A global Entity Name System (ENS) for data ecosystems

Paolo Bouquet
OKKAM srl – University of Trento
Via Segantini 23
38122 Trento (TN), Italy
+39 0461 283649
bouquet@okkam.it

Andrea Molinari
OKKAM srl
Via Segantini 23
38122 Trento (TN), Italy
+39 0461 283649
molinari@okkam.it

ABSTRACT

After decades of schema-centric research on data management and integration, the evolution of data on the web and the adoption of resource-based models seem to have shifted the focus towards an entity-centric approach. Our thesis is that the missing element to achieve the full potential of this approach is the development of what we call an Entity Name System (ENS), namely a system which provides a collection of general services for managing the lifecycle of globally unique identifiers in an open and decentralized environment. The claim is that this system can indeed play the coordination role that the DNS played for the document-centric development of the current web.

1. INTRODUCTION

Research and development on data integration has been traditionally dominated by work on integrating schemas from different data sources. This is true not only for databases, but was inherited also by the semantic web community, where the problem is typically addressed under the title of ontology matching and mapping.

However, two relatively recent developments showed that there was the need for a different approach:

- on the one hand, the RDF data model is designed around the concept that the building blocks of data publication on the web are the identifiers (URIs) for the resources one needs to name. In other words, instead of starting from a schema, the RDF data model starts from naming "things" (a superclass of what we call "entities", like people, organizations, locations, events, products, etc.), which are then connected through binary relationships (again identified by a URI), which are eventually defined in an explicit vocabulary or ontology;
- on the other hand, the large-scale adoption of the Linked Data principles for the publication of data on the web forced again data producers to focus on the names of resources about which data are provided and in particular on the identity statements which allow to connect information about the same entity from different datasets. This means that, before addressing the issue of what is said about an entity, we need to address the issue of what entities one is talking about, as this is a key element to merge data from different sources about the same thing.

Proceedings of the VLDB Endowment, Vol. 6, No. 11

The argument of this short paper is that to achieve the full potential of this entity-centric approach for querying, retrieving and integrating data in networked data production settings, we need to develop a coordinated solution to naming entities, in analogy to what was done exactly 30 year ago when the DNS for domain names on the Internet was made available. To this end, we will shortly outline the concept and architecture of the Entity Name System (ENS) as it was developed in the OKKAM project[1]. We will also provide a few examples of applications which can support the claim.

2. ENTITY NAME SYSTEM (ENS)

As part of the OKKAM project (2008-2010), a first instance of an Entity Name System (ENS) was developed based on the following general requirements:

1. Maintaining a large-scale *entity repository* which can ensure the uniqueness and persistence of the binding between a token (entity identifier) and a singular entity.
2. Making these identifiers *searchable* and easily *retrievable* by humans and – much more important – by any application which needs them for data production and publication.
3. Storing known mappings between the ENS identifier and any other available identifier for the same entity.
4. Supporting a few basic operations (create, edit, merge, split) which enable the lifecycle management of each identifier.

The details of these components and the available APIs are described and documented in past papers [1] and in the dedicated web portal[2]. In a recent paper [2], we also showed how the initial OKKAM's approach has been modified to support the distinction between the pure identifier (*token*) and the multiple services which can return data about it (*resolvers*). This was done to offer in a single package a solution for some apparently conflicting requirements on identifiers coming from the linked data and the persistent identifiers communities.

Here we want to stress only a few general features of the ENS. First of all, the ENS should not be thought of as a kind of entity-base (*à la* DBpedia or *à la* Freebase), like the DNS cannot be

[1] See http://project.okkam.org/

[2] See http://api.okkam.org/

confused with an index (or a search engine) for web contents: the only purpose of an ENS is to store the binding between a token and an entity and make it usable; information about entities will always reside outside the ENS, where it belongs to. Second, the ENS can only be the result of a fully distributed and social process, where millions of users create new identifiers and reuse existing ones in their data. Third, using a ENS identifiers in web data must not create a dependence between external applications and the ENS (or even worse a single point of failure): once the id for an entity is retrieved from the ENS and used in local data, the data lifecycle should be completely independent from the ENS itself. Finally, the ENS must be managed by an independent authority, which is not in competition with any of the potential users of the ENS identifiers[3].

