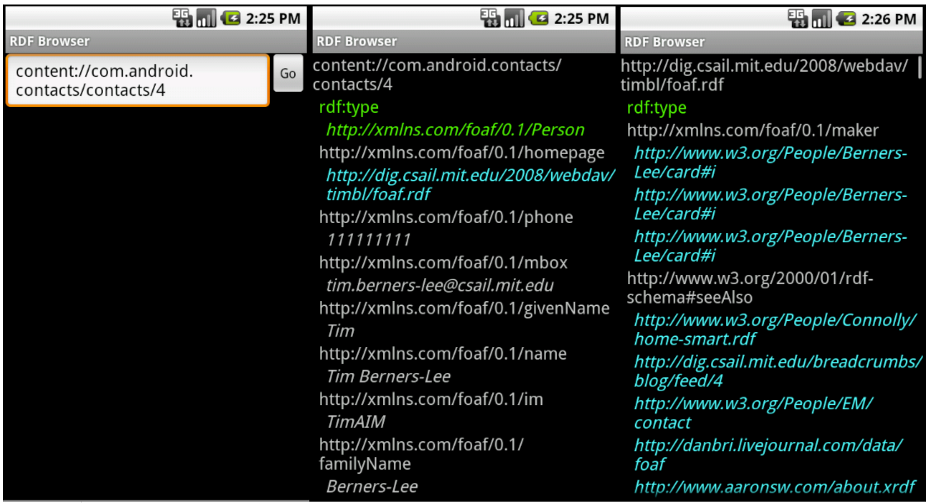
5 RDF Server

The RDF Server is a new component added to the architecture. This server provides to the outside world the data stored into the device as RDF. Due to the fact that the server must maintain a permanent connection to the Internet without user interaction, we implemented it as an Android service, i.e., a background process.

One important issue appears when one would like to get data from a device because the URI used to query the content providers has a local meaning. In the outside world,

[1] RDF Calendar vocabulary: `http://www.w3.org/TR/rdfcal/`.

[2] Geo location vocabulary: `http://www.w3.org/2003/01/geo/` .



content://com.android.contacts/contacts/4

content://com.android.contacts/contacts/4
rdf:type
http://xmlns.com/foaf/0.1/Person
http://xmlns.com/foaf/0.1/homepage
http://dig.csail.mit.edu/2008/webdav/timbl/foaf.rdf
http://xmlns.com/foaf/0.1/phone
111111111
http://xmlns.com/foaf/0.1/mbox
tim.berners-lee@csail.mit.edu
http://xmlns.com/foaf/0.1/givenName
Tim
http://xmlns.com/foaf/0.1/name
Tim Berners-Lee
http://xmlns.com/foaf/0.1/im
TimAIM
http://xmlns.com/foaf/0.1/familyName
Berners-Lee

http://dig.csail.mit.edu/2008/webdav/timbl/foaf.rdf
rdf:type
http://xmlns.com/foaf/0.1/maker
http://www.w3.org/People/Berners-Lee/card#i
http://www.w3.org/People/Berners-Lee/card#i
http://www.w3.org/People/Berners-Lee/card#i
http://www.w3.org/2000/01/rdf-schema#seeAlso
http://www.w3.org/People/Connolly/home-smart.rdf
http://dig.csail.mit.edu/breadcrumbs/blog/feed/4
http://www.w3.org/People/EM/contact
http://danbri.livejournal.com/data/foaf
http://www.aaronsw.com/about.xrdf

Fig. 2. An example of using the RDF Browser.

the URI used to query the address book of two different persons will be the same, but the content of the address book will be different.

The server principles are quite simple. In the beginning, the server receives a request from the outside. Then, it dereferences the requested URI, i.e., it translates the external URI into an internal one, which has meaning inside the Android platform. The RDF Server sends it further to the `RDFContentResolver`. In a manner similar to the one explained for the RDF Browser the set of triples is obtained but, before sending this set to the server, the URIs of the triples are externalized and the graph is serialized using a port of Jena under the Android platform.

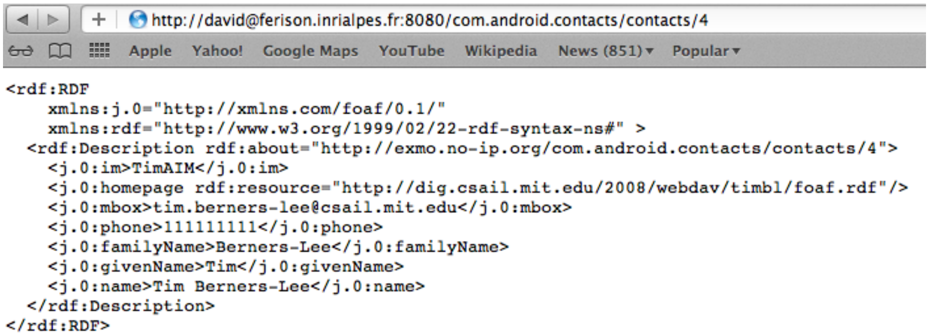
The URI externalization process translates the local URI `content://authority/path/to/data` into the dereferenceable one `http://deviceIPAddress:port/authority/path/to/data`. Reversing the translation of such a URI is possible since both the authority and the path are kept during the externalization process.

Usually, mobile devices do not have a permanent IP address and thus, the externalized URIs are not stable. To overcome this, a dynamic DNS client[3][4] can be used.

In addition, the server supports a minimal content negotiation mechanism. If one wants to receive the data in RDF/XML, it will set the MIME types of the Accept-type header of its request to "application/rdf+xml" or to "application/*". In the opposite case or when the client sets the MIME type to "text/plain", the data will be transmitted in an N-TRIPLE format. Not only the requester has the opportunity to express its preferences regarding the format of the received data, but the default format of the transmitted data can be specified in the server settings, as well the port on which the server can listen on and the domain name server for it.

[3] Dynamic DNS Client: `https://market.android.com/details?id=org.l6n.dyndns&hl=en`.

[4] DynDNS: `http://dyn.com/dns/` .



http://david@ferison.inrialpes.fr:8080/com.android.contacts/contacts/4
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (851)
Popular

```
<rdf:RDF
    xmlns:j.0="http://xmlns.com/foaf/0.1/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <rdf:Description rdf:about="http://exmo.no-ip.org/com.android.contacts/contacts/4">
    <j.0:im>TimAIM</j.0:im>
    <j.0:homepage rdf:resource="http://dig.csail.mit.edu/2008/webdav/timbl/foaf.rdf"/>
    <j.0:mbox>tim.berners-lee@csail.mit.edu</j.0:mbox>
    <j.0:phone>111111111</j.0:phone>
    <j.0:familyName>Berners-Lee</j.0:familyName>
    <j.0:givenName>Tim</j.0:givenName>
    <j.0:name>Tim Berners-Lee</j.0:name>
  </rdf:Description>
</rdf:RDF>
```

Fig. 3. RDF Server response.

An example can be found in Figure 3. In this scenario, the user retrieves information about the fourth contact from the device address book. The request is processed by the RDF Server in a manner similar to the one of the RDF Browser.

6 Technical Details

The RDF Server included in our architecture eases the access of the user to the RDF data found on the web. For that purpose, we wanted to reuse an existing semantic web framework, such as Jena or Sesame. Yet they are not suitable to be employed under the Android platform (the code depends on some libraries that are unavailable under Android). There are a few ports of these frameworks to Android: Microjena[5] and Androjena[6] are ports of Jena and there exists a port of Sesame to the Android platform mentioned in [1]. We use Androjena.

A problem that arises when we use this framework is that the size of the application increases substantially. This problem could have been avoided by reimplementing only the Jena modules that are needed in our architecture. Still, we would like to improve our architecture by adding more features (such as a SPARQL query engine) that require additional modules to those used to read/parse/write RDF, available in Jena.

A tool that we found useful in our development process was ProGuard. ProGuard[7] is a code shrinker, optimizer, and obfuscator. It removes the unused classes, methods or variables, performs some byte-code optimizations and obfuscates the code. The tool proved to be efficient in reducing the size of our application (our framework including Androjena) by half, i.e., its initial size was 6.48MB, and after we applied the tool it diminished up to 2.98MB.

The existence of such tools as ProGuard, is a step forward in the continuous battle between applications that require a considerable amount of space for storing their code and devices with a reduced memory storage.

We are currently examining how to query the device data using SPARQL. There are two main ways of doing this:

[5] http://poseidon.ws.dei.polimi.it/ca/?page_id=59.
[6] http://code.google.com/p/androjena/.
[7] http://proguard.sourceforge.net/.

– creating a new RDF content provider which relies on a triple store to deposit the data [5], and then using SPARQL to query it; or
– translating SPARQL queries into SQL queries, and further decompose it in a form compatible with the ContentProvider interface.

At the moment, we are investigating the second option. There are several available tools that can make the translation from SPARQL to SQL, like Virtuoso or D2RQ. However, these tools solve only half of the problem because the SQL queries have to be adapted to the ContentProvider interface, i.e., the queries have a particular format, different than the SQL one. This interface allows for querying only one view of a specified table at a time, hence it is not possible to ask Content Providers to perform joins.

Further challenges regarding security must be taken into account. The user of the application should be able to grant or to deny the access to its personal data. A specific vocabulary should be used in order to achieve this [8]. More that that, the dangers of granting system access to a third-party user can be avoided by using a secure authentication protocol [9].

As can be seen, there are still technical problems in implementing a full RDF framework at the core of Android. Specific solutions must be developed.

7 Conclusion

Involving Android devices in the semantic web, both as consumers and providers of data, is an interesting challenge. As mentioned, it faces the issues of size of applications and URI dereferencing in mobility situations.

A next step is to provide a more fine grained and structured access to data through SPARQL querying. This promises to raise the issue of computation, and thus energy, cost on mobile platform.

A further issue will be the control of privacy in such a framework. But here too, we think that semantic technologies can help.

References

1. Mathieu d'Aquin, Andriy Nikolov, and Enrico Motta. Building sparql-enabled applications with android devices. 2011.
2. Jérôme David and Jérôme Euzenat. Linked data from your pocket: The android rdfcontentprovider. In *Proc. 9th demonstration track on international semantic web conference (ISWC), Shanghai (CN)*, pages 129–132, 2010.
3. Marko Gargenta. *Learning Android*. O'Reilly Media, Inc., 2011.
4. Reto Meier. *Professional Android 2 Application Development*. Wrox, 2011.
5. Danh Le Phuoc, Josiane Xavier Parreira, Vinny Reynolds, and Manfred Hauswirth. RDF On the Go: An RDF Storage and Query Processor for Mobile Devices. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.

[8] `http://www.w3.org/wiki/WebAccessControl`.
[9] `http://www.w3.org/wiki/Foaf+ssl`.

The Web of Radios - Introducing African Community Radio as an interface to the Web of Data

Anna Bon[1], Victor de Boer[1], Pieter De Leenheer[1], Chris van Aart[1], Nana Baah Gyan[1], Max Froumentin[2], Stephane Boyera[2], Mary Allen[3], Hans Akkermans[1]

[1] Network Institute, VU University, Amsterdam, The Netherlands
{a.bon, v.de.boer, pieter.de.leenheer, c.j.van.aart, n.b.gyan}@vu.nl, hans.akkermans@akmc.nl
[2] World Wide Web Foundation
{maxf,boyera}@webfoundation.org
[3] Sahel Eco, ACI 200 Rue 402, 03 BP 259, Bamako, Mali
mary.saheleco@afribonemali.net

Abstract. The World Wide Web as it is currently deployed can only be accessed using modern client devices and graphical interfaces, within an infrastructure compassing datacenters and reliable, high speed Internet connections. However, in many regions in developing countries these conditions are absent. Many people living in remote rural areas in developing countries will not be able to use the Web, unless they can produce and consume voice-based content using alternative interfaces such as (2G) mobile phone, and radio. In this paper we introduce a radio platform, based on a use case and requirements analysis of community radio stations in Mali. The voice-based content of this radio platform will be made publicly available, using Linked Data principles, and will be ready for unexpected re-use. It will help to bring the benefits of the Web to people who are out of reach of computers and the Internet.

Keywords: community radio, voice-based interfaces, Web of Data, radio platform

1 Introduction

The World Wide Web is perfectly adapted for use by people in developed countries. It is visual, text-based, and mainly written in English or a few other world languages [4]. The Web depends on the availability of computers, datacenters, glass fiber backbones, fixed and wireless networks, 3G mobile telephony and transport of large volumes of data at high speed. In remote rural areas in many developing countries, conditions are different. Poor infrastructure, lack of equipment, low levels of literacy, and use of under-resourced local languages, seriously hamper the access to the Web for many people.

There is a general consensus that the global Information Society must benefit all people in the world. The United Nations Millennium Declaration contains a commitment for developing a *people-centred, inclusive and development-oriented Information Society*

[4] Wikipedia, http://en.wikipedia.org/wiki/Global_Internet_usage,Global Internet Usage

so that people everywhere can create, access, utilize and share information and knowledge to attain the internationally agreed development goals and objectives, including the Millennium Development Goals.[5]
Yet, in many rural regions in Africa community radio is the only source of information. People have radios at home and listen to programs broadcast in local languages every day. Many people have access to simple voice-based (2G) mobile phone, but text messaging is hardly used [1].
The availability of both mobile phone and radio is opening opportunities for new services. E.g. radio listeners phone to the radio station and leave voice messages that they want to have broadcast, or react to popular radio programs leaving news, opinion, regional information etc. Community radio here operates as an important local information hub, where people bring information for further dissemination.
Radio stations in rural areas in Africa operate under harsh conditions. Only the largest and state financed radio stations have a computer and an internet connection. Due to lack of funds many radio stations still use old-fashioned, analogue equipment, such as tape recorders. Yet, it is in the line of expectation that more and more radio stations will have computers and an internet connection in the coming years.
In the current situation the information broadcast by the community radio is volatile: it is not stored and kept for later access or re-use. Radios do not have means to manage, reuse and index this voice-based content.
In this paper we introduce a radio platform as a new interface to the Web. It enables management of radio content in an efficient way, making it accessible and searchable, so that it can serve a broad audience, e.g. Africans in the diaspora, who want to have news from their home villages[6]
Additional, the voice-based radio content on this radio platform might be linked to other data sources on the Web, enabling community radios in Africa to become an interface to the Web of Data. An example of a system that manages market information based on Linked Data principles and produces voice-output as broadcasts for African community radios, is described by De Boer et al. [2]. In the future new applications providing locally relevant information from the Web of Data, such as pluviometric data, agricultural data, market prices etc. might become available through the radio platform.

The radio platform described in this paper not only facilitates production, consumption and management of voice-and web-based radio content, but it also enables access to the Web for people who do not have a computer or the Internet.

Contributions of this paper are:

- A radio platform with both a web and a voice-based mobile interface that allows content creation, retrieval and indexing of spoken radio content.

[5] UNMD, United Nations Millenium Declaration, General Assembly resolution 55/2. United Nations, New York, 2000

[6] Communication possibilities with people living in the diaspora, are described by Serigne Mansor Tall in: Les émigrés sénégalais et les nouvelles technologies de l'information et de la communication. http://www.unrisd.org

- African community radio, introduced as a new interface to the Web of Data

This paper is structured as follows. In section 2 we describe related work. In section 3 the architecture of the radio platform is described, the use cases collected from three different radios in Mali, as well as the principles used to manage the content. In section 4 we describe the challenges related to the organization of the voice-based radio content. In section 5 we discuss future work that must be done on the Web of Radios, including the sustainability aspects.

2 Related Work

Related work on the development of a similar platform was done in the Freedom Fone[7]. Freedom Fone is a project initiated by The Kubatana Trust of Zimbabwe, a civil and information activist platform from Zimbabwe. Freedom Fone is open source software for creating audio content using phone. Freedom Fone provides a voice platform similar to the basic setup proposed in this paper, but without the Linked Data enabled data management.

Research on speech recognition started in the 1930s and resulted in commercial deployments of voice-based services in the 1970s. Major achievements on language recognition, mainly for English, took place in the 1980s and 1990s and culminated in the development of VoiceXML by the W3C Voice Browser group, in 1999, facilitating and standardizing the development of voice applications [3].

Sheetal Agarwal et al. from IBM Research India, developed a system to enable authorship of voice content for 2G phone in a web space, they named the WWTW or World-Wide Telecom Web. The system is not connected to the Web, therefore not allowing access by third party search engines. The system represents closed web space, within the phone network. Especially the lack of open search possibility constrains its growth [4].

From Burundi a system has been reported [5] to use tagging software and multimedia mobile data collection. The software is named EthnoCorder[8]. The NGO that co-developed this app was Help Channel Burundi. However, because of the current unavailability of multimedia devices in the given rural context, this technical solution may be still out of reach of community radio stations targeted in this study.

A related project on the Semantic XO and Linked Data for developing countries is described by Guéret et al. [6]. The Semantic XO is a system that connects rugged, low-power, low-cost robust small laptops (aka the XO promoted by the One Laptop Per Child organization) for the empowerment of poor communities, based on Linked Data principles in order to publish previously unpublished data.

De Boer et al. [2] describe a distributed voice- and web-based market information system, named Radio Marché, aimed at stimulating agricultural trade in rural areas of Africa. This system connects to regionally distributed market information systems, using Linked Data approaches.

[7] Freedom Fone, http://www.freedomfone.org

[8] http://www.ethnocorder.com/

3 The radio platform

The design of the radio platform described in this section is based on extensive use case and requirements analysis, performed in Mali, with the collaboration of radio journalists from community radio stations. The research was done as part of the Foroba Blon[9] project[10], funded by the International Press Institute, and the VOICES project[11], partially funded by the EU, within the 7th Framework Programme. The Foroba Blon project is aimed at supporting and promoting citizen journalism in developing countries. The VOICES project is aimed at developing innovative mobile voice services to support users in underprivileged communities in African countries.

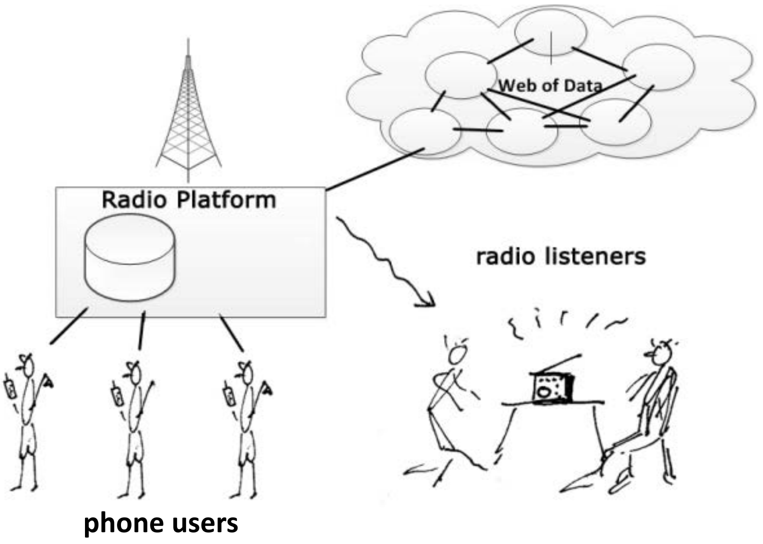


Fig. 1. Conceptual design of the radio platform as a voice-interface to the Web for people who are out of reach of computers and the Internet, but do have phone or radio.

[9] Foroba Blon in Bambara language refers to a large space, where everyone has the right to speak in front of the village chief; the truth must be told here, but only respectfully, without insulting anyone.

[10] Foroba Blon, Citizen Journalism: http://worldplantage.wordpress.com/2012/01/14/community-radio-in-tominian-and-segou-mali/

[11] VOICES: http://www.mvoices.eu

3.1 Community radio in Mali and how it operates

In Mali many community radio stations exist. Some are state funded and connected to the national broadcasting service ORTM (Office Radio Télévision du Mali). Others are privately funded or completely self-supportive. According to their business, funding scheme, size and location some radio stations do have computers and internet, some have computers without internet connection and some do not have any computer facilities at all. All these radio stations are situated within the coverage area of mobile telephony.

The Malian community radios have large bases of listeners and the radius of coverage ranges between 100-200 km. These radio stations create their own programs and broadcast local and regional news, music, informative programs, round table programs and paid announcements. Three radio station stations are involved in the projects described in this paper. These are: Radio ORTM Ségou, a state owned radio, that has computers and a 2 Mbps fixed line (DSL) internet connection. Radio ORTM Ségou broadcasts programs in French and Bambara, the most widely spoken language in Mali.

The second radio station is Radio Moutian, in Tominian. This radio is independent and its funding is based on paid airtime for announcements and private gifts from third parties. Radio Moutian has a computer but no internet connectivity. Programs are mainly broadcasted in Bomu, a local language fro the Tominian region. The third radio is Radio Seno in Bankass. This radio is independent from the Malian state and has only analogue equipment. There are no computers, there is no internet connection here, but the radio has many listeners in the region around Bankass. The main language spoken here is Dogon. The activities of the three above mentioned radio stations are related to three types of end-users or customers:

- NGOs, that buy airtime to broadcast public announcements about informative and educational topics, such as agriculture and public health information. This type of service is usually based on fixed monthly subscriptions to airtime for recurring broadcasts.
- Non-commercial listeners from the region, who buy a few minutes of airtime and pay a broadcast fee per minute airtime. The information is usually brought to the radio, or communicated via phone and subsequently written down on paper by the radio staff. Some listeners call in on a given time slot (one hour per week) and leave a short voice message (few seconds only) as a reaction to a program that was broadcast on a certain popular topic. These messages are named letters to the editors (LTE).
- Journalists or trusted village reporters that phone to the radio and leave local news or interviews on a regular base. In the current situation, all incoming phone calls are attended by a radio staff member and annotated in tabular form on paper.

3.2 The radio platform architecture

The proposed radio platform, which we named Foroba Blon (FB), consists of a data store containing recorded voice messages and related meta-information.

The interface to the FB radio platform is either purely voice-based, through mobile phone, for entering new content, or web-based. Users of the mobile interface are the listeners from the region, who enter letters to the editor (LTE). These people only have mobile phones but no access to the Internet. Their calls are answered by the system with a pre-recorded welcome message in a local Malian voice inviting them to leave their message. For the sake of user-friendliness, the user interface and the dialogue for this category of users is kept as short and simple as possible, since the expected callers will be unfamiliar with interactive voice response systems and may not respond to a complex computer-generated dialogue asking to press buttons etc.

Another category of users of FB are the trusted reporters calling from the field, and also using the mobile interface. They phone in and leave their spoken report for broadcasting. These users are previously registered, having their phone number, name, address and preferred language in FB. These users will be trained to navigate the voice-menu, and use the IVR system, asking to press a button on the phone to confirm or answer a question about their current location, subject of the message, etc. The FB system always answers the registered caller in his/her preferential language.

The voice messages are stored as audio files in the FB data store, together with meta-information being the date and time of the call, the length of phone call in seconds, the phone number of the caller. Messages from trusted users are linked to the owner, his/her address, and his/her preferred language. For all users of the system, confidentiality and anonymity will be ensured, according to the broadcast policies used by the radio stations in Mali.

The FB radio platform also has a normal web interface, where internet-connected end-users/customers can access and upload a voice message. Depending of their customer relationship to the radio, they can login to the radio-platform as (i) registered users such as NGOs, and trusted reporters, or (ii) as unregistered users. There is an option to sign up and create a user account by registering the name, phone number, village and preferred language. Unregistered users can access former broadcasts since this is public information.

For the radio user, FB provides a web-based interface, enabling them to manage the data in the data store. It provides a file list where they can access, listen, broadcast, delete files, and add/update/delete meta-information.

The radio station that has no computer nor internet, has only a very limited interface to the radio platform, since this is the constraint of a voice interface. The radio user receives a welcome message asking if she wants to hear the last 10 messages, or if she wants to manage the welcome messages to the end-users.

The FB radio platform is hosted either locally, on a stand-alone computer, or *in the cloud*. The FB consists of a voice platform running an open source web server and a local voice browser that handles the voice interaction. The FB radio platform uses a GSM gateway device, e.g. OfficeRoute,[12]a device that handles incoming and outbound calls and streams the voice messages to and from the phone.

[12] OfficeRoute: http://en.flossmanuals.net/freedom-fone/connecting-officeroute

The FB radio platform could in theory be physically hosted anywhere in the world, on any web server, connected to the Internet. However, in this actual case in rural regions of Mali, this is not possible. Firstly, the radio platform has to be accessible using an inexpensive local Malian phone number, so it must be connected to a Malian phone network. Secondly, the web service accessed over the Internet must also be accessible locally. Since the internet connectivity in Mali is usually of low bandwidth and of high latency, voice web services hosted in datacenters in the US or Europe, are too slow for proper deployment in Mali. For these two reasons, the system has to be preferably hosted locally in Mali. In the absence of good and reliable datacenters or hosting providers in Mali, the radios can decide to deploy the FB radio platform on a local computer at their own premises. Obviously before this can be done, the radio staff members have to be trained how to do operational maintenance of the FB platform, and especially how they can cope with frequent power outages, and bring the system back to a consistent state

4 Organizing the radio content

The next challenge is how to manage the spoken content of un-resourced languages such as Bomu, Bambara and Dogon. Since up to present no interactive voice response (IVR) systems exist for these languages, the voice-based content cannot be indexed by conventional search engines. Therefore collecting as much meta-information as possible is essential. Very simple ways of indexing the messages are based on owner (known through phone number) automatic language recognition, time slot, (e.g. *all messages collected on January 13 between 10 and 11 a.m. are related to the radio program on harvesting sheanuts).* The radio journalist can manually enter meta-information such as keywords, village region, language, name or any other attribute to an audio file using her radio-web interface. In the future existing tagging systems such as EthnoCorder may be considered, to facilitate meta-data collection.

