A business value driven approach to a new L(O?)D definition

Dr. Giovanni Tummarello (giovanni@siren.io)
Siren.io

Abstract

In this short position statement, we start from the discouraging state of the "LOD" efforts and discuss a more humble "catalyst" research approach. Moving slightly away from the concept of "Web", the research would start from studying successful "D Communities" - communities around data - to identify ideal candidate for first pragmatic data interoperability "bridges" forming a new LD cloud. Just later, if possible and not as primary concern the research would look at the O part. It is envisioned that technically, the result might be completely different from the mechanisms envisioned in the current LOD cloud (e.g. might not use RDF, OWL and will include all sort of exchanges including emails and manual web form submission) but the less elegant approach impact on society and business might be enormous and ultimately a revolutionary success for Semantic Web research.

Keywords: Linked Data, Data Interoperability, Open Data, Semantic Web

Introduction

In this paper we use the term "failure" or "success" for an initiative as they would be defined in a reasonably economic self sustained environment: success means adoption, self sustainability, and measurable business value - failure the opