OKKAM: Enabling an Entity Name System for the Semantic Web[*]

Paolo Bouquet & Heiko Stoermer

Dipartimento di Ingegneria e Scienza dell'Informazione – University of Trento
Via Sommarive, 14 – 38050 Trento, Italy
{bouquet, stoermer}@disi.unitn.it

Abstract. In this paper, we argue that realizing the Linked Data concept would greatly benefit from the availability of a global service which may enable the systematic reuse of URIs across independently produced Semantic Web contents. Such a service, which we call an *Entity Name System* (ENS), might play for the Semantic Web the role that the DNS played for interlinking hypertexts on the Web. The key idea behind the ENS is that any tool/application for creating content (not only Semantic Web) may be empowered through simple plugins or extensions to interact with a network of ENS servers. Here we show how this has been done in a few prototypes for MS Word, Protégé and an editor for FOAF profiles.

1 Introduction

In a note from 1998, Tim Berners-Lee described the grand vision of the Semantic Web as follows:

> Knowledge representation is a field which is currently seems to have the reputation of being initially interesting, but which did not seem to shake the world to the extent that some of its proponents hoped. It made sense but was of limited use on a small scale, but never made it to the large scale. This is exactly the state which the hypertext field was in before the Web [...]. The Semantic Web is what we will get if we perform the same globalization process to Knowledge Representation that the Web initially did to Hypertext.

We understand this parallel as follows. Like the WWW provided a global space for the seamless integration of small hypertexts (or local "webs of documents") into a global, open, decentralized and scalable publication space; so the Semantic Web should provide a global space for the seamless integration of small knowledge bases (or local "semantic webs") into a global, open, decentralized and scalable knowledge space.

[*] This work is a revision of a paper currently under review at the 5th European Semantic Web Conference (ESWC2008).

Today, as a result of many independent research efforts and commercial initiatives, relatively large and important knowledge repositories (like DBpedia.org, GeoNames, DBLP) have been made available which actually are (or can be easily transformed into) local "semantic webs", namely graphs of resources connected through properties which are defined in some schema or vocabulary. Some of these repositories have already been interlinked through the Linked Data initiative (see `linkeddata.org`). However, it is difficult to claim that the interlinking of Semantic Web data is proceeding as fast as expected in a fully decentralized and open process. Why?

The argument we put forward in this paper is the following. On the one hand, the interlinking of local "webs of documents" into the WWW was largely made possible by a key enabling factor: the introduction of a global reference and addressing mechanism for locating and retrieving resources. This mechanism heavily depends on the availability of a global service, the DNS, which can resolve resource identifiers (URLs) into physical locations on the Internet. This is how one can be sure that, for example, a `href` link to a Web resource will be always resolved to the appropriate location on the Internet. On the other hand, the interlinking of local "semantic webs" is based on (i) the introduction of uniform identifiers for anything which can be named, including concrete entities (people, geographical locations, events, artifacts, etc.) and abstract objects (concepts, relations, ontologies, etc.); and (ii) on the introduction of a generalized notion of link from simple hypertext links to any binary relation between resources. Whenever a statement is made about an entity e_1 in any location of the network, then such a statement is in principle connected with any other statement made about the same entity elsewhere and independently, provided that (a) the same identifier (URI) is consistently used for it; or (b) an explicit mapping between the different URIs referring to the same entity is available. However, as to the former, to date no scalable and open service is available to make possible and to support a consistent reuse of identifiers for entities; as to the latter option, it seems to us that it cannot easily scale to a scenario in which billions of entities are identified through thousands of different URIs, as it would require a huge number of `owl:SameAs` statements and their use would be computationally too expensive. This situation undermines the practical possibility of a seamless interlinking of local knowledge into the global knowledge space envisage in the initial quote.

In this paper, we will try to defend the view that the practical realization of the grand vision of the Semantic Web as a huge graph of interlinked data would be much easier and faster if we could count on a service which, by analogy with the DNS, we call an *Entity Name System* (ENS), namely a service which stores and makes available for reuse URIs for any type of entity in a fully decentralized and open knowledge publication space. In this paper, we first describe how such a ENS should work; then we discuss the main issues and challenges associated with the design and implementation of a scalable and sustainable ENS; we present some preliminary data on the advantages of this approach for information integration on the Semantic Web, and briefly sketch some simple example of applications which are enabled to interact with the ENS for creating new content ready for being interlinked with pre-existing content.