4.1 Prepare the data for unexpected re-use

For the moment, the data of the radio platform will be only used locally, but we want to prepare the platform for the future. According to Tim Berners-Lee it is the unexpected re-use of information which is the great value, added by the Web. We will use Linked Data principles for the audio data on the FB radio platform, as following the golden rules by Berners-Lee [7] for publishing and connecting data on the Web, while adhering to its architecture and standards: 1. Using URIs as names for things; 2. Using HTTP URIs so that people can look up those names; 3. Providing useful information, using the standards like RDF, SPARQL; 4. Including links to other URIs, so that people can discover more things, especially across regions and future similar community radio platforms. If the FB radio platform proves to be a success, other instances of FB may be installed at local radio stations in Mali, across borders, in neighbouring countries, Burkina Faso, Ghana, Senegal, Guinée where conditions with regards to illiteracy, local languages, mobile telephony and community radio are similar to those in Mali. This will create a Web of African community radios that are linked to each other and that will eventually become part of the Web of Data.

4.2 Organize an open source community of developers to create applications for the radio platform

In the VOICES and Foroba Blon projects one instance of the radio platform is developed by a small team of developers, in collaboration with end-users[13] sponsored by the International Press Institute as a pilot project. However, to enable further development of the radio platform, and to expand the scale of the web of radios, it is important to look at new ways of production and consumption of data and services. African community radios operate in a low-income region where the sustainability of a system relies on the underlying business model. Community radios do not have enough earnings to invest in new systems, and their listeners-base is large, but poor. Application development will therefore to be organized in a cost-effective way. We propose to organize an open source community of developers and to rely on commons-based peer production for the development of applications that will open the Web of Data to radio using voice-modality.

5 Discussion and future work

From this paper it becomes clear that the Web of African Radios can only emerge as an interface to the Web of Data, when sufficient applications are built, that link voice-based content. For the navigation of voice menus and other voice-based dialogues small subsets of the local languages such as Bambara and Bomu have to be recorded and resourced using time-consuming techniques and efforts. The user interfaces have to be extensively tested and validated with end-users in the local situation, since these are culturally sensitive. For the resourcing of more local languages crowd-sourcing techniques may be applied. The issue of meta-information is another important topic. In the model presented for the FB radio platform in this paper, only a small amount of meta-data is collected. When the repositories of spoken content start to become larger, new innovative ways of describing spoken content have to be developed. To contribute to a critical mass of content and applications that are necessary in this rural domain, a socio-technical network has to be put in place, that must be supported by a community of contributors: web developers, listeners that provide meta-information, local ICT-entrepreneurs, people who are willing to produce and consume data. According to Kazman and Hong-Mei Chen [8] organizing a community of developers around an open source service requires a consolidated kernel infrastructure, allowing peripheral services to be created by a de-centralized community of developers. Specific social and technical mechanisms are needed to ensure long-term participation and to encourage community engagement. In this case this is justified by the aim to open the Web of Data to people who are out of reach of computers and the internet.

References

1. Akkermans, H., Grewal, A., Bon, A., Tuyp, W., Allen, M., Gyan, N.B.: W4RA-VOICES

[13] At the moment of writing, the use cases have been collected in Mali, and the FB platform is being built accordingly. However, no feed-back has yet been received from the users

eld report. Tech. rep., Web Alliance for Regreening Africa (2011), http://www.mvoices.eu/2011/03/25_Voices-W4RA_Public_Report.pdf

2. De Boer, V., De Leenheer, P., Bon, A. Gyan, N.B., Van Aart, C., Guéret, C., Tuyp, W., Boyera, S., Allen, M., Akkermans, H. Radio Marché: Distributed Voice en Web Interfaced Market Information System under Rural Conditions, in prep.
3. W3C: Voice Extensible Markup Language VoiceXML Version 2.0, W3C Recommendation 16 March (2004), http://www.w3.org/TR/voicexml20/
4. Agarwal, S.K., Jain, A., Kumar, A., Rajput, N.: The world wide telecom web browser. In: Proceedings of the First ACM Symposium on Computing for Development. ACM DEV 10, (2010) New York, NY, USA, ACM 4:14:9.
5. Horst, N.: EthnoCorder in Burundi: innovation, data collection and data use. In: Participatory learning and action. IIED (2011). http://pubs.iied.org/pdfs/14606IIED.pdf
6. Guéret, C., Schlobach, S.: SemanticXO : connecting the XO with the worlds largest information network. In: Proceedings of the First International Conference on eTechnologies and Networks for Development, ICeND2011. (2011)Communications in Computer and Information Science, Springer LNCS.
7. Berners-Lee T. : Linked Data, the four rules (2006). http://www.w3.org/DesignIssues/LinkedData.html
8. Kazman, R. & Hong-Mei, C.: The Metropolis Model. A new logic for development of crowd-sourced systems. Communications of the ACM, (2009), Vol. 52, No 7.

Voice-based Access to Linked Market Data in the Sahel

Victor de Boer, Nana Baah Gyan, Anna Bon, Pieter de Leenheer, Chris van Aart, Hans Akkermans

Dept. of Computer Science, the Network Institute, VU University, Amsterdam, The Netherlands
{v.de.boer, n.b.gyan, a.bon, pieter.de.leenheer, c.j.van.aart, j.m.akkermans}@vu.nl

Abstract. In this paper, we present our ongoing efforts to bring the Web of Data to rural communities in the Sahel region. These efforts center around RadioMarché, a market information system (MIS) which can be accessed using first-generation mobile phones. We argue that linking the locally produced and consumed data to (external) Linked Data sources will increase its value. We describe how RadioMarché data is available as Linked Open Data and present a prototype demonstrator with voice-based access to this linked market data. Through this interface, the Linked Data can be accessed using first generation mobile phones. As such, these are first steps towards opening the Web of Data to local users that do not have appropriate hardware to produce and consume Linked Data. We present a number of use cases as well as the current deployment state. We also discuss our current efforts to leverage the creation of Linked Data in development regions and build applications on this Linked Data.

1 Introduction

Development and use of the Web of Data has until now mainly focused on developed countries, as was the case with the Web of Documents before it. 4.5 billion people - mainly in developing countries- currently can not access the World Wide Web. The reasons for this include infrastructural ones such as a lack of high bandwidth Internet connections and reliable power supplies as well as socio-economic issues such as the high cost of buying Personal Computers, language mismatches and lack of reading and writing abilities. For our case study in Mali, only 1.8% of the population has Internet access[1], only 10% has access to the electricity network[2], and only 26.2% is literate[3]. Currently, a number of efforts are being undertaken to bridge this so-called 'digital divide' in the World Wide Web, including the recent forming of the Web Foundation. As was argued in [1], while the Web of Documents has been around for 20 years, as engineers of the much newer Web of Data, we have the opportunity to not let the "digital Linked Data divide" grow too large. To avoid a seemingly unbridgable gap, we should consider the underprivileged majority as we design Linked Data architecture, describe use cases and provide access to that Linked Data. In this paper, we describe our ongoing

[1] http://www.internetworldstats.com/ Internet World Statistics, Miniwatts Marketing Group.

[2] http://www.developingrenewables.org/energyrecipes/reports/genericData/Africa/061129%20RECIPES%20country%20info%20Mali.pdf

[3] http://www.indexmundi.com/facts/indicators/SE.ADT.LITR.ZS Index Mundi 2011.

investigations the implementation of Linked Data-backed solutions for the rural Sahel regions.

1.1 RadioMarché

Our efforts center around a Market Information System, RadioMarché [2], a web-based market information system being developed within the VOICES project [4] aimed at stimulating agricultural trade in the Sahel region. The RadioMarché system augments an already running Market Information System (MIS), that was introduced by our partner NGO, Sahel Eco[5], in the Tominian Area in Mali.

Within RadioMarché local market information about Non-Timber Forest Products (NTFPs) such as honey, tamarind and shea nuts is stored. A local instance of RadioMarché has a data store with rudimentary market information such as product offerings (including product type, quality, quantity, location and logistical issues) and contact details from sellers and buyers. This information is sent to community radios for radio broadcast and is made available for individual potential buyers and sellers. To overcome interfacing and infrastructural issues, RadioMarché has a voice interface which can be accessed through the normal telephone network using first-generation mobile phones.

A first version of RadioMarché has been deployed in November 2011 in the Tominian region. In this paper we describe a prototype demonstrator developed in parallel which exposes the market data from this prototype instance of RadioMarché using Linked Data approaches, so that new opportunities for product and service innovation in agriculture and other domains can be unleashed. The prototype demonstrator also features rudimentary voice-access to the Linked Data.

1.2 Why Linked Data?

We believe that Linked Data as a paradigm is very much suitable for knowledge sharing in developing countries. Linked Data approaches provide a particularly light-weight way to share, re-use and integrate various data sets using Web standards such as URIs and RDF. It does not require the definition of a specific database schema for a dataset [3]. We assume that the majority of the use of the locally produced data will also be consumed locally. Although the specifics of the locally produced data will differ from use case to use case and from region to region, Linked Data provides us with a standard way of integrating the common elements of the data. Also, because we do not impose a single overarching schema on the data, data reuse for new services is easier, both within a region and across regions. The aggregated data can be used by NGO's to assess running programs and increase their own transparency and accountability. We will provide examples in Section 4.

An additional advantage is that Linked Data is well-suited to deal with multiple languages as its core concepts are resources rather than textual terms. Where the Web of Documents, by design, is language-specified, Linked Data is designed to be "language

[4] http://www.mvoices.eu

[5] http://www.saheleco.net

agnostic", which suits our purpose of multilingual and voice-based access well. A single resource, identified by a URI (ie. http://example.org/shea_nuts) can have multiple labels (eg. Shea Nuts@en and Amande de Karite@fr). Other than textual labels, for our voice-services we add audio to the resources with language-specific voice snippets, also identified through URIs.

2 Related Work

Agarwal et al. from IBM Research India, developed a system to enable authorship of voice content for 2G phone in a Web space, they named the WWTW (World-Wide Telecom Web). The whole system creates a closed web space, within the phone network. Linking from one voice site to the other is done through a protocol HSTP, created by IBM. Especially the lack of open search possibility constrains its growth [4].

Several automated market information systems have been developed and built to support farmers and agricultural trade in developing countries. One of the well-known market information systems is ESOKO [5], an online market system, developed and built in Ghana. ESOKO enables sellers and buyers to exchange market information. Google started a project in Uganda in 2009, partnering with MTN and Grameen Foundation to develop mobile applications that serve the needs of poor and other vulnerable individuals and communities, most of whom have limited access to information and communications technology [6]. This system is based on SMS but does not allow voice access.

The Web Foundation has started the Open (Government) Data to "Conduct country level actions and global actions to increase the impact and benefits of Open Data worldwide" [7]. This effort focuses on opening government data in developing countries such as Ghana. Our data is initially designed to be produced and consumer by the regional farmers themselves. Linking our regional data to the (Linked) Open government data could increase the value of both datasets. A related project on Linked Data for developing countries is described by Guéret et al. [8]. The SemanticXO is a system that connects rugged, low-power, low-cost robust small laptops for empowerment of poor communities in developing countries.

3 The RadioMarché linked market data Demonstrator

3.1 The Linked Market Data

For our experimental demonstrator, we transcribed a copy of the up-to-date market information from the RadioMarché prototype deployed in the Tominian region to RDF triples and stored in a ClioPatria triple store [9]. The transcription process is done using XMLRDF rewrite rules [6], the conversion can be run when the RadioMarché database is updated to ensure the database of the deployed version and the linked data store of our prototype are in sync.

[6] http://cliopatria.swi-prolog.org/packs/xmlrdf

Currently, we use PURLs for the resource URIs. The temporary namespace chosen is http://purl.org/collections/w4ra/radiomarche/. An HTTP request to these PURL URIs is redirected to the ClioPatria server, running at http://semanticweb.cs.vu.nl/radiomarche/.

Through ClioPatria's Linked Data package, the RDF data is accessible as Linked Open Data. The result of an HTTP request for a resource is either a human-readable web page [7] or the raw RDF triples describing the resource (in the case of a browser request or an RDF request respectively). A SPARQL 1.0 endpoint is also provided at http://semanticweb.cs.vu.nl/radiomarche/sparql/.

As of February 2012, 31 market offerings are in the triple store. These market offerings have been done by 15 different farmers, living in 13 different villages spread across 6 regional "zones". The market offerings contain the quality, quantity and type of the product the price and contact information. In total, the market data consists of 721 triples.

In the current version of the demonstrator, the FOAF ontology has been used to describe persons. Additionally explicit links from the dataset to external data sources were made manually. These include links from zones and villages to GeoNames concepts, DBpedia geographical resources and DBPedia product descriptions.