3. EXAMPLES OF APPLICATIONS

3.1 The web of data: Sig.ma

The first application of the ENS concept was the development of the Sig.ma information mashup engine[4]. The idea is to allow web users to search and navigate the web of data not by file (or dataset), but by entity. In short, if you search by keyword (e.g. "Paolo Bouquet") or by ENS id (e.g. "ok200706301185791252056"), what you get is a profile which is built runtime by querying all the data sources on the web which are indexed by Sindice[5]. The ENS plays two roles when the query is by ID: it makes sure that the token "ok200706301185791252056" is persistently associated to the same person; and returns the list of known URIs which are stored as equivalent on the ENS for the entity "ok200706301185791252056" (this is used for query expansion).

3.2 Tax investigation

A second application is for data mashup in the context of tax investigation. In this project, which is run in collaboration with the regional tax agency in Trentino (Italy), the concept of the ENS is used for creating a single, virtual knowledge base out of hundreds of databases which are available to tax inspectors. A local ENS was started and then the entities named in each database were matched and aligned with the ENS identifiers via a process of automatic entity matching. The outcome of the process (which of course includes the alignment with a common domain ontology for classes and properties) is a system where data can be navigated and searched by entity as if we information was stored in a single graph.

This entity-centric approach has proven to be very effective and – most importantly – is extremely flexible when the need arises of integrating new data sources into the knowledge base. In addition, as a side effect, the entity repository which was built from data can be used by the regional public administration to develop other applications, like a social policy planner, tax redistribution plans, and the like.

[3] Indeed, the OKKAM's ENS will be run as a public Trust, and OKKAM s.r.l. will act as the Trustee under the surveillance of an international Board of Protectors.

[4] See http://sig.ma/

[5] See http://sindice.com/

3.3 Data marketplace

As part of a EU co-funded project called DOPA[6], we are integrating our ENS technology into a platform for creating a data marketplace with other important partners. The ENS will be used to cross-link data from different sources in a financial and economic domain, and create value added for users who have access to the single data sources but do not have the resources to discover and use the links between them.

3.4 ObjectLinks

A last example of how we use this entity-centric vision is in the development of ObjectLinks[7], a platform which can be used to "link" data and services to real world objects through tags (QR codes, NFC tags, etc.). For simplicity, imagine a QR code on a bus stop in a city; and imagine that the bus stop is assigned a persistent ID which can be used to enrich data in other information systems (timetables, real time positioning systems for buses, relevant event in the area, ticketing, etc.). Then the QR code, thanks to the dynamic features of ObjectLinks, can become a terminal where these data can be integrated and used in applications which convey useful services to customers in a highly flexible way. In other words, simple physical objects (business cards, product envelops, brochures, books, newspapers, buildings, …) become first order citizens in the development of data intensive smart city services and applications.

4. CONCLUSIONS

In the last five years, we were told many times that a global ENS is not possible (perhaps not even desirable), or that the problem will be solved bottom up as people converge on the usage of DBpedia URIs or a persistent ID platform (e.g. DOI or ORCID) or even Facebook accounts for managing IDs for specific categories of entities. Our position is that none of these solutions will work for large-scale, decentralized data ecosystems, as they do not offer a clear separation between the basilar ID management services of the ENS and other value added (for sure more profitable!) services built on top of the IDs themselves. This mix leads to unsustainable costs and eventually social rejection as a neutral ID management system. That's why, even though the OKKAM's ENS project may fail, we believe that something like the ENS will be eventually realized to fully exploit the potential of data which are currently made available on the web and in large-scale information systems.

5. REFERENCES

[1] Bouquet P., Stoermer H., Niederee C. and Mana A. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the Second IEEE International Conference on Semantic Computing* (4-7 August, 2008). IEEE Computer Society Press, p. 554-561.

[2] Bortoli S., Bazzanella B. and Bouquet P. Can Persistent Identifiers Be Cool. In *Proceedings of the 8th International Digital Curation Conference (IDCC-2013)* (Amsterdam, The Netherlands, January 14-17, 2013

[6] See http://www.dopa-project.eu/

[7] See http://ol.objectlinks.biz/ for a simple web frontend of the platform