From Linked Data to Linked Entities: a Migration Path

Giovanni Bartolomeo
University of Rome Tor Vergata
Via del Politecnico, 1
00133 Rome, Italy
giovanni.bartolomeo@uniroma2.it

Stefano Salsano
University of Rome Tor Vergata
Via del Politecnico, 1
00133 Rome, Italy
stefano.salsano@uniroma2.it

ABSTRACT

Entities have been deserved special attention in the latest years, however their identification is still troublesome. Existing approaches exploit ad hoc services or centralized architectures. In this paper we present a novel approach to recognize naturally emerging entity identifiers built on top of Linked Data concepts and protocols.

Categories and Subject Descriptors

H [**Information Systems**]: Models and Principles; H.1 [**Models and Principles**]: Miscellaneous; H.1.m [**Miscellaneous**]

Keywords

Entity, Identity, Linked Data

1. INTRODUCTION

The "Identity and Reference on the Web" (IRW) ontology [1] classifies non information resources into `AbstractResources`, `ConceptualResources` and `PhysicalEntityResources`. In order to explain what an abstract resource (e.g. the weight force) or a conceptual resource (e.g., a RJ45 plug or a Margarita cocktail) is, it is sufficient – and necessary – to provide a formal definition, or a technical specification or maybe even a receipt. On the contrary, physical entity resources (henceforth entities) typically have more complex and vague "explanations". Despite entities have been deserved special attention in the latest years, capturing their identity on the Web is still the subject of an open and lively debate[1].

In this paper we will address the main conceptual aspects behind entity identification, as emerging from current state of the art solutions. Based on a better understanding of these aspects, we will rethink at the principles governing the association of descriptions to non information resources. In particular, we will explain why some of these principles might represent an obstacle to pass from the current Linked Data to a world-wide global knowledge space that we call "Linked Entities". Finally we will propose a possible "migration path" based on a methodology which allows to recognize emerging entity identifiers by inspecting the natural evolution of equivalence links.

2. RELATED WORKS

The current Linked Data practice of using the `owl:sameAs` predicate to interconnect two "similar" resources supposed to represent the same entity is often a oversimplification which produces inconsistency between the statements asserted in the associated descriptions of the connected resources [2]. In fact, entity attributes may change over the time[2], and more generally over different contexts. Since RDF does not allow to model *n-ary* relations (in particular, those including time variables), corresponding individual property values (such as age, work position, etc.) may vary in different associated descriptions. Merging different subjects into a single node usually destroys contextualization and creates inconsistency. To mitigate the context-loss effect (a well known issue common to other predicates such as `owl:imports`) the W3C Technical Architecture Group (TAG) has recommended to treat RDF statements as claims by different information providers rather than as actual facts.

Many efforts have been devoted in finding alternatives to `owl:sameAs`. For instance, Hayes and Halpin [3] present four "alternative readings" of the way this predicate is currently used: misplaced references, referential opacity, identity in different contexts and similarity. They also present possible alternative predicates (mainly from the SKOS vocabulary[3]) for each of these cases, but admit that in some cases choosing suitable alternatives to `owl:sameAs` might be difficult: "*their use may be a matter of opinion, as someone's close match may be another person's identical match*".

Jaffri [4] addresses the problem of coreferences in Linked Data. Coreferences may arise i) when multiple URIRefs point at the same resource and ii) when a single URIRef points at more than one resource. The author suggests a solution based on the introduction of a local "Consistent Reference Service" that groups together URIRefs referring to the same resource from different contexts. Jaffri also highlights that some URIRefs may change their "meaning" depending on the context in which they appear.

Bouquet et al. [5] observe that entity identification is difficult because the "*good practice of associating the same URIRef to the same entity* [and using it consistently] *is not supported by any large-scale Web infrastructure*". Therefore, they propose a community-supported entity profile repository called Entity Name System (ENS). The ENS contains profiles of entities that have been assigned invariant and consolidated URIRefs. To issue new

[1] See for instance P. Hayes. Message to www-rdf-comments@w3.org,2003.http://lists.w3.org/Archives/Public/www-tag/2003Jul/0198.html.

[2] See for instance the "Dilibert's cubicles" example presented by Dan Brinkley in his blog post on November, 3 2011: http://danbri.org/words/2011/11/03/753.

[3] See http://www.w3.org/2004/02/skos/vocabs.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

RDF statements about an entity, an information provider should contact the ENS in order to get the "universally unique identifier" associated to the entity, and then use this identifier in her RDF statements. The authors present an implementation in the context of the OKKAM project[4] and discuss the main challenges of this approach: i) finding the right granularity to classify entities; ii) providing suitable invariant attributes in each entity profile to allow to univocally identify the entity; iii) managing the centralized name system (in terms of ownership, privacy, scalability and maintenance).