3.2 Voice-based access to Linked Data

The linked market data can be browsed through the web using the above mentioned URLs. However, as stated, our goal is to provide a voice-based interface that allows non-intrusive market information access for all users having a first-generation mobile phone.

We have implemented a rudimentary version of a voice-based interface to the linked market data as described in the previous section. The voice service is built using VoiceXML [10], the industry standard for developing voice applications. Although in a deployment version we cannot assume that text-to-speech (TTS) libraries are available for the local languages, we here only implement English-language access to the data, using English TTS. Within the VOICES project, TTSs are currently being developed for local dialects of French as well as local languages such as Bambara and Bomu.

The prototype voice application is running on the Voxeo Evolution platform [8]. The platform includes a voice browser, which is able to interpret VoiceXML documents, includes (English) TTS and provides a number of ways to access the Voice application. These include the Skype VoIP number +990009369996162208 and the local (Dutch) phone number +31208080855.

When any of these numbers is called, the voice application accesses a VoiceXML document hosted on a remote server. This document contains the dialogue structure for the application. In the current demonstrator, the caller is presented with three options, to browse the data by product or region, or to listen to the latest offering. The caller presses the code on his or her keypad (this is Dual Tone Multi-Frequency or DTMF). The voice application interprets the choice and forwards the caller to a new voice menu.

[7] For example http://purl.org/collections/w4ra/radiomarche/village_Samoukuy/ shows all information about the Samoukuy village

[8] http://evolution.voxeo.com

For products, the caller must select the type of product ("press 1 for Tamarind", "press 2 for Honey", etc.), for regions the caller is presented with a list of regions to choose from. Based on the choice the application then accesses a PHP document on the remote server, the choice is copied as a HTTP GET variable.

Based on the choice, a SPARQL query is constructed. This SPARQL query is then passed to the RadioMarché Linked Data server, which returns the appropriate results. For a product query, all (recent) offerings about that product are returned. The SPARQL result is then transformed into VoiceXML and articulated to the caller.

The prototype demonstrator and the ways of accessing it are shown in Figure 1.

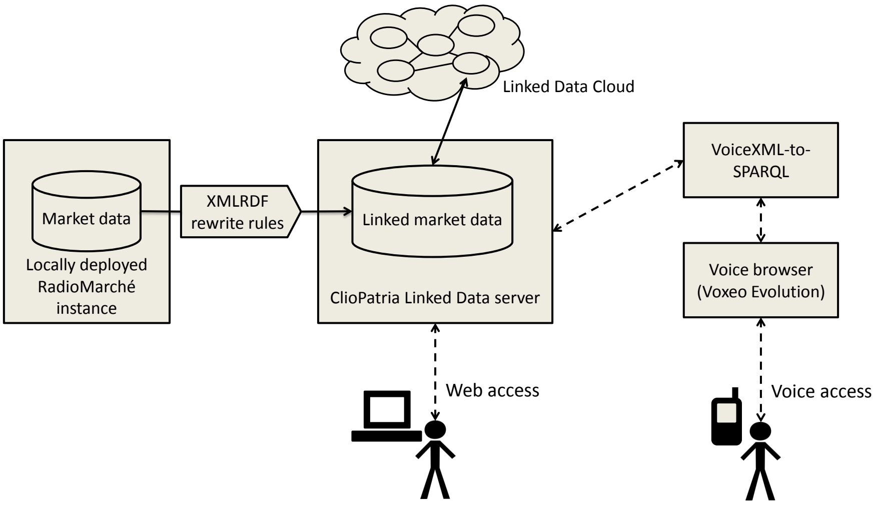


Fig. 1. Schematic representation of the linked market data prototype demonstrator.

Of course, the current method of accessing the data is only one of many possible actions. The caller can be presented with advanced filtering options ("enter the maximum price for offerings of product X", "enter a date range for product offerings") or combinations of data queries. However, because of the slow and linear nature of voice interfaces -when compared to visual UIs- options have to be limited more than with visual interfaces. This means that in our research we will identify useful services on this data and provide Voice-to-SPARQL mappings for these services.

4 Current Work

Voice-access to the linked market data as described above is still very much in an early prototype state. We are currently working on multiple projects to a) expand number of interlinked datasets produced and consumed in the region and b) investigate use cases that use this Linked Data and build services and applications for those use cases. In this section, we describe the current status of these efforts.

4.1 Other Linked Data sets

The following is a list of Linked Data sets currently being realized. Each of these will be related to the linked market data as well as to external sources.

Meeting Scheduler Within the VOICES project a second use case is to develop a voice-accessible meeting scheduling system. The goal of this system is to provide local NGOs with a more effective way to transfer agricultural knowledge about non-timber forest products to their farmer community. The services developed in this case study provide voice access to personal and scheduling information. By integrating this information with the market information from RadioMarché, personal profiles can be enriched with information about the type of products that specific farmers have been producing within a given period. Here a new scheduling and notification service can re-use the market information within a region.

Citizen Journalism Data A second use case that is currently under development by the same team is a voice-based journalism platform, IPI innovation fund, which allows both professional and citizen journalists to send voice-recorded news items to local community radios. The target region for this use case consists of agricultural communities and there is a large possibility for re-use of both technical infrastructure as well as data. To do this, the re-usable resources (e.g. person data, geographical or product information) in the market information data are linked to the relevant resources in the target data set using Linked Data standard relations.

Pluvial Data We are also developing a crowdsourcing platform to transform photocopied data about rainfall in the Bankas area in Mali to Linked Open Data. This platform targets the 'diaspora', e.g. people originally from the region that have since moved to developed countries, where they might have better access to web browsers. The pluvial Linked Data acquired in this way will be linked to the aforementioned data. This can be exploited by our partner NGO as well as other NGOs to analyze for example patterns between rainfall and market offerings.

IDS Data The Institute for Development Studies recently published an API exposing more than 30.000 publications about development research [9]. As part of a recent agreement, we will develop a wrapper around the IDS API to expose its content as high quality Linked Data, enriching it with connections to other Linked Data datasets. These include both general datasets such as DBPedia or GeoNames as well as datasets with information from developing countries that are currently being realized. We will also develop a client application showing the advantages of this publication, exploiting the integration of the IDS data with other Linked Data sets in an information mashup.

[9] http://api.ids.ac.uk

Links to External datasets At the same time, local and national governments as well as NGOs can exploit the linked market data for analytic purposes, monitoring the trade in NTFPs within and across regions. By linking the market information to existing agricultural vocabularies such as FAO's Agrovoc thesaurus[10], the CAB Thesaurus[11], or the USDA's National Agricultural Library NAL[12], the aggregated market data can be used for specific analyses for government or NGO purposes.

4.2 Linked Data Applications

We are currently building a client web application where we use the various Linked Data. This application will use the market data and exploit its links to GeoNames for displaying the market offerings on a map. The links to other dataset (IDS data, pluvial data, etc.) will also be exploited to provide the user with additional information about products or regions.

The application will allow the local NGO to perform various types of analysis based on market data that are useful on the basis of their educational programs. The application will aim to demonstrate the added value of the linked data approach through the re-use of and integration with existing market data from various sources with differing schemes and through the re-use of and integration with market data with publicly available knowledge from the web on agriculture and economics.

At the same time, we will continue to work on client applications for users in the developing regions themselves, focusing on voice-based access to the data. Within the VOICES project, (limited) TTS systems for smaller languages are being developed.

5 Discussion

We have presented the Linked Data version of the RadioMarché system, its data and the voice-based access. This system represents our first steps to brining Linked Data to producers and consumers in developing countries. We describe a demonstrator with locally produced Linked Data which provides rudimentary voice-based access, in addition to browser-based and Linked Data-application access.

Currently, the demonstrator is implemented on commercial-grade and University-provided web servers including the Voxeo Evolution platform, PURL servers and the VU University Amsterdam web server. The voice application is also only reachable through a Dutch local phone number or Skype access. To ensure sustainability of the Linked Data and the client applications, this infrastructure needs to be moved to the developing regions itself as much as possible. The Orange Emerginov platform [13] can provide the web server and voice browser technology needed for this infrastructure and include local Malian phone numbers. The Linked Data servers, voice-interfaces and client applications can be moved to this platform at testing or deployment time. A second option is entirely local. This version has the data and applications running

[10] http://aims.fao.org/website/AGROVOC-Thesaurus/sub

[11] http://www.cabi.org/cabthesaurus/

[12] http://agclass.nal.usda.gov/

[13] http://www.emerginov.org/

on a web-connected dedicated laptop that is be deployed locally. The voice channel is provided by a local voice browser and a GSM gateway (2N OfficeRoute) device connected to the laptop that allows phone calls to be handled by the system on the laptop.

As was discussed in Section 1.2, we aim to include the audio language resources to the Linked Data itself. We are currently gathering language snippets that act as audio labels for resources. These will be added to the data itself so that they can be interpreted by a voice browser directly.

Acknowledgements The authors would like to thank Mary Allen and Amadou Tangara of Sahel Eco. This research is partly funded by the European Union through the 7th Framework Programme (FP7) under grant agreement Num. 269954.

References

1. Guèret, C., Schlobach, S., de Boer, V., Bon, A., Akkermans, H.: "is data sharing the privilege of a few? bringing linked data to those without the web". Outrageous Ideas at International Semantic Web Conference (ISWC 2011). Jury award winning paper. 1st Place (2011)
2. de Boer, V., Leenheer, P.D., Bon, A., Gyan, N.B., van Aart, C., Guèret, C., Tuyp, W., Boyera, S., Allen, M., Akkermans, H.: Radiomarché: Distributed voice- andweb-interfaced market information systems under rural conditions. In: Accepted for publication in Proceedings of 24th International Conference on Advanced Information Systems Engineering, CAiSE'2012, Gdansk, Poland, 25 29 June 2012. (2012)
3. Domingue, J., Pedrinaci, C., Maleshkova, M., Norton, B., Krummenacher, R.: Fostering a relationship between linked data and the internet of services. In Domingue, J., Galis, A., Gavras, A., Zahariadis, T., Lambert, D., Cleary, F., Daras, P., Krco, S., Mller, H., Li, M.S., Schaffers, H., Lotz, V., Alvarez, F., Stiller, B., Karnouskos, S., Avessta, S., Nilsson, M., eds.: The Future Internet. Volume 6656 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2011) 351–364
4. Agarwal, S.K., Jain, A., Kumar, A., Rajput, N.: The world wide telecom web browser. In: Proceedings of the First ACM Symposium on Computing for Development. ACM DEV '10, New York, NY, USA, ACM (2010) 4:1–4:9
5. ESOKO: Esoko. http://www.esoko.com/ (2011)
6. AppLab, G.: Google sms to serve needs of poor in uganda. http://blog.google.org/2009/06/google-sms-to-serve-needs-of-poor-in.html (2009)
7. Foundation, W.W.W.: Open government data. http://www.webfoundation.org/projects/ogd/ retrieved 14-03-2012 (2012)
8. Guéret, C., Schlobach, S.: Semanticxo : connecting the xo with the world's largest information network. In: Proceedings of the First International Conference on eTechnologies and Networks for Development, ICeND2011. Communications in Computer and Information Science, Springer LNCS (2011)
9. Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., Wielinga, B.: Multimedian e-culture demonstrator. In: The Semantic Web - ISWC 2006, Athens, Georgia, volume 4273 of LNCS, pages 951-958, Winner Semantic Web Challenge 2006, Springer Verlag, November 2006 (2006)
10. W3C: Voice extensible markup language (voicexml) version 2.0. W3C Recommendation 16 March 2004 http://www.w3.org/TR/voicexml20/ (2004)

Semantic Web in a Constrained Environment

Laurens Rietveld and Stefan Schlobach

Department of Computer Science, VU University Amsterdam, The Netherlands
{laurens.rietveld,k.s.schlobach}@vu.nl

Abstract. The semantic web is intrinsically constrained by its environment. These constraints act as a bottlenecks and limit the performance of applications in various ways. Examples of such constraints are the limited availability of memory, disk space, or a limited network bandwidth. But how do these bounds influence Semantic Web applications? In this paper we propose to study the Semantic Web as part of a constrained environment. We discuss a framework where applications adapt to the constraints in its environment.

Keywords: downscaling, ranking, constraints, resource bounds

1 Introduction

Agreement that the Semantic Web has become a huge success story is widening: standardised languages exist for representing web data and ontologies, together with tools to deal with such semantic information. This facilitates applications that publish and consume *triples* in the billions, in domains as various as life-sciences, cultural heritage, e-business and journalism. Additionally, the number of robust commercial triple stores is constantly increasing, and successful DBMS such as Oracle become RDF aware. Parallelization and distribution have made expressive semantic reasoning massively scalable, and the movement towards storage and reasoning in the cloud suggest that even the last technological barriers can soon be overcome.

