Abstract

With the advent of RDFa and the at least partial support by major search engines, semantically structured data is more and more appearing on the Web. To enable high value use cases, links between entity descriptions need to be established. The linked data model suggests that links should be state explicitly by those who expose entity descriptions, but unlike on the normal web, incentives for doing so are unclear so that the model ultimately seems to fail in practice. In this position paper, we make the conjecture that explicit links are not needed for realizing the semantic web. We propose discuss how Record Linkage techniques are in general very well suited for the task but argue the need for a tool would allow data publishers to have an active role in producing entity descriptions that can then be linked automatically.

The need for linkage, the absence of links

The dream that inspired many Semantic Web researchers is that of a web where bits of information are discovered and connected automatically because they “matter” for the task at hand, possibly coming from any web location and ultimately reused well beyond the purpose for which they were originally created. Applied to news, this vision would allow a reader to get “second and third” points of views when reading about anything. Applied to commerce it would ideally eliminate the need for advertising: sellers and suppliers would simply “be found” for the characteristics of the offer.

Given no expected imminent breakthrough in the ways machine can understand content meant for human consumption, the idea of the Semantic Web initiative has been that of proposing that Web Site “lend a hand” to machines by encoding semantics using RDF.

For years, however, RDF descriptions on the Web have been made available almost exclusively by web data enthusiasts, i.e. by the Semantic Web community itself. Despite this, the community has been able to made available a remarkable amounts of information, known as the Linked Open Data cloud, to the point that many entities, e.g. encyclopedic entities but also the people participating in the community, are often “described” (have metadata about then in RDF) in several dozen different independent RDF sources on the web.

The existence of descriptions alone, however, it is not sufficient condition for this data to be discovered automatically. For this reason the LOD community has been advocating the reuse of URIs of other sites as a way to create interlinks. In [1], it is explained that to allow crawlers and agents to understand that a description is about something described also elsewhere, URIs from other sites should be used. For these URIs to be found, one should first manually select datasets from a maintained list of known datasets, then explore these to find suitable URIs to link to, this for each entity to be linked. It is suggested that automatic methods be used when linking multiple entities, e.g. [2] but especially in this case it is necessary to know a priori which specific dataset to link to and to perform manual configuration of the matching algorithms, something that requires a high degree of expertise.

This complexity, together with the – arguably temporary - lack of immediate incentives for doing this, makes it so that even among the LOD community formal data quality [5] and interlinks are scarce. A quick query on Sindice, currently indexing approximately 65 M semantic documents shows that less than 4 million RDF documents (usually entity descriptions from the LOD cloud) exhibit at least one sameAs link.

In the last year however LOD is becoming no more the only source of large amounts of RDF structured content. Thanks to the support of Google and Yahoo for RDFa encoded content for advanced snippets, it is safe to say that tens of millions of pages of database generated content have appeared, none of these, to the best of our knowledge, providing interlinks among descriptions on different websites.

Missing Links are here to stay

We believe that the problem of missing interlinks in RDFa descriptions - and of the little number of interlinks also in the Semantic Web enthusiasts community datasets is “here to stay” for many reasons. Links to the web increase the value of a site by making it “more useful” to visitors, the
same cannot be said of invisible RDF links. Also to be of use one would expect to have many links from a single entry point, something which is requires manual or semi assisted 1 to 1 connections. Finally, for being consistent on the Semantic Web, links would have to be bidirectional and maintained by multiple parties at the same time, something which is in essence contrary to the principles of decoupling which made the very web successful. Furthermore there could be political and commercial reasons why a dataset provider might not have any incentive to put links, e.g. when in a dominant market position.

For the Semantic Web Interlinking, AI could actually work

Creating meaningful, entity centric, aggregates of textual Web information is a dauntingly difficult NLP task. On the one hand one has to do named entity extraction, and extract attributes and facts about the entity on the other one would have to be sure that both the attributes and the entity itself are the same or mean the same thing as those mentioned in other websites. 

Doing the same thing on semi structured RDF, on the other hand, can reduce to the well known problem of duplicate record detection or Record Linkage (RL). RL between distributed databases is considered a difficult problem still [3] but the RDF case is in a sense simplified. While data can still be subjects to problem typical of RL scenarios - transcription errors, incomplete information etc - the use of RDF and the shared used of ontologies, something fostered by common practices and by recommendations by big search engines, makes RL on Ontologically structured content an easier task than distributed linkage on heterogeneous databases.

- There should at least in theory be no ambiguity in the mapping between different “fields” in each dataset.
- Ontologies should be able to strongly help the RL algorithms by suggesting that certain properties e.g. Inverse Functional Properties be taken in very strong consideration (if not being taken for certain) as identifiers for the same entity.

Under these conditions, one could envision interlinks between semantic descriptions simply created by a centralized entity, e.g. a Search Engine, and consumed by clients and or data publishers themselves (which would embed it in their data representations).