2 An ordinary day on the Semantic Web

Imagine an ordinary day on the Semantic Web:

- the University of Trento exports in RDF its bibliographic database;
- the WWW2008 conference organizers make available the metadata about authors and participants as part of the Semantic Web Conference (SWC) initiative;
- participants at WWW2008 upload their pictures on `http://www.flickr.com/`, and tag them;
- some participants atWWW2008 attend a talk on FOAF and decide to create their FOAF profile at `http://www.ldodds.com/foaf/foaf-a-matic` and publish it on their web server;
- …

At the end of the day, a lot of related material has been created. In principle, the newly created RDF content should allow Semantic Web programs to answer questions like: “Find me which of my friends is attending WWW2008”, “Find me pictures of Mike’s friends who are attending WWW2008”, “Find me the papers published by people of the University of Trento (or their friends) accepted at WWW2008”, and so on. But, unfortunately, this can’t be done. And the reason is that any time an entity (Mike, WWW2008, University of Trento, …) is mentioned in one of the data sets, they are referred to through a different URI. And this does not allow browsing multiple RDF graphs based on the fact that the same resource is referred to by the same URI.

This scenario is quite typical of how Semantic Web content is produced today. Nearly like hypertexts in the pre-WWW era, every tool for creating semantic content mints new URIs for any resource, and this makes really difficult the boot-strap of the global knowledge space envisaged by Tim Berners-Lee. More and more attention is paid in reusing existing vocabularies or ontologies, but statements about specific resources (instances, individuals) cannot be automatically integrated, as there is nothing practically supporting the (desirable) practice of using a single global URIs for every resource, and reusing it whenever a new statement about it is made through some content creation application.

The Linked Data initiative is a very relevant and promising attempt to make this vision real. However, we believe that we need something more structural to support the process of interlinking RDF data. The ENS we will discuss in the next section is our proposed approach and solution for addressing this issue in a systematic and scalable way.

3 ENS: architecture

Our current prototype implementation of an ENS service is called OKKAM and represents one node in a federated architecture, which is depicted as a cloud in the center of Fig. 1. The aim of the OKKAM prototype is to provide a basic set of ENS functionality, i.e. searching for entities, adding new entities and creating new identifiers. The identifiers that OKKAM issues are *absolute URIs* in the sense of RFC3986 [1], which makes them viable global identifiers for use in all current (Semantic) Web data sources; they furthermore are valid UUIDs, i.e. identifiers that guarantee uniqueness across space and time[1], which prevents accidental generation of duplicates and thus also enables their use as primary keys e.g. in relational data sources.

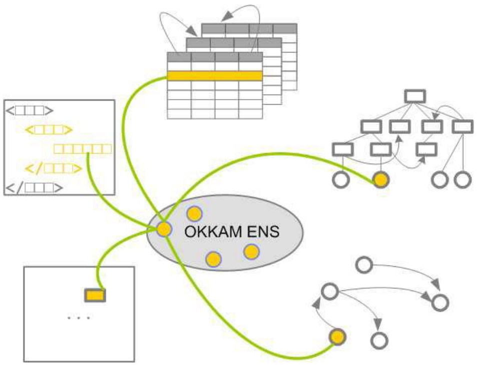


Fig. 1. The ENS providing entity identifiers across system boundaries.

What is illustrated in Fig. 1, and currently implemented as a single node, is planned to become a distributed system that is fully in line with the distributed nature of the (Semantic) Web. It is important to note that what we are propagating is an *entity-centric* approach, not an ENS-centric approach; this means, for example, that data sources which have issued their entities with OKKAM identifiers will continue to be integratable on the entity level, disregarding the existence or availability of an ENS server.

A critical feature of an ENS is to provide means for searching for the identifier of an entity. This step is strictly connected to the algorithm that supports the population of the system's repository with new entities. Indeed, when a query is submitted to the system, it has to decide if the query corresponds to an entity already stored (and return the information about it) or if a new entity has to be generated.