Although we partially share this optimistic view, it only paints a part of a larger picture. In fact, the apparent scalability critically depends on a computational infrastructure that is out of reach for the vast majority of the human population. Powerful servers, supercomputers and clusters are the privilege of few. Often, users of the Semantic Web cannot rely on a constant and reliable Internet connection with sufficient bandwidth. In more remote regions, even electricity supply is not always guaranteed, with restricted battery runtime having to be considered when building applications. But it is not just the infrastructure, or its lack of, that provides unsurpassable boundaries for the Semantic Web to be used and useful. Think, e.g., of real-time user interaction interaction that restricts the computation time to the often very short attention span of humans. After all, who wants to wait more than a few seconds for search results? Additionally, such interaction has to take the human processing bandwidth into account. Anybody wants to see more than 10 search results? No. In short: the Semantic Web is bound by resource availability.

Although those bounds look diverse at first glance, we suggest to study information access and processing on the Semantic Web in the context of these bounds more systematically. The purpose of such an analysis is not an end in itself: such an analysis of the relation between Semantic Web applications and the way they are resource bound can help when building better applications, or to build good applications more easily. A promising approach is to study the explicit and intrinsic orderings and rankings in data and the information need. This information can then be used to deal with the resource-bounds. Take as example the human attention span, which requires applications to produce results extremely fast. This is often impossible in computationally expensive representation languages and for applications involving complex data, unless the intrinsic rankings are used to produce *good* results first and fast. Ranking of goodness is at the basis of such any-time approaches.

This paper is a first attempt to investigate the dependencies between particular resource bounds and the type of orderings and rankings used to overcome the boundaries. Is there a more generic relation between ranking and resources? And if so, can we explicate this relation, and use it to guide the process of building Semantic Web applications in a resource bound world? The rest of this paper introduces a number of assumptions underlying our approach (section 2). Following, we study those assumptions in the context of two Semantic Web applications(section 3). We conclude with future work and open questions(section 4).

2 Constraints & Ranking

In the introduction we claim that resource bounds and rankings are related. This claim constitutes the foundations of an unifying idea for more easily building good application for the Semantic Web: the idea of using orderings in the data and application requirements to deal with explicitly defined resource bounds. This section will introduce those claims more systematically. We will first elaborate on these statements. Afterwards, in the next section, we will study them in light of two very different use-cases.

(1.) The world is resource bound. Our environment is full of constraints. These bounds constrain us, applications, and the way we interact with, and have to build, these applications.

(2.) Deal with these constraints. Applications have to (implicitly) deal with these constraints, which often means trading functionality of the application to remain within the given bounds (e.g. switch to off-line mode when there is no connectivity).

(3.) We need ranking Applications can deal with these constraints by ranking results and/or tasks. This enables the application to take a sub-selection (top-k) part of the results or tasks, and process them. In doing so, the application does not process the complete result or task set. An example where the ordering of tasks is used to adapt to changing bounds, is that of an operating system. Most operating systems will rank the running tasks in importance, and give the higher ranked tasks precedence. This ensures the most important tasks will still run whenever there is a high CPU load (resource bound).

(4.) Ranking: it's all in the data Ranking of results and/or tasks depends on rankings of data. This data often contains explicit ordering such as recency. Or there is a more implicit ordering covering items such as importance or relevance. For this framework we selected a set of ranking measures which are easy to retrieve and calculate:

- *Recency:* How old is this entity
- *Size:* How large is this entity
- *Frequency:* How often is this entity used or accessed
- *Similarity:* How similar is this entity compared to others

(5.) We need explicit bounds In doing so, an application is able to use these bounds as input, and link these bounds to other functionality in the application. The following list is a first attempt at distinguishing between the generic types of constraints relevant for Semantic Web applications:

Hardware Constraints We consider any machine (e.g. server, pc, laptop) hardware restriction as a machine constraint. Examples of such constraints are *Memory*, *Hard disk space* or *Battery power*. Not all constraints are applicable to every domains. For example, the hardware restriction *Battery power* only applies to environments where laptops are used.

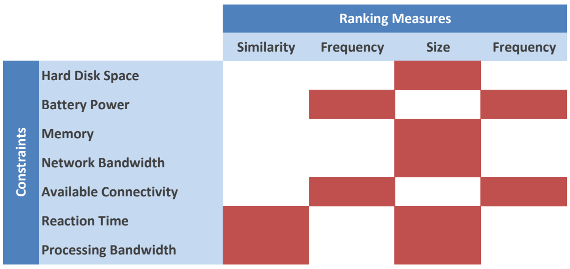
Network Constraints The Semantic Web needs network connections. The connections between the nodes in such a network are constrained by for example *Network Bandwidth* or *available connectivity.*

Interaction Constraints We consider interaction constraints as the constraints imposed by the limits of the agent using the application. These constraints are more difficult to obtain, but they might be retrievable from access logs or user profiles. Examples are *reaction time*[1] or *Processing Bandwidth*, i.e. the maximum information an agent is capable of processing in a given situation.

(6.) Tell me your bounds and ranking, and I deal with them. This requires an explicit link between the constraints and the ranking measures. By using this link, applications can adapt their ranking measures to the changing resource bounds. The ability to adapt is particularly useful, because resource bounds are not static: Network bandwidth might fluctuate, just as available memory can change. Applications and the way they use the ranking measures should change as well.

The picture shown in figure 1 illustrates the plausible connections between the constraints and the ranking measures. The case study descriptions in the next section will elaborate on these connections in more detail.

[1] The maximum time limit between a *request* of the agent to the SW application, and the desired *response* from the application.



		Ranking Measures			
		Similarity	Frequency	Size	Frequency
Constraints	Hard Disk Space			🟥	
	Battery Power		🟥		🟥
	Memory			🟥	
	Network Bandwidth			🟥	
	Available Connectivity		🟥		🟥
	Reaction Time	🟥		🟥	
	Processing Bandwidth	🟥		🟥	

Fig. 1: Ranking Measures vs. Constraints

3 Case Studies

We discuss two very different case studies: the SemanticXO and visualization of census data, and the way rankings and resources bounds relate in them.

3.1 SemanticXO

An environment with obvious constraints is given in the SemanticXO project. The XO laptop is part of the One Laptop per Child (OLPC) project. The aim of OLPC is to create educational opportunities for the worlds poorest children by providing each child with a 'rugged, low-cost, low-power, connected laptop'. The SemanticXO is a project which aims "to provide an infrastructure that is needed to integrate semantic information from the Web of Data into programs (...) running on an XO computer" [2].

Software Stack Currently, the SemanticXOs software stack contains [2]:

- A triple store, in charge of storing triples with meta-data.
- An HTTP server, which serves as a public interface to the triple store, provide de-referenceable URIs for the resources, and serve the files. The HTTP server is accessible by other XOs.
- A common interface (API) for accessing the triples stored locally, on another SemanticXO, or elsewhere on the Web of Data.

The SemanticXO uses this software stack to store information as a Data Graph in the triple store. After the graph is stored in the triple store, it is available to other XOs via the HTTP Server or the API.

Problem description XO's are often used in a network with similar XO's, where a school server functions as an access point with internet connection. Currently, distributing information within such difficult. Moving information from one XO to another requires both XOs to be on-line, and run the same program (called 'activity'). Situations where the recipient of information is off-line, where information may be shared asynchronously, or where information needs to be shared regardless of any running activity, are currently impossible to deal with.

The SemanticXO offers a framework with which to approach this problem. Currently, objects containing meta-information are stored locally as a data store object (DS-Object). These objects contain meta-information, and contain (a combination of) attribute-value pairs. More complex information such as images are stored separately. Generic shipping of objects requires a DS-Object to be wrapped as a Semantic DS-Object (SDS-Object), which is stored as a graph in the triple store. Via the HTTP server or the API, this information is automatically available to other XO's in the network. Using this infrastructure, the nodes can share and sync objects. This creates a network of loosely coupled triple stores between which information has to be distributed.

Constraints & Ranking We will discuss the constraints and rankings in this case study, using the claims from the previous section.

(1.) The world is resource bound Internet connections are unreliable or have a very limited bandwidth, and hardware may not be on-par with regular laptops and desktops, to name but a few.

(2.) Deal with these constraints Not dealing with constraints such as network bandwidth will most likely result in an inadequate distribution of objects. Not all objects may be distributed to all nodes, which means SDS-Objects run a risk of not being delivered to the intended recipient.

(3.) We need ranking Constraints may limit the number of SDS-Objects to be shared, which means a subsection of the SDS-Objects should be synced. This is essentially a top-k ranking problem, where only the most useful objects are cached and synced, and others are ignored.

(4.) Ranking: it's all in the data Relevant ranking measures are *Recency* (i.e. how long ago is this graph created or modified), *Size* (i.e. how big (in bytes) is the graph), and *Frequency* (i.e. how often is this graph synced)

(5.) We need explicit bounds The resource bounds of the SemanticXO are either caused by the hardware specification[1] or the XO network connections. We consider the following resource bounds for the SemanticXO (a subset of the bounds described in section2): Available memory, available hard disk space, battery power, network bandwidth and available connectivity.

(6.) Tell me your bounds and ranking, and I deal with them By combining the resource bounds **(5)**, the rankings **(4)** and the links between both (fig 1), the application can deal and adapt to its resource bounds. We consider the following relations between the constraints and these rankings measures:

1. **Hard Disk Space:** With limited disk space available, giving precedence to smaller (*size*) objects will allow the SemanticXO to store a larger quantity of objects.
2. **Battery Power:** When battery power is running low, giving precedence to *recency* and *frequency* will make sure the old objects which have not been synced that often are ordered higher than others, because older objects in the queue might indicate these objects are not spread through the network.
3. **Available Memory:** Processing large graphs might require more memory. When available memory is low, precedence should be given to smaller graphs.

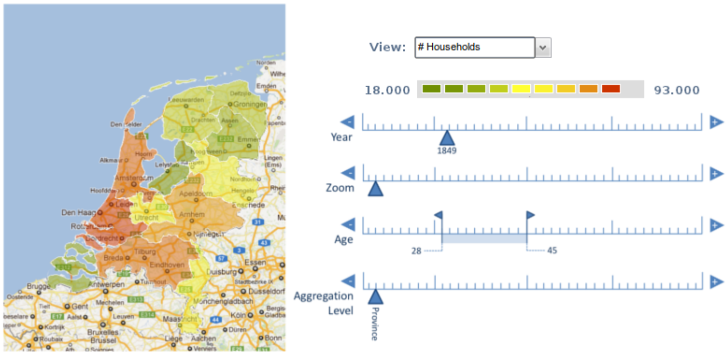


Fig. 2: Census Visualization

4. **Network Bandwidth:** A low network bandwidth should decrease the importance of the *size* of SDS-Objects. This way, at least the smaller objects will still be distributed via this node.
5. **Available Connectivity:** When the SemanticXO is connected to a very limited number of other nodes, the graphs which are not very well distributed in the network should be given precedence. This means objects which are not often synced (*frequency*) and older object (*recency*) are ranked higher than others.

This framework is particularly useful for the SemanticXO, because hardware may differ between XO's. An XO server has completely different hardware (bounds) than a regular XO laptop. Using the described framework, the server and laptop can both optimize their performance based on their own resource bounds.

3.2 Visualizing Census Data

To show that resource bounds are not limited to the environment of developing countries we describe a totally different application where Dutch historical census data is visualized. This use case is part of the Data2Semantics project[2], which focusses on the problems of 'how to share, publish, access, analyse, interpret and reuse scientific data'.

Visualizing Census Data This dataset contains Dutch census counts from between 1795 and 1971, and covers demographic information, such as gender, age, location, marital status, occupation, household and religion. An example of an application making use of the (convert to RDF) dataset is shown in figure 2. This visualization has several selectors:

1. *Variable selection:* The variable of which the distribution is shown on the map
2. *Time:* The year of the census count

[2] www.data2semantics.org

3. *Zoom:* Zoom in or out of the map
4. *Age:* Age range for which to visualize the data for
5. *Aggregation Level:* The aggregation level for the visualization. Changing this value changes the granularity of the colors on the map.

Because this visualization makes use of a SPARQL endpoint, these selectors correspond to SPARQL queries. In this respect, we can consider the *variable selection*, *time*, and *age* as filters, and the *aggregation level* as a 'group by'.

Constraints & Ranking We will discuss the constraints and rankings in this case study, using the claims from section 2.

(1.) The world is resource bound This visualization application is clearly resource bound. Users will expect near to any-time responses when changing the sliders. Additionally, the application will have to deal with current (consumer) hardware and network constraints.

(2.) Deal with these constraints Not adapting to the constraints leads to situations where the response time (when changing a slider) is too long.