3. PROPOSED APPROACH

In Linked Data any resource is "identified" by a HTTP URIRef. The choice of the HTTP protocol has its rationale into two main advantages: i) the facility to create identifiers in a totally decentralized fashion, which allows anyone to issue new URIRefs and avoids the disadvantage of maintaining a centralized naming authority; ii) the HTTP own ability of making information accessible by dereferencing the URIRefs. The second facility has been exploited in the successful introduction of a technique [6] allowing to redirect an URIRef assigned to a non-information resource to an URIRef accessing an "associated description" of the non-information resource. Unfortunately, this technique makes harder entity identification at a global scale as it leaves the description of a resource to single URIRef owners. This leads to the open problem of objectively expressing the degree of matching between similar resources.

As opposite to these URIRefs, which he calls `RDFURIs`, Bouquet introduces `OkkamIDs`, identifiers that *directly refer to* entities [7]. Appealling to Kripke[5], the notion of *direct reference* is realized by means of an "entity profile" (`OkkamProfile`) which contains information agreed by the Web community (and not simply provided by a single owner). Delivering information about the "normative" use of an entity identifier as agreed by the community, the entity profile answers one of the main arguments raised by Hayes and Halpin [8], i.e. that the user tends "*to observe* [only] *a small portion of* [an URIRef] *use"* and thus to maintain an implicit ambiguity about the referent of an associated description.

Having understood the fundamental disambiguation function performed by `OkkamIDs`, in the following we illustrate a possible methodology allowing to introduce the concepts of entity identifier and entity profile on top of Linked Data, without "breaking" the deployed base and without introducing external systems or specialized identifiers.

4. METHODOLOGY

We reuse classes and properties defined in the IRW ontology[6] and introduce (Fig. 1) the concept of entity identifier (`EntityIDs`) and entity profile (`EntityProfile`), performing identical functions as, respectively, `OkkamID` and `OkkamProfile`. `EntityID` is a subclass of `SemanticWebURI`. Unlike `OkkamIDs`, `EntityIDs` are decentralized and do not introduce any syntax restriction but the ones defined for their parent class. `EntityProfile` is a subclass of `ldow:AssociatedDescription`. `EntityProfiles` contain only information agreed by the Web community, thus they fix the referent of an `EntityID` by community agreement.

To answer the question of which `irw:SemanticWebURIs` should become `EntityIDs` (and, consequently, which associated descriptions should become `EntityProfiles`), once more we refer to Kripke's "chain of communication", a natural process which occurs to entity names when they are transmitted from people to people through the time and the space. Hayes refers that this process is not causal, rather it has the character of a communication process and may also fail if the provided information about an entity is not accurate enough or is not accurately reported; some names may be lost, others might even change their referent[7]. Nevertheless, we believe that this is part of a natural evolution of the language itself and that when the new referent is finally agreed by the community, its name arises to the role of an (asymptotically) stable identifier, eventually becoming part of the shared human knowledge about the reality. Therefore, rather than creating universally unique entity identifiers, we think that there should exist a natural tendency of some URIRefs to emerge and to become more popular and stable than others. The Linked Data practice to connect resources which are claimed to be similar is, in our opinion, a Web based realization of Kripke's "chain of communication". Following these connections it should be possible to find URIRefs that are natural candidates to become entity identifiers. The search for community agreed entity identifiers then turns into the investigation of the most connected nodes in a RDF graph where the nodes are the target URIRefs and the arcs are RDF links expressing equivalence or similarity. From the analysis of the connectivity properties of this graph we expect to find the small-world and scale-free structure that characterizes natural networks[8]. We plan to discover potential entity identifiers by splitting this RDF graph into clusters and by computing relevant properties of the nodes [10] in each cluster. After an entity identifier has been detected, its associated description could be easily turned into an entity profile conveying community agreed information useful to characterize the entity at a global scale.

5. CONCLUSIONS

"*The identity resolution problem in Linked Data will be naturally solved by a distributed and evolutionary strategy*" [11]. In order to effectively realize this vision, some underlying mechanisms need to be slightly modified. Our approach introduces the concept of entity identifier and community agreed entity profile on top of Linked Data, without "breaking" the currently deployed base. Entity identifiers are taken among the authorities of "similarity networks" which we conjecture to be small-world and scale-free.

[4] The OKKAM project co-funded by the European Commission (GA 215032), ran from January 2008 to June 2010, http://www.okkam.org/.

[5] See Kripke, S.: Naming and necessity. Cambridge, MassaHarvard University Press, 1980.

[6] In the IRW ontology, "`RDFURIs`" are modelled as `irw:SemanticWebURIs`.

[7] See Evans, G.: The Causal Theory of Names. in Martinich, A. P. ed.: The Philosophy of Language. Oxford University Press, 1985.

[8] Ding [9] has recently proved that networks consisting in `owl:sameAs` statements are scale free and that they contain "hubs" and "authorities" from organizations such as DBpedia, OpenCyc, GeoNames and Semanticweb.org.