Figure 2 illustrates the standard use-case for the *okkamization*[2] of content, namely to query OKKAM for the existence of the entity at hand, and the re-use of a global identifier for this entity. This would usually be achieved through functionality provided by a client application, such as FOAF-O-MATIC or OKKAM*4P* (see Sect. 4) which accesses the OKKAM API, and presents (if available) a list of top candidates which match the description for the entity provided within the client application. If

[1] See `http://java.sun.com/j2se/1.5.0/docs/api/java/util/UUID.html` for details

[2] We call *okkamization* the process of assigning an OKKAM identifier to an entity that is being annotated in any kind of content, such as an OWL/RDF ontology, an XML file, or a database, to make the entity globally identifiable.

the entity is among these candidates, the client agent (human or software) uses the associated OKKAM identifier in the respective information object(s) *instead* of a local identifier. If the entity cannot be found, the client application can create a new entry for this entity in OKKAM and thus cause an identifier for the entity to be issued and used as described before.

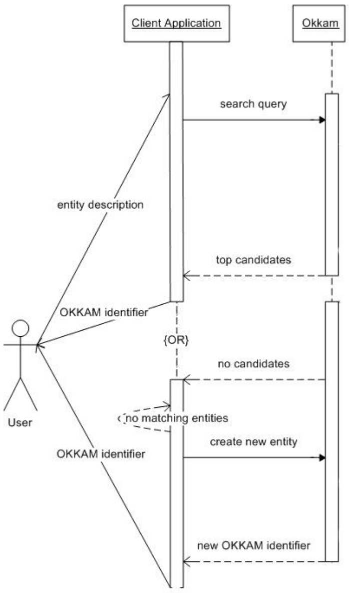


Fig. 2. Sequence diagram of a standard interaction with OKKAM.

4 Three ENS-enabled Applications

To illustrate the viability and the usefulness of the approach, we have developed three exemplary prototypes that have been strategically selected from the area of content creation, and serve as means to achieve an *a priory* alignment of identifiers that we propagate in our approach. The reason for this selection was the fact that the success of ENS approach entirely depends on the availability of a critical mass of OKKAMized content, and of course on the availability of tools for the creation of such content.

The three prototypes presented here are all available at `http://www.okkam.org` for download and test.

4.1 Okkam*4P*

The first tool is called Okkam*4P* [5], a plugin for the widely-used ontology editor Protégé. This plugin enables the creator of an ontology to issue individuals with identifiers from OKKAM, instead of assigning local identifiers that bear the risk of non-uniqueness on a global scale. The choice for this tool was made based on two criteria, namely the target audience being rather ‘expert’ users of the Semantic Web, and, secondly, the very wide usage of the Protégé editor, which makes it a promising candidate for a rapid distribution of the tool.

The plugin essentially supports the assignment of a global unique identifier (the “OKKAM ID”) to a newly created individual, rather than relying on manual input of the user or the standard automatic mechanism of Protégé. To this end, it implements the use-case illustrated in Fig. 2: based on the data about an individual that are already provided in the KB developed by the user, it queries OKKAM to see whether an identifier already exists which can be assigned to the new created individual, otherwise a new identifier would be created and used.

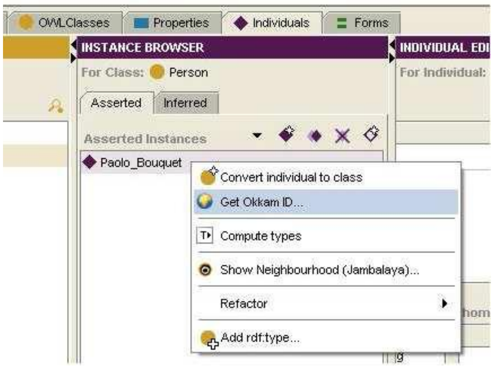


Fig. 3. Assigning a global identifier to an individual.

Access to the plugin is given through the context-menu of an individual, as depicted in Fig. 3. The plugin then guides the user through the search and selection process and finally replaces the local identifier for the entity with the one retrieved from the ENS. The result is an OWL ontology that is equipped with globally unique and re-usable identifiers and thus enables vastly simplified, automatic integrations with high precision. The plugin is available at the following URL: http://www.okkam.org/projects/okkam4p.