While this has been done on Semantic Web data services such as Sameas.org [7], and Okkam [8], these have so far been operating mostly based on a limited number of preselected datasets. Can open web data consolidation work reasonably well, and if so under which conditions?

To gain insight about this, we developed http://sig.ma - a Web of Data search and browser based on a combination of Semantic Web principles based data consolidation (e.g. looking for the same URI reused in different documents, looking for SameAS) and simple soft computing methods for selecting potentially useful data sources and for performing Record Linkage and data deduplication [6].

As a result of this, it is possibly to empirically verify how doing a Sig.ma search on entities for which there is data on the semantic web (e.g. names of community researchers – or projects) provides a visibly much greater wealth of information than when applying Semantic Web principles only (e.g. searches by URI, IFPs, following direct links etc). This can be verified by doing a query on Sig.ma and looking at the sources that sig.ma aggregates, noticing that, usually, very few if any of these are linked by actual semantic web links as previously defined.

The curse of scaling it

As any reader might notice by trying the system, Sig.ma’s simplistic entity consolidation will quickly fail once queries are ambiguous e.g. refer to multiple persons by the same name e.g. a “John Smith” query or even different kind of entities e.g. Galway (notably a city but also a last name).

For human users, Sig.ma provides tools to overcome this. Users can select information that seem unrelated to his search and this will cause the related data sources to be removed accordingly. As a result, a clean profile can be generated usually in a few clicks and used as needed, e.g. embedded in a blog post.

Machines and software clients using sig.ma, however, will inevitably suffer from the noise. While it is safe to say that Sig.ma’s algorithms have very big room for improvements, the task suffers from a fundamental “curse”; as the Internet data space becomes increasingly more populated, more and more entity descriptions will having overlapping and ambiguous attributes making so that results of automatic linkage will not hold in time.

A conjecture for a “Web Linkage” validator

The conjecture is that RL systems alone are never likely to be able to scalably interlink the semantic web, if not somehow sustainably “helped” by data producers themselves.

We therefore imagine the creation of a tool to help data producers assess how well can Record Linkage be performed between their data and the rest of the Semantic Web Data – something we will here call Web Linkage. The workflow would be as follows:

A) The dataset owner submits a link to the dataset. The dataset is fetched the following steps are performed:

- Calculating how “close” dataset entities are to others previously known by the system to be on the Semantic Web. E.g. The system would warn when a person Giovanni Tummarello is inserted being this close both to Giovanni Tummarello and Giovanna Tummarello.
- Evaluating how ambiguous are the properties and values that are being used. E.g. a perfect match with a record whose name is “John Smith” would raise much more attention than a record match for the much rarer full name Giovanni Tummarello.

B) A report is given to the dataset owner, possibly highlighting suggestions in case the Web Linkage of the data resulted poor. For example the owner could be suggested to:
• add more attributes, or more specific attribute values
• better specify classes and properties
• in extreme cases, to manually put sameAs links where the records seem extraordinarily difficult to disambiguate.

C) As unambiguous records can become ambiguous as more data joins the Semantic Web, the validator will likely offer services such as continuous dataset review, in this case the datasets would be periodically reviewed and warnings sent to the dataset owner as needed.

Discussion and conclusions
In this short position paper we present a Semantic Web data publishing model where:

• data is produced with no regard for interlinks per se. Owners who care about being correctly interlinked use a “Web Linkage” validator.
• interlinks are instead generated by centralized services using state of the art Record Linkage specifically optimized to use machine learning together with ontology driven reasoning rules e.g. as in [4]. Interlinks are generated if the records meet a certain Web Linkage confidence threshold. If the dataset passed the Web Linkage validator stage it is expected that each of its entities should be above the Web Linkage confidence threshold.
• links are consumed by clients, e.g. browser extensions, robots, aggregators (e.g. sig.ma) looking them up in the search engine.

We then make the conjecture that such model can indeed be sustainable and scalable, possibly enabling many of the wanted Semantic Web scenarios. The bases for this conjecture are the following arguments:

• The technical feasibility of high quality automatic record linkage service and a matching validator service which could give very effective advices (e.g. “add the homepage, or a phone number or a date of birth” to be almost certain to correctly linking any person)
• The simplicity for data publishers: a single tool which does not require the data publisher to know or care about any other dataset (e.g. as it would be required by direct dataset to dataset interlinking tools). Furthermore a dataset which has good Web Linkage will continue to automatically generate good links (in the automatic record linkage system) in the future with no intervention by the data publisher whatsoever.
• The value and incentives for data publishers: the method allows universal interconnection, not directly links to previously known datasets, the perceived value in doing the operation will be much higher.

References
3. Ipeirotis P. , Verykios V, Elmagarmid A Duplicate Record Detection, a Survey IEEE Transactions on Knowledge and Data Engineering Volume 19, issue 1 – 2007
4. Sais F. , Pernelle N, Rousset MC Combining a Logical and a Numerical Method for Data Reconciliation Journal on Data Semantics, 2009