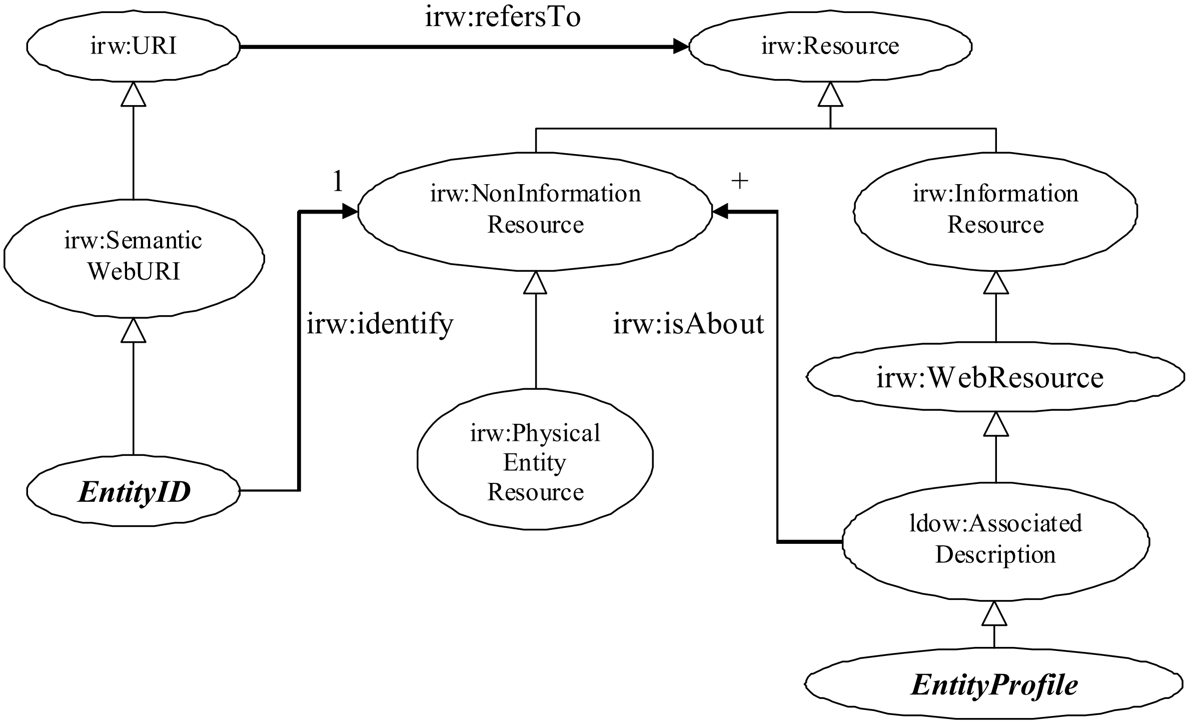

Figure 1. `EntityID` **and** `EntityProfile` **classes introduced in the IRW ontology.**

This choice is fault tolerant: through the years, some authorities might naturally disappear, whereas new ones might arise. Even if entire organizations such as DBpedia or OKKAM could be dismissed, new ones could take their place and their URIRefs could arise to the role of new entity identifiers.

6. ACKNOWLEDGMENTS

This work is partially sponsored by the European Commission under research grant 'Convergence' (FP7-257123).

7. REFERENCES

[1] Halpin, H., Presutti, V. 2009. An ontology of resources: Solving the identity crisis. In: ESWC 2009, Research Studies Press/Wiley.

[2] J. McCusker, J., McGuinness, D. 2010. owl:sameas considered harmful to provenance. In Proceedings of the ISCB Conference on Semantics in Healthcare and Life Sciences.

[3] Halpin, H., Hayes, P.P., McCusker, J., Mcguinness, D., Thompson, H. 2010. When owl:sameas isn't the same: An analysis of identity in linked data. In Proceedings of the 9th International SemanticWeb Conference.

[4] Jaffri, A., Glaser, H., Millard, I. 2008. URI disambiguation in the context of linked data. In Proceedings of the 1st International Workshop on Linked Data on the Web.

[5] Bouquet, P., Stoermer, H., Niederee, G., Mana, A. 2008. 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 554-561 IEEE Computer Society.

[6] Sauermann, L., Cyganiak, R., Volkel, M. 2007. Cool URIs for the Semantic Web, Technical Report, TM-07-01, DFKI.

[7] Bouquet, P., Palpanas, T., Stoermer, H., Vignolo, M. 2009. A Conceptual Model for a Web-scale Entity Name System, In Proceedings of 9th the Asian Semantic Web Conference.

[8] Hayes, P., Halpin, H. 2008. In defense of ambiguity. International Journal of Semantic Web and Information Systems, 4(3).

[9] Ding, L., Shinavier, J., Shangguan Z., McGuinness, D.: SameAs Networks and Beyond. 2010. Analyzing Deployment Status and Implications of owl:sameAs in Linked Data. Lecture Notes in Computer Science, Volume 6496/2010, 145-160.

[10] Watts, D.J, Strogatz, S. 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684): 440-442.

[11] Bizer, C., Health, T. 2011. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory & Technology, 1:1,1-136. Morgan Claypool.