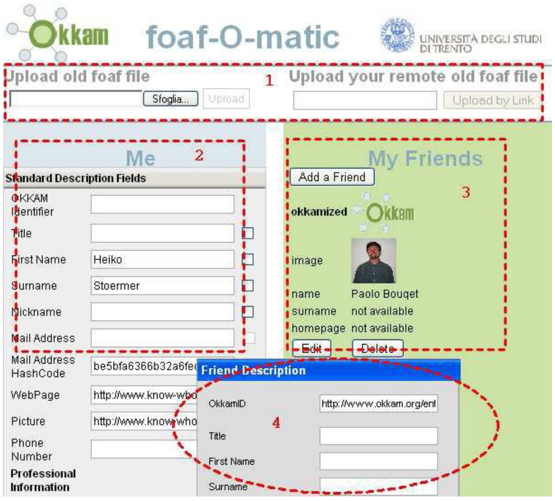
4.2 Foaf-O-Matic

The second application is called FOAF-O-MATIC [2], a WWW-based service for the creation of OKKAMized FOAF[3] profiles. Indeed, FOAF is in our opinion one of the few real success stories of the Semantic Web so far, as it is one of the few applications that really contributed to the creation of a non-toy amount of RDF data, with the special restriction that the agreement on URIs for persons is extremely low [6]. As content creation tools for FOAF are mostly rather prototypical, we decided to

[3] http://www.foaf-project.org

create a completely new application that both serves the user due to state-of-the art technology and at the same time creates OKKAMized FOAF profiles.

As we have explicated in [2], what is currently missing from FOAF is a reliable and pervasive way to identify "friends". The aim of creating the FOAF-O-MATIC application is not only to provide an alternative to the well-known foaf-a-matic application[4]. The focus of the new application is to allow users to integrate OKKAM identifiers within their FOAF document in a user-friendly way. In this way, it will be possible to merge more precisely a wider number of FOAF graphs describing a person's social networks, enhancing the integration of information and reach more easily the goal of the FOAF initiative.



OKKAM Identifier
Title □
First Name Heiko □
Surname Stoermer □
Nickname □
Mail Address
Mail Address HashCode be5bfa6366b32a6fe
WebPage http://www.know-who
Picture http://www.know-who
Phone Number

name Paolo Bouqet
surname not available
homepage not available

OkkamID http://www.okkam.org/ent
Title
First Name
Surname

Fig. 4. FOAF-O-MATIC The main interface of FOAF-O-MATIC.

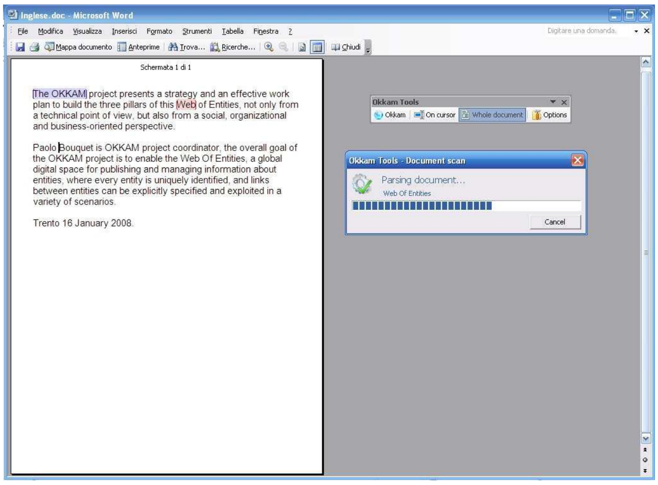
A view of the application layout is given in Figure 4: it includes functions to re-use existing FOAF profiles (1), a form for describing oneself (2), the list of friends (3), and the form for adding friends (4) which initiates the ENS search process. The application is deployed and usable at the following URL: `http://www.okkam.org/foaf-O-matic`.

4.3 Okkam4MSW

The last ENS enabled application we present is a tool called OKKAM4MSW, a MSWord plugin for the globally unique identification of individuals in MSWord. Strictly speaking, this tool is not a Web (or Semantic Web) tool; however, it illustrates how the OKKAM concept may go beyond RDF/OWL content and support also the annotation of entities in non-structured formats/contents with global identifiers, this way (i) making possible a progress in traditional information retrieval with standard search engines, (ii) making easier the task of named entity recognition

[4] `http://www.ldodds.com/foaf/foaf-a-matic`

on information extraction and (iii) enabling the integration with RDF/OWL knowledge.