(3.) We need ranking A way to deal with these constraints is by pre-executing queries. Executing all possible setting combination will not be feasible, however: there are too many possible combinations of queries. Deciding which queries to pre-execute, though, is a ranking problem/

(4.) Ranking: it's all in the data Relevant ranking measures are:

1. *Query similarity:* Difference in similarity between current view, and the possible query. The bigger the difference, the wider the 'window' becomes which is being pre-executed.
2. *Frequency of usage:* Selectors often used are more interesting to pre-execute. Therefore, queries where the filter value of a frequently used selector is different from the filter value in the SPARQL query of the current view, should have precedence.
3. *Size of expected number of results for a given query:* e.g., decreasing the aggregation level increases the retrieved number of triples.

(5.) We need explicit bounds The resource bounds of this visualization involve: available memory, network bandwidth, reaction time of the user, and processing bandwidth of the user.

(6.) Tell me your bounds and ranking, and I deal with them By combining the resource bounds **(5)**, the rankings **(4)** and the links between both (fig 1), the application can deal and adapt to its resource bounds. We consider the following relations between these constraints and the rankings measures:

1. **Memory:** Limited memory implies limited available space to store the query results. This should give precedence to queries where the expected query result is low.
2. **Network bandwidth:** The smaller the network bandwidth is, the smaller the expected *size* of query results should be.
3. **Reaction time:** More time might allow for a smaller *query similarity*, which results in a bigger window size for the pre-executed queries. A larger interval allow for a larger query result, as there is more time available for the response to be generated and processed.

4. **Processing bandwidth:** The more information a user is able to process, the larger the *size* of the query result may be, and the bigger the allowed windows size may be (*query similarity*).

4 Future Work & Conclusion

This paper showed a perspective with which to consider applications in a (dynamic) resource bound environment. The presented framework should allow semantic web applications to adapt to changing constraints, by making use of the explicit and intrinsic orderings and rankings in the data. We showed the generic link between ranking and constraints, and put these into context of two very different case studies. This framework is not complete though. A roadmap towards completion requires discussion on some unanswered questions, and an implementation/evaluation of this framework. Current open questions are:

- The ranking measures mentioned in section2 are very broad: recency, size, frequency and similarity. Can we keep these notions on such a generic level? Or are these ranking measures open to interpretation, depending on the application domain?
- What are the generic relations between constraints and ranking measures? For both use cases we show some links. Are these generic? And does dealing with one constraints influence other constraints? Is there a cost in dealing with constraints.
- How do the ranking measures relate to each other in such a dynamic system? How is the final ranking score determined?
- How to define interaction constraints? Different users interact differently with a system.
- How does this approach hold up against other methods such as the use of access control, differentiated services, or formal maximization under constraint?

The main question seems to be: does explicating these concepts and relations work in practice? This requires further implementation and evaluation of this framework in one of the use cases.

Acknowledgements. This publication was supported by the Dutch national program COMMIT.

References

1. CL1 Hardware Design Specification (2008)
2. Guéret, C., Schlobach, S.: SemanticXO : connecting the XO with the World's largest information network. In: Proceedings of the First International Conference on e-Technologies and Networks for Development (ICeND2011). "Communications in Computer and Information Science", Springer LNCS, Dar-es-Salaam, Tanzania (2011)

Information Ecosystems and International Development

Mike Powell

IKM Emergent
m.powell@pop3.poptel.org.uk

Abstract. This paper looks at the purpose of developing linked open data applications of value to poorer and less well connected potential users and considers some of the challenges that need to be met if this purpose is to be achieved. It stresses the different societal contexts in which such applications will be deployed and describes some of the potential negative and unintended consequences of enthusiastic but ill prepared initiatives. It describes the development information environment as one of overlapping information ecologies, which, as the metaphor implies, contains both interdependencies and examples of predatory, counter developmental behaviours. From this analysis it suggests a few key areas which combine the potential of real value for users with exciting and ground breaking technological challenges.

1 Introduction

The Downscale 2012 workshop asks software producers[1] to think of the four billion people who do not have access to the internet when designing linked open data platforms and to aim to reduce rather than add to the digital divide. This request seems entirely consistent with the purpose behind many existing linked open data initiatives, which seek to promote greater equity in governance through enabling higher levels of transparency and accountability. It could indeed be said that it is hard to see the point of an interest in linked open data solutions without having an interest in their potential social impact. If this is indeed a common perspective within the linked open data community, then the questions of what social impact is intended and how it may be achieved need to be taken seriously.

This paper, whilst commending the Downscale initiative, argues that the relationship between information and poverty is far more complex than a simple lack of access to communications channels or indeed to information itself. It suggests that deeper study of the potential of information to combat poverty is required. Such deeper study needs, in each situation, to include the specific dynamics of any particular type of information and of the situation in which any

[1] To avoid confusion, we use the word "development" to relate to the field of international development, struggles against poverty etc. and thus must use another word to describe the development and developers of software and other technical solutions

intended user group finds itself. However, some general points can perhaps be outlined here. The desired result of such preparatory work is to identify issues in people's informational needs which pose new technical challenges as well as to become aware of and build meaningful collaborations within the social and organisational contexts in which any technical solutions will be applied.

2 Information and International Development

This paper is also located within a discourse about "international development". The author participated at a panel discussion at ICTD 2011 at which considerable scepticism was expressed by a predominantly technical audience about the relevance of the international development sector. Some argued that the idea of top down investment in processes of poverty alleviation involved outsiders telling other people what they should do and was therefore patronising. Others suggested that such efforts had been largely ineffective and therefore need not be taken into account. In this author's opinion, there is some merit in both arguments. It is also undeniably the case that much economic and social development, including innovation with ICT, emerges from the internal dynamics of local societies and has no relation to the purposeful efforts of international development actors. That said, the sector is a significant presence in many parts of the developing world. Definitions of what constitutes "official development assistance" vary, but the OECD[2] calculated that it amounted to USD 127.6 billion in 2010. When we consider that some USD 2-3 billion of that amount is spent by professional development organisations just on their own information systems, including their linked open data platforms, we can see that the sector is inevitably going to impact on the information landscape. It has the potential both to offer resources and co-ordination to solutions which will help the poor or, as has been argued elsewhere, to worsen the digital divide[3]. It is also the case that history of international development efforts is littered with examples of "technical experts", including in the ICT4D field, failing to recognise or adapt to the socio-cultural or physical specificities of the new environments in which they were working and thus failing spectacularly. Thus for the lessons of its failures as much as its successes, as well as for its exchanges of ideas and the opportunities it offers to build collaborations, it would be unwise to ignore the work of the development sector.

Indeed one starting point for any information and development initiative should be a clear understanding of what is meant by development. Leaving aside the large number of examples which could legitimately be given of development assistance being applied for the direct economic or political benefit of the donor,

[2] OECD, Query Wizard for International Development Statistics, accessed September 15th 2011

[3] See Powell, Davies and Taylor, 2012 "ICT for or against development? An introduction to the ongoing case of Web 3". IKM Emergent Working Paper, in press. The paper takes an historical overview of the use of ICT within the devlopment sector and looks more particularly at early linked open data initiatives within it.

there are longstanding and legitimate debates on what types of development strategy are likely to be most effective. Is it more useful to attempt to support the general development of a poorer country in the expectation that resulting growth and educational improvements will spread to all sectors of society and, because of their wider base, be more sustainable? Or is it better to target interventions at the specific needs and challenges of the poorest and most marginalised? As Wright *et al* [1] explain in their recommendations on Open Government Data in India:

> "The meaning of "open government data" and the purposes it serves will have to be re-examined from an Indian perspective. The reasons that work well in the US and the UK may not work well in India. We also have to be very careful about how we imagine the end-users of open government data. Do we visualise open data as being for the benefit of individual middle-class citizens by helping them to consume (processed) data themselves (with bus routes, for instance), or do we visualise them as being for the benefit of the poor, and thus target NGOs? Do we visualise them as being hackers or laypersons?" (P 40)

This question is fundamental for any information-led development initiative, including those concerned only with data. If "development" initiatives are intended to be of direct benefit to the poorest, this implies improving the relative position of the poor in relation to the rest of society, or, if we are thinking in terms of societies, the relative position of a poor country in relation to global levels. Given that information by itself is of little value without the capacity to make use of it and that such capacity is strongly linked to educational levels, it follows that simply making information more generally accessible is likely to worsen not improve such relative positions. Wright *et al.*, in the paper cited above, highlight the danger of what they call "elite capture of transparency"[4], whereby information coming from or intended to benefit poor people is captured and exploited by others, a process which would be familiar to students of the uptake and use of public health services in Europe. Thus simply enabling the provision of data or information alone is not adequate. Questions need to be posed as to what information would be of particular value to particular groups.

Equally important are issues of how information is provided and curated and to what extent the people using and producing the information have any control over the process. These issues relate to power and profit but also to culture. All societies are based on sets of cultural understandings which may be more or less homogeneous, more or less dynamic in each situation. In recent years changes in informational behaviours have led to new understandings of the interaction of information and culture in many parts of the world. However, as an UNRISD

[4] A full case study, also referred to by Wright, is Benjamin, S. Bhuvaneswari, R. Rajan, P. and Manjunatha 2007 "Bhoomi: 'E-Governance', Or, An Anti-Politics Machine Necessary to Globalize Bangalore?" A CASUM-m Working Paper http://casumm.files.wordpress.com/2008/09/bhoomi-e-governance.pdf

workshop in the run up to the first global summit on the "Information Society" (WSIS) concluded: "it is a serious mistake to assume that they constitute a uniform process globally or share a common destination, rather than a variety of new processes each influencing and being influenced by the society in which they are taking place." [2]. Nor are such changes likely to be uniform within individual societies. In a country such as the UK, for example, informational behaviour of an individual might be expected to depend in large degree on their age, education, social and professional networks as well, possibly, on their class, gender and ethnic origin. This is of particular importance in the context of linked open data which, in its European and North American manifestations, is very much associated with ideas of openness of information and with the role of the engaged "hacktivitst" committed to interpreting and putting to public use the data made available. In our view, this model of open data leading to new analysis and real change remains largely aspirational even in those environments with the greatest quantity of linked open data and the highest density of "hactivists". It is a long way from being a realistic model for change in other environments as both the Web Foundation's "Open Government Data Feasibility Studies"[5] and Wright *et al.*'s study in India indicate. Whilst some such attitudes and skills may be shared especially amongst the most computer literate in many places, it is a mistake to assume that such ideas will always be seen as positive and beneficial. This is not intended as an argument against initiatives which promote the use of linked open data but to reinforce the point that, as with any form of "aid", what is done and how needs to be negotiated with those whom it is intended to benefit.

3 Information Ecology

The development information environment is multi-faceted and complex. It extends from the global to the entirely local, it covers a multitude of disciplines, cultures and languages. It operates on many levels. It can be understood as a series of overlapping information ecologies. The value of this metaphor is threefold. As predators form part of natural ecologies, so power relations impact on human ones. Ecologies are complex adaptive systems, so that an action in one part of them will have impacts elsewhere. Arguably, they are also systems which can benefit from being purposefully looked after: it is certainly possible to damage them, perhaps also to nurture them.

One feature of the power imbalances in the development field is the influence of large, well resourced institutions using large data sets from which to generate generalised policy approaches which are then applied to local situations. This process can ignore the specificities of local circumstances and limit the freedom of people working at more local levels to develop policy which seems appropriate to them. In a similar vein, such organisations (and many others) justify their roles by making extravagant claims about the value of their knowledge and

[5] In Chile and Ghana. See http://www.webfoundation.org/projects/ogd-feasibility/

implying that the main challenge of development is getting this knowledge to the people who need it. This was an explicit argument behind the World Bank's claim to be a "knowledge bank" in the late 1990s. Such attitudes lead to one-way information flows which are often not based on sufficient information about their intended recipients for the information to be of much value. As importantly, it can be hard for information about the realities on the ground, those realities which the whole effort is intended to improve, to get fed into and used in these policy making and communications processes.

Another aspect of the development information ecology is that although there is a certain attempt to maintain a common purpose of "co-operation for the common good" across the sector, most development organisations are in an increasingly competitive environment when it comes to seeking funds. This means both that they try to emphasise the origin and value of "their" information, for example by seeking to encourage direct traffic through their own web site and also that they often have few resources to invest in collaborative information initiatives. This, as well as the complexity of the quantities and ranges of information that is relevant to development, has resulted in a very fragmented information environment. With a few exceptions, mainly limited to certain well defined communities of interest, information resources are highly dispersed, metadata is poor, and the inbuilt algorithms of most search engines tend to recognise and reproduce the influence of the more powerful information providers.

If the above concentrates on the information ecology as it exists for professional development organisations working at international level, the situation for community level organisations and for individuals, marginalised or living in poverty is even worse. A fragmented information environment is inconvenient and inefficient for someone working in an office in Geneva, for someone relying on a pay per minute connection in an internet café in Africa it is impossible. Despite all the rhetoric about knowledge equating to power and the large ICT and communications budgets of the sector as a whole, there are incredibly few resources made available to support information processes which go beyond the idea of the passive recipient. "Passive recipients" are of no value to development. What is needed is for people to respond to information, understand it, adapt it to their own circumstances and to use it: or, in other words, to go through the process whereby information becomes knowledge. As has been argued elsewhere:

> Formal education undoubtedly helps, but so can many other forms of human interaction. People need to be able to validate information and to think through if and how any of it may be useful to them. In this context, and whatever other mechanisms may be available to help the process, connections with other people are essential, not only as sources of information, but also as means of validation, reflection and action. This is true for everyone - from the fraternity clubs of elite US universities to networks of the most poor and marginalised. "Ki raflé du ki amul yeeré wayé moy ki amul nit", as a Senegalese proverb has it, "the poor person is not the one without clothes but the one without anyone." [3]

The same chapter argues that most people, however marginalised, take part in a number of informational spaces, “some geared to family matters and social obligations, others related to work and income, some perhaps related to politics and governance and to faith” (p 134). These spaces, each of which will have their own rules and norms depending on their character and purpose, are seldom seen as part of the development knowledge ecology, but in fact they are key to knowledge being created and used by the people who’s actions determine whether or not any development actually takes place. At least some of these spaces are also open to interaction with external supporters but very little work has been done at this level to really understand what types of information are most useful to participants, to what extent and how digital platforms can be used, what issues of privacy and restricting access arise. It is an area of huge potential but also of many difficulties.

Value: in many countries much government data is inaccurate or incomplete, often out of date and, whatever the legal requirements may be, often hard to hold of[6]. In the absence of reliable free information, some have sought to develop enterprise models which either reward the generation of data and/or seek to charge for its use[7]. These models may offer better value than public information but, if the end-user has to pay for the data thus provided, are likely to be of less benefit to the very poor.

Security: in the wrong circumstances even the most apparently innocuous information can be misused. Ushahidi[8], in Kenya, rightly attracted a lot of attention by its ability to use crowd-sourcing to map emerging troublespots in the violence that erupted after that country’s elections in 2008. In doing so, the organisation, locally based and in touch with the many of the elements in Kenyan society working to end the violence, will have had to exercise judgement as to what information to make public. The same information, used differently, could have had disastrous consequences as was demonstrated by the use of Radio Mille Collines as an agent of the genocide in Rwanda.

Cost: the factor of cost is ever present. Amidst the marvelling at the growth of mobile phone usage in Africa, issues of inclusion and exclusion can easily get lost. As mobile phones become the main vehicle for communication across extended families, the poor are left with the choice of paying up - sometimes over 20% of their income - or removing themselves from their most life affirming community. Likewise, and in a cautionary tale for any proposed support of informational spaces, poor women in Zambia found themselves excluded from networks specifically set up to “empower” them (See, for example, [4]).

[6] In addition to the two Web Foundation reports cited in vii above, see Raman, N (2012) “Collecting data in Chennai City and the limits of openness” and Raman, B. 2012, “The Rhetoric of Transparency and its Reality: Transparent Territories, Opaque Power and Empowerment” both in Community Informatics 8:2, Special Issue: Community Informatics and Open Government Data http://ci-journal.net/index.php/ciej/issue/view/41

[7] See for example http://www.esoko.com/about/

[8] http://www.ushahidi.com/

Control: much current discourse about information, particularly in Europe and North America, concerns the desirability of its freedom. However other cultures may have other perspectives. For example, in New Zealand the indigenous Maori community believe that information about their ancestors belongs to them and is a vital part of their identity. The insistence of the colonial authorities in collecting and retaining information about individual families was a historic bone of contention which has been re-ignited by the idea that the government had the right to put such records on-line[9].

4 Implications for Developmental Linked Open Information Initiatives

All of the above pose a number of challenges for anyone wanting to develop linked information solutions in a way which goes beyond the basic connection to actually the poor and the marginalised to improve their conditions. It underlines the need for real understanding of local contexts, the desirability of working with local partners, and the value of participatory engagement with information providers and users if good choices are to be made about process, content and platform. Some of the challenges are of a general nature, others are more directly connected to issues of linking data and information.

- The need for caution in becoming over reliant the data of powerful institutions like the World Bank and OECD. It is hard to avoid using their data and it is of course positive that they are increasingly making their data openly available, so that it becomes a valuable resource to query. On the other hand, constant reference to these sources can appear to reinforce their dominant roles in the development information ecology and privilege the types of evidence favoured by global policy makers over evidence which enables learning from local contexts.
- In the same vein, there is a pressing need to make sure all relevant voices are heard and that data and other information from the grass roots becomes more visible
- The term “linked open information” rather than “data” is used in the subheading above because, if the aim is to support the information needs of the poor, there is no point in privileging data over other types of information. In the situations we are discussing, people are generally lacking many types of information. This poses the challenge, which to some extent has been explored in the IKM Emergent Programme[10], of using RDF and other data-linking tools to create links to metadata about other forms of information and to related social media discussions

[9] Plenary floor exchange at First Global Congress on Community Networking, Barcelona, Nov 2000

[10] See http://wiki.ikmemergent.net/index.php/Workspaces:1._Information_artefacts and http://wiki.ikmemergent.net/index.php/Workspaces:1:Linked_Open_Data

- Work on data itself, especially government data, in developing countries is likely to involve political and technical issues to improve its availability, accuracy and timeliness as much as technical development. It may also involve initiatives aimed at deriving data from crowd sourcing, a practice of which there are already examples from India and Kenya.
- The issues of the platform and the affordability of its use is crucial. Mobile Phones are an obvious choice but the cost of their use, particularly for higher bandwidth applications can be prohibitive. It may make sense to explore the potential for partnerships with resource centres, telecentres or, where they exist, libraries.
- Reusing schema can generate links simply and efficiently but again raises issues of politics and influence in whose schema are used. Standards also need to be established to avoid the development of an "official" development linked open data in which the powerful professional organisations exchange data and which excludes interaction with other sources of information and data
- Developing thesauri and ontologies capable of transcending the multi-lingual, multi-disciplinary realities of development will be a real challenge for what are referred to in the literature as "emergent ontologies", "heterogeneous ontologies" or "dynamic networked ontologies". The issue of "semantic interoperability" is also highlighted in the "Report on Open Government Data in India" cited above.

5 Conclusion

Finally, as the authors of a report on women's use of ICT in Mozambique conclude, there are limits to what can be achieved by the use of ICTs alone. Their impact can be immeasurably strengthened if they can be deployed alongside other more traditional tools of empowerment.

> "Literacy is key - without literacy there can be no empowerment, particularly for women and girls. We therefore strongly recommend the improvement of women's literacy in rural areas. We believe that women's literacy, combined with increased relevance of content, could result in computer-related ICT tools becoming an asset to women's pursuit of the means for survival and for control of their lives." [5]

What this conclusion suggests, apart from the point it so clearly makes, is that the process of linking information with meaningful change is far from straightforward and is unlikely to be taken very far by the development of one set of technologies, produced in isolation. Collaboration in more broadly based change efforts perhaps offers a better chance of greater impact as well as providing an opportunity for the discussions of assumptions, needs and options with people, rooted in their communities, to better understand their needs and create a process of mutual learning.

References

1. Wright, G, Prakash, P. Abraham, S. and Shah, N. 2011, “Report on Open Government Data in India”, Centre for Internet and Society, Bangalore. http://www.transparency-initiative.org/reports/open-government-data-study-india
2. UNRISD. (2005). Understanding Informational developments: a reflection on key research issues. Conference Report. p. 2.
3. Powell and Cummings, 2011, Missed understandings: How ICT might yet prompt change in Development in Zavazava and Perez-Chavolla (eds) “The Role of ICT in Advancing growth in Least Developed Countries: Trends, Challenges and Opportunities”. International Telecommunication Union (ITU), Geneva, http://www.itu.int/pub/D-LDC-ICTLDC.2011. p132
4. Abraham, K.B. (2009). The names in your address book: are mobile phone networks effective in advocating women’s rights in Zambia? In Buskens, I. and Webb, A. (Eds.), African women and ICTs: investigating technology, gender and empowerment (Chapter 9). IDRC, Ottowa. http://www.idrc.ca/EN/Resources/Publications/Pages/IDRCBookDetails.aspx?PublicationID=61
5. Macueve, G., Mandlate, J., Ginger, L., Gaster, P. & Macome, E. (2009). Women’s Use of information and communication technologies in Mozambique: a tool for empowerment. In Buskens, I. and Webb, A. (Eds.), African women and ICTs: investigating technology, gender and empowerment (Chapter 2, p. 30).

Downscaling Entity Registries for Ad-Hoc Environments

Philippe Cudré-Mauroux[1], Gianluca Demartini[1], Iliya Enchev[1], Christophe Guéret[2], and Benoit Perroud[3]*

[1] eXascale Infolab, University of Fribourg, Switzerland
{firstname.lastname}@unifr.ch

[2] Vrije Universiteit Amsterdam, the Netherlands
c.d.m.gueret@vu.nl

[3] VeriSign Inc., Fribourg, Switzerland
bperroud@verisign.com

Abstract. Web of Objects and Linked Data applications often assume that connectivity to data repositories and entity resolution services are always available. This may not be a valid assumption in many cases. Indeed, there are about 4.5 billion people in the world who have no or limited Web access. Many data-driven applications may have a critical impact on the life of those people, but are inaccessible to those populations due to the architecture of today's data registries. In this paper, we point out the limitations of current entity registries when deployed in poorly connected or ad-hoc environments. We then sketch new architectures based on IPV6, structured P2P networks and data replication for entity registries that could run in ad-hoc environments with limited Internet connectivity.

1 Introduction

Data registries are critical components of the Web architecture and are widely used in every-day web activities. For example, domain name registries are databases containing registered Internet domain names. They are necessary for all web users wishing to visit a website knowing its URL (e.g., hostname) rather than its IP address. Thanks to the Domain Name System (DNS) infrastructure, such information can be obtained by recursively resolving a domain name to an IP address.

Another example of registry is the Digital Object Architecture (DOA, see Section 2). It allows to assign unique identifiers to digital objects (e.g., scientific publications) which can then be univocally accessed by users. Their identity will thus last in time even if their physical location may change (similarly to the IP address associated to a domain name, which often changes over time).

In situations where data registries are not continuously accessible, the user experience can be strongly limited. As a basic example, if the DNS server used

* Authors are listed in alphabetical order.

by a client computer is not connected to the rest of the DNS hierarchy, then a very restricted set of Internet domains can be resolved to their IP addresses.

In addition to those traditional registries offering hash-table like functionalities, online infrastructures and applications are increasingly turning to more flexible registries containing information about general objects or entities (i.e., *object registries* and *entity registries*) to power data-driven applications. Emerging examples of that trend are DBPedia[4] and Freebase[5] which, given an entity identifier (e.g., a URI), provide semi-structured metadata about the entity. Another example are the registries used for the Web of Objects, which mediate information between networks of digital devices connecting to each other, enabling information publication or integration in sensor networks or smart building contexts for example. In those cases, the infrastructure needed is more complex than a traditional "Hostname-IP" DNS system and is closer to a global registry mapping unique identifiers to arbitrary structured data.

In situations where such registries are not continuously accessible, however, the user experience can be strongly limited. As a basic example, if the DNS server used by a client computer is not connected to the rest of the DNS hierarchy, then a very restricted set of Internet domains can be resolved to their IP addresses. In an object registry context, discontinued access to the registry typically results in the impossibility to publish data or issue object queries.

In this paper, we argue that data registries increasingly represent an essential part of today's Internet ecology (see Section 2 for a few examples), but that their current architecture precludes their use in many important contexts. For example, we can envision ad-hoc environment where the nodes self-organize without having access to third-party registries. In such transient and poorly-connected environments, nodes have a clear need to discover, connect, and exchange (structured) information with related entities locally and should be able to do so without resorting to any outside registry. Another interesting context is data-intensive object applications, where nodes have to discover and exchange data about very large numbers of entities and should be able to do so in a peer-to-peer manner whenever possible.

We propose in this paper different approaches to overcome the limitations of existing registry solutions for poorly connected environment or localized environments. First, we point out the limitations of current registry architectures, before proposing solutions that take into account the limited connectivity of the peers and enable the management of digital information in ad-hoc environments.

The remainder of this paper is structured as follows. In the rest of this section we motivate our work: we show why it is important to consider entity registries for ad-hoc environments and which are their benefits. In Section 2 we briefly describe existing architectures for data and entity registries. Section 3 highlights the problems of current solutions when applied to poorly connected or ad-hoc environments. Section 4 presents two alternative solutions to exploit entity registries based on IPV6 addresses and on P2P networks respectively. Finally, we conclude and discuss future work in Section 5.

[4] `http://dbpedia.org/`

[5] `http://www.freebase.com/`

1.1 Use case: Internet-less Mesh Networks

A rapidly increasing number of applications—such as open social applications, applications relying on governmental data (data.gov) or Linked Open Data[6]—assume an ubiquitous and continuous access to the Internet in order to power data sharing and data-driven applications. As pointed out in our previous work [6], this is not a safe assumption as there is more than half of the world's population who is cut-out from wide area networks. However, people who do not have access to the Internet still generate data and need to consume knowledge. For example, children who benefit from the "One Laptop Per Child" project, which aims at providing low-cost laptops (called XO) to developing countries, can be connected to each other. XO laptops are used at school while connected to school servers, but also at home where connectivity typically cannot be ensured. In such networks, access to centralized registries (e.g., DNS or global object registries) is intermittent at best. When functionalities based on such registries are needed (e.g., entity resolution or entity linking), they have to be emulated or replaced within the local network, and then possibly integrated to the centralized infrastructure when the link to the wide-area network is reestablished.

We can for example imagine a XO laptop connected to a server (e.g., at school) downloading some Web pages or Wikipedia articles. Afterwards, the laptop is moved to a different location with no Internet connectivity but with the possibility to connect to other XO laptops. This scenario enables the sharing of previously downloaded documents with others who had no possibility to obtain them from the Internet, and the local publication of new entities.

In this context, it is often important to identify entities in all documents to enable entity-centric document aggregation, semantic of faceted search. Such aggregations may, for example, support learning applications that present the set of documents users should read when they want to learn about a specific entity (e.g., "Malaria").

Extraction entities from HTML text may be performed automatically by tools running on the XO laptops or even manually exploiting crowdsourcing, which can address the problem of limited computational resources available [4]. After an entity occurrence is identified in the text, it has to be uniquely identified by associating it with the right entity ID in order to foster automated processing. XO users can also create new entities themselves (e.g., through data acquisition, document authoring or document enrichment), which then should be propagated and shared to the rest of the local community. For those different tasks, an entity registry containing a relevant set of entity descriptions and identifiers is necessary to streamline and support all data-driven applications in ad-hoc networks.

2 Entity Registries Today

There already exist many solutions to resolve entity names and/or get structured information about entities. One example of entity registry has been proposed in the context of the Okkam project[7], where the envisioned system stores a number

[6] http://linkeddata.org/

[7] http://www.okkam.org/

of entity profiles which can be accessed via keyword or structured queries. More recently, the popularity of Linked Data made it possible to connect large entity datasets and to make them accessible via SPARQL endpoints. Additionally to these, we can imagine the adoption of well established platforms like, for instance, DNS or DOA and to extend such technologies to entity registries.

2.1 DNS

The Domain Name System (DNS) is the system used on the Internet to resolve domain names to their corresponding IP addresses. Domain resolution works in a hierarchical manner; the top of the domain name space is served by so-called *root* name servers, pointing to *authoritative* name servers maintaining authoritative information for the top-level domains (a.k.a. "TLDs", such as ".ch" or ".com"). The authoritative name servers responsible for the TLDs point in turn to further name servers, responsible for second-level domains (e.g., "unifr.ch"), and so on and so forth to process each domain name label iteratively until the last iteration, which return the IP address of the domain name queried. In practice, domain names are often cached at various levels, for instance at the client-side, or at the level of the DNS server provided by the Internet Service Provider in order to limit the load on authoritative DNS servers.

Though originally not designed for this purpose, it is be possible to extend the current DNS infrastructure to create a full-fledged entity registry. In that context, we recently suggested an extension of the DNS [3] to serve authoritative metadata about Internet domains, leveraging both the DNS Text Record field ("DNS TXT") and new cryptographic features ("DNSSEC").

2.2 DOA

The aim of the Digital Object Architecture (DOA)[8] is the management of digital entities over potentially very long timeframes. There are three distinct components in DOA:

- the Resolution System (Handle System)
- the Digital Object Repository (DORepository)
- the Digital Object Registry (DORegistry)

The principal function of the Handle system is to map known identifiers into handle records, containing useful information about the digital object being identified (e.g., IP address, public key, URL etc.). Every identifier has two parts: a naming authority (or prefix) and a unique local name under the naming authority suffix, separated by "/" (e.g. "10.1045/january99-bearman").

The collection of all local names defined under a certain prefix defines the local handle namespace under that prefix (something similar to a root zone in the case of DNS). All the local namespaces (i.e., all prefixes) define the handle namespace and a prefix can be considered as a top level domain. More namespaces for Local Handle Services (LHS) can be defined in a hierarchical fashion

[8] `http://www.cnri.reston.va.us/doa.html`

under the Global Handle Registry (GHR), thus the Handle system provides a hierarchical service model.

The DORegistry provides services like browsing, searching, repository and federation for collections of digital objects that can be distributed across multiple sites including other DO Registries. A DO registry may manage metadata of objects from a certain repository. Another possibility is managing both metadata and actual digital object content stored by the registry, and a third scenario is managing metadata from multiple repositories. The DO Registry can be set for different types of metadata schemata and can be customized to provide different search, federation, handle registration, event management and other services.

The most popular application of this system is the use of Digital Object Identifiers (DOIs) to identify digital versions of written publications (e.g., scientific articles). Such identifiers, by means of an ID resolution, will lead not only to the digital object but also to its metadata. The important benefit of using DOIs are persistent citations (i.e., the location of the digital object may change over time but the identifiers will remain the same and its resolution will lead to the new location).

2.3 Linked Data

The Linked Data movement has been pushing towards publishing and interlinking public data in standard formats, which enables the automated discovery, management and integration of structured resources online. The adopted technology is based on HTTP URIs and RDF. The resolution of an entity given its identifier boils down to three steps in that context:

1. discovering the IP address where the HTTP URI is supposed to be hosted (for example using the DNS)
2. contacting the corresponding server and negotiating the content (e.g., to serve a human-readable version of the RDF data if the client is a Web browser)
3. retrieve the structured description of the entity over HTTP.

This process is commonly called entity *dereferencing* since it is similar to general URI dereferencing on the Web[9].

2.4 The Entity Name System

In the context of the Okkam project, the Entity Name System (ENS) [2] has been proposed. It is defined as a service to resolve entity names to their global identifiers (called Okkam IDs). This is made available thanks to a repository of entity profiles described as a set of attribute-value pairs, and a mix of matching components that select the correct identifiers for an entity request which may be submitted in the form of a structured (i.e., attribute-value) or unstructured (i.e., keyword) query.

[9] `http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14`

3 Limitations of Current Entity Registries

The existing data and entity registries are a critical asset for many Internet applications. However, all current architectures present limitations when we consider a situation with limited network connectivity or ad-hoc environments. If we imagine networks of computers with no connectivity to external Internet resources, it becomes clear that DNS-based entity registries typically do not work properly as only partial information will be cached in local and accessible DNS nodes. Moreover, data in the cache may not be up-to-date as DNS nodes may not frequently communicate with the rest of the DNS infrastructure. Similarly, clients may not be able to resolve DOI if the servers of the naming authority which issued such IDs is not reachable through the network. Entity registries like the Okkam ENS or LOD end-points are all based on centralized solutions which limit their reliability: in ad-hoc environments such central resolution points may or may not be reachable at a given point in time.

For the above mentioned reasons, we claim that a decentralized solution enriched with full replication of some seed content would be a better approach for an entity registry in ad-hoc environments. A decentralized system based, for example, on structured P2P networks can provide better connectivity thanks to cheap P2P communication and can tolerate the situation in which registry servers are not constantly available. Thus, in the following section, we envision solutions that consist of multiple distributed registry instances to optimize the availability of entity-registries in the ad-hoc environments.

4 Approaches to Downscale Entity Registries

We sketch out below three possible avenues that we have considered for downscaling entity registries in ad-hoc environments. Our solutions are based on IPV6, structured P2P and full replication technologies.

4.1 IPV6 Addresses

For many, the first solution to our problem may be provided by IPV6 technologies. IPV6 guarantees directly-resolvable IP addresses, irrespective of the network topology (e.g., rending NAT irrelevant). Hence, local entities could be made available using local IP addresses only. One could even envision some address ranges reserved for such a use.

While practical and simple, we however think that solutions directly built on bare IPV6 technologies are limited for two reasons. First, we are missing the indirection layer offered by the DNS or similar registries, which is often essential when migrating or evolving entities (migrating a resource to another IPs would require setting up some redirection mechanism between the previous identifier and the new one). Second, from a networking perspective, this would boil down to mixing two separate layers, namely the networking and the application layers, which would certainly never be endorsed by the central networking companies and bodies.

4.2 Structured P2P Proxies

P2P technologies are another potential solution to our problem. Distributed Hash-Tables (DHTs), such as Chord[10] or our P-Grid system [1], provide decentralized, scalable hash-table-like functionalities that could be used to store entity identifiers as well as related meta information in ad-hoc environment. Through dynamic load-balancing and replication, those networks provide fault-tolerant and efficient networking primitives where arbitrary requests can typically be resolved in $O(log(N))$ messages, where N is the number of nodes in the P2P network, from any entry point to the network.

P2P technologies have been proposed in the past to enhance or supplant DNS infrastructures[11], most often to provide an alternative to ICANN or to support P2P file exchange. Such efforts had limited success so far. We think that the CoralCDN [5] system, in particular, is relevant to our scenario, since it takes advantage of highly efficient P2P mechanisms (P2P DNS and distributed sloppy hash tables) to create P2P content distribution networks. It however suffers from several severe limitations in our context, including some reliance on high-bandwidth and wide-area connectivity and the lack of any mechanism to serve structured entity content.

One could hence consider a DHT-based CDN as a starting point to solve our problem, and enhance the infrastructure with a native entity storage system (such as our recent dipLODocus system [7]), and with semi-structured capabilities (e.g., supporting declarative queries). Using such P2P infrastructures, we could for example explore the best possible way to support both entity publication and entity search in ad-hoc networks.

4.3 Entity Nucleus and Lazy Replication

Even though supporting a full-fledged entity registry in ad-hoc settings is a necessity, there are many cases where some of the nodes might connect to centralized infrastructures intermittently. Thus, we believe that it is also essential to be able to cache authoritative or centralized information, and to be able to dynamically synchronize data with such infrastructures.

Depending on the context of the application, some core nucleus of the entity data can be identified (e.g., DBPedia entity data for LOD). In such a case and if the entity nucleus can be pre-installed on each node, then many of the entity operations can be resolved locally without resorting to any third-party infrastructure.

Networked search and updates, however, still require distributed mechanisms to be resolved. If such operations are deemed relatively infrequent, then semi-structured P2P technologies like those described above can be applied: both distributed queries and updates can for instance be (lazily) propagated or broadcasted across the network at regular intervals.

[10] `http://pdos.csail.mit.edu/chord/`

[11] see for instance `http://blogs.computerworld.com/17444/p2p_dns_to_take_on_icann_after_us_domain_seizures`

5 Conclusions

Current entity registry solutions are often based on global hierarchies or centralized online directories. Such solutions are inapplicable to many contexts, including ad-hoc networks and environments that have limited access to a wide-area connection. In this paper, we described some of the key problems related to using current entity registries in an ad-hoc context and suggested a few possible alternatives. Three potential solutions were specifically sketched, based on IPV6 technologies, scalable and structured P2P technologies, and (lazy) content replication. We now plan to implement, test and combine both current entity registry solutions and our new alternatives to determine in practice which architecture is most useful given a specific ad-hoc environment. As a start, we plan to focus on the OLPC XO context to provide a working solution and enable Open Data and Linked Entity applications for the billions of people who are currently cut-out of wide area networks.

6 Acknowledgment

This work was supported (in part) by the Swiss National Science Foundation under grant number PP00P2_128459.

References

1. Karl Aberer, Philippe Cudré-Mauroux, Anwitaman Datta, Zoran Despotovic, Manfred Hauswirth, Magdalena Punceva, and Roman Schmidt. P-grid: A self-organizing structured p2p system. *ACM SIGMOD Record*, 32(3), 2003.
2. Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and Antonio Mana. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, number CSS-ICSC 2008-4-28-25 in CSS-ICSC, pages 554–561. IEEE Computer Society, August 2008.
3. Philippe Cudré-Mauroux, Gianluca Demartini, Djellel Eddine Difallah, Ahmed Elsayed Mostafa, Vincenzo Russo, and Matthew Thomas. A Demonstration of DNS^3: a Semantic-Aware DNS Service. In *ISWC 2011*.
4. Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudre-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In *International World Wide Web Conference (WWW)*, 2012.
5. Michael J. Freedman, Eric Freudenthal, and David Mazières. Democratizing content publication with coral. In *Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation - Volume 1*, NSDI'04, pages 18–18, Berkeley, CA, USA, 2004. USENIX Association.
6. Christophe Guéret, Stefan Schlobach, Victor De Boer, Anna Bon, and Hans Akkermans. Is data sharing the privilege of a few? Bringing Linked Data to those without the Web. In *ISWC 2011 - Outrageous Ideas.*
7. Marcin Wylot, Jigé Pont, Mariusz Wisniewski, and Philippe Cudré-Mauroux. dipLODocus[RDF] - Short and Long-Tail RDF Analytics for Massive Webs of Data. In *International Semantic Web Conference*, pages 778–793, 2011.