Inglese.doc - Microsoft Word
File Modifica Visualizza Inserisci Formato Strumenti Tabella Finestra ?
Digitare una domanda.
Mappa documento Anteprime Trova... Ricerche... Chiudi
Okkam Tools
Okkam On cursor Whole document Options
Okkam Tools - Document scan
Cancel

Schermata 1 di 1

The OKKAM project presents a strategy and an effective work plan to build the three pillars of this Web of Entities, not only from a technical point of view, but also from a social, organizational and business-oriented perspective.

Paolo Bouquet is OKKAM project coordinator, the overall goal of the OKKAM project is to enable the Web Of Entities, a global digital space for publishing and managing information about entities, where every entity is uniquely identified, and links between entities can be explicitly specified and exploited in a variety of scenarios.

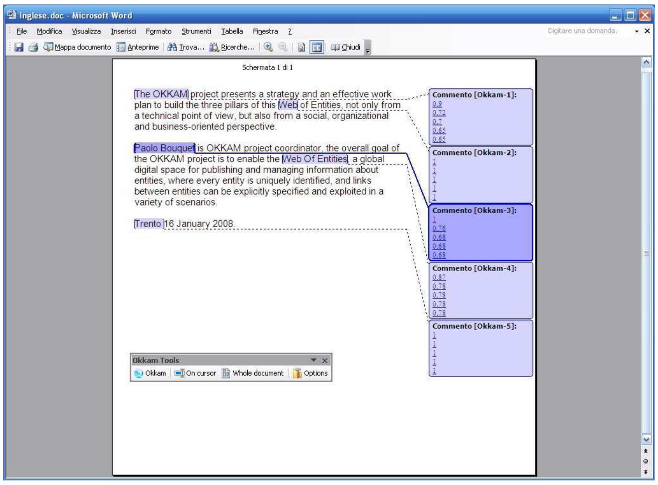
Trento 16 January 2008.

Parsing document...
Web Of Entities

Fig. 5. The search for recognized entities.

The great circulation of MS Word as a word processor is the main reason for the choice of implementing OKKAM4MSW. It allows for the unique identification of entities inside text documents created with MS Word. NLP and semantic technologies are used to detect the mention of an entity in the text and to extract contextual information that enables the matching decision within the OKKAM entity repository.

The plugin works as follows. The user writes a text using MS Word. The plugin identifies precisely or automatically entities by means of a thorough processing and linguistic disambiguation, which includes morphological, grammar, syntactic and semantic analyses of the text. All fundamental info for the disambiguation process, i.e. the whole system knowledge, is represented as a concept-based semantic network. The plugin can be accessed by means of a MS Word-integrated toolbar. Once the entity is identified, the plugin returns its unique identifier from the ENS – if found (Figure 5); otherwise a new identifier is generated. The analysis can be performed in real-time during document typing, or offline. The result is the option of uniquely choosing the entity which need be inserted into the text, or the a posteriori analysis of the whole document entities (Figure 6). The OKKAM URIs are stored in the MS Word document, and becomes available for any use, including indexing from search engines.


Inglese.doc - Microsoft Word
File Modifica Visualizza Inserisci Formato Strumenti Tabella Finestra ?
Digitare una domanda.
Mappa documento Anteprime Trova... Ricerche... Chiudi
Schermata 1 di 1
Okkam Tools
Okkam On cursor Whole document Options

The OKKAM project presents a strategy and an effective work plan to build the three pillars of this Web of Entities, not only from a technical point of view, but also from a social, organizational and business-oriented perspective.

Paolo Bouquet is OKKAM project coordinator, the overall goal of the OKKAM project is to enable the Web Of Entities, a global digital space for publishing and managing information about entities, where every entity is uniquely identified, and links between entities can be explicitly specified and exploited in a variety of scenarios.

Trento 16 January 2008.

Commento [Okkam-1]:
0.9
0.72
0.7
0.65
0.65

Commento [Okkam-2]:
1
1
1
1
1

Commento [Okkam-3]:
1
0.76
0.68
0.68
0.68

Commento [Okkam-4]:
0.87
0.78
0.78
0.78
0.78

Commento [Okkam-5]:
1
1
1
1
1

Fig. 6. The outcome of the search for entities.

5 Related Work

There are currently two major approaches which can be considered relevant for the topic described in this paper.

The first is, of course, the *Linking Open Data Initiative*[5], which has the goal to "connect related data that wasn't previously linked". The main approach pursued by the initiative is to establish `owl:sameAs` statements between resources in RDF. While the community has made a huge effort to link a significant amount of data, their approach depends on specialized, data source-dependent heuristics[6] to establish the `owl:sameAs` statements between resources, and it requires the statements to be *stored* somewhere, along with the data. As we said in the introduction our main concerns with this approach (without the ENS) are the following: first, in most Web scenarios, we don't see standard web users making an effort to create `owl:SameAs` statements for their data; second, an error in an identity statement might have long ramifications on the entire Web of Data; finally, reasoning over massive numbers of owl:sameAs statements in distributed ontologies is computationally a complex and highly expensive task, which may lead to the conclusion that these linked data are more suitable for *browsing* than for reasoning or querying, and thus do not fully attempt to realize the vision of the Semantic Web as a large, distributed knowledge base.

[5] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

[6] `http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining`

The second is presented in Jaffri et al. [7]. In their work resulting from the ReSIST project, these authors recently came to a conclusion similar to the one we expressed in [3, 4], namely that the problem of proliferation of identifiers and the resulting coreference issues should be addressed on an infrastructural level. As a solution, they propose what they call a *Consistent Reference Service*. While we share this general view, their point about URI potentially changing "meaning" depending on the context in which they are used, is philosophically disputable: the fact that several entities might be *named* in the same way ("Spain" the football team, "Spain" the geographic location) must not lead to the conclusion that they can be considered *the same* under certain circumstances[7]. Furthermore, their implementation of "coreference bundles" which establish identity between entities, are in fact very similar to a collection of `owl:sameAs` statements, that we discussed below.

6 Challenges and conclusions

In the paper, we presented the idea and the results of a test on ontology integration with a simple prototype of the ENS. However, designing, implementing and making available the ENS on a global scale involves some very difficult scientific and technological challenges. Here we list some challenges, and discuss how we think to address them in the FP7 project OKKAM.

The first big challenge has to do with the idea of *recognizing* that an entity named in some content (e.g. in an RDF graph) is the same as an entity stored in the ENS repository. This entity matching problem, which is obviously related to well-known problems is several disciplines (e.g. named entity recognition, coreference, object consolidation, entity resolution), is one of the most important reasons why developing a working and scalable ENS is intrinsically more difficult than the DNS. The solution which we will explore in the project is based on two ideas:

- on the one hand, any entity will be stored in a ENS server with its URI and a non fixed set of attribute-value pairs. This "profile" of the entity is not meant to be a source of correct and up-to-date information about the entity, but as a minimal collection of information that a human or a program can use to identify an entity, or to distinguish between two entities. The matching algorithm we used for the test in this paper is quote simple, and is based on string similarity algorithms; still, the results are encouraging. We plan to refine this method with heuristic rules, possibly specialized for broad categories of entities (e.g. people, organizations, locations, events, products, etc.);
- on the other hand, we will investigate how we can automatically build a contextual profile for entities named in content specified in different formats (e.g. text, HTML or XML files, RDF/OWL databases, relational databases) and how such a profile can be used for matching the entity against the profile available in an ENS server.

[7] see e.g. Kripke [8]

A second issue has to do with bootstrapping the service. This problem has two dimensions. First, we need to make sure that the ENS is pre-populated with a significant number of entities, so that there is a reasonable chance that people find a URI to reuse in their application; this will be done by implementing tools for importing entities (and their profiles) from existing sources (e.g. from DBpedia.org, DBLP, factbooks, GeoNames, etc.). Second, and even more important, we need to make sure that the interaction with the service is integrated in the largest possible number of common application for creating content. In Section 4 we described two simple examples of how we imagine this interaction should happen; however, it is our plan to extend the idea also to non Semantic Web tools, like office applications, web-based authoring environments (including forums, blogs, multimedia tagging portals, and so on). This approach should make the interaction with the ENS very easy (ideally, the use need not be involved, though he or she of course must be aware of it, see below for this issue) and will slowly introduce the good practice of OKKAMizing any new content which is created on the Web.

A third big issue has to do with the scalability of the proposed solution. Indeed, the number of entities which people might want to refer to on the Web is huge, and the number of requests that the ENS might receive from OKKAM enabled applications can be extremely high. For this reason, the architecture which we envisage for the ENS is totally distributed and decentralized, and is very similar to the DNS architecture. However, there is a problem here: while the DNS is organized hierarchically (mainly geographically), we don't see any obvious way to subdivide the domain of entity names in a hierarchical way. By design, we decided not to impose any conceptual schema on entities, as this would mix the naming service with a viewpoint on what is true about an entity. But we can't assume that the ENS architecture is completely flat, as for example it would be extremely hard to check whether an entity which someone is trying to create through a ENS server already exists somewhere else. This is perhaps the most important open issue in the proposed approach.

Last but not least, we want to mention two non-technical related issues. The first has to do with acceptance: how will we convince people to adopt the ENS? We already had some feedback on the concern that the ENS might become a way to track people or organizations, and therefore we need to make sure that the benefits outnumber the concerns by proving the advantages of the service in a few very visible and popular domains. The second issue has indeed to do with the general problem of guaranteeing privacy and security of the ENS. As to this respect, it is important that we do not raise the impression that the ENS is about storing lots of information about anything. The profiles which we will store will be minimal, and will have the only function of supporting reasonably robust matching techniques. Also, we need to make sure that people have some degree of control on what can be stored in a profile, what cannot, and on what can be stored for improving matching but should never be returned as the result of a query to the ENS.

We are aware that the challenges are quite ambitious, but in our opinion the ENS has the potential to become the factor which will enable the creation of the global, open and decentralized knowledge space which Tim Berners-Lee envisions in the quote we reported at the beginning of the paper.

7 Acknowledgments

This work is partially supported by the by the FP7 EU Large-scale Integrating Project **OKKAM – Enabling a Web of Entities** (contract no. ICT-215032). For more details, visit the project web site at `http://fp7.okkam.org`.

References

1. T. Berners-Lee, R. Fielding, and L. Masinter. *RFC 3986: Uniform Resource Identifier (URI): Generic Syntax*. IETF (Internet Engineering Task Force), 2005. `http://www.gbiv.com/protocols/uri/rfc/rfc3986.html`.
2. S. Bortoli, H. Stoermer, and P. Bouquet. Foaf-O-Matic - Solving the Identity Problem in the FOAF Network. In *Proceedings of the Fourth Italian Semantic Web Workshop (SWAP2007), Bari, Italy, Dec.18-20, 2007,* 2007.
3. P. Bouquet, H. Stoermer, and D. Giacomuzzi. OKKAM: Enabling a Web of Entities. In *i3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.*, CEUR Workshop Proceedings, ISSN 1613-0073, May 2007. online `http://CEUR-WS.org/Vol-249/submission 150.pdf`.
4. P. Bouquet, H. Stoermer, M. Mancioppi, and D. Giacomuzzi. OkkaM: Towards a Solution to the “Identity Crisis” on the Semantic Web. In *Proceedings of SWAP 2006, the 3rd Italian Semantic Web Workshop, Pisa, Italy, December 18-20, 2006. CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-201/33.pdf*, December 2006.
5. P. Bouquet, H. Stoermer, and L. Xin. Okkam4P - A Protégé Plugin for Supporting the Re-use of Globally Unique Identifiers for Individuals in OWL/RDF Knowledge Bases. In *Proceedings of the Fourth Italian Semantic Web Workshop (SWAP2007), Bari, Italy, Dec.18-20, 2007,* 2007.
6. A. Hogan, A. Harth, and S. Decker. Performing object consolidation on the semantic web data graph. In *i3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.*, 2007.
7. A. Jaffri, H. Glaser, and I. Millard. Uri identity management for semantic web data integration and linkage. In *3rd International Workshop On Scalable Semantic Web Knowledge Base Systems*. Springer, 2007.
8. S. Kripke. *Naming and Necessity*. Basil Blackwell, Boston, 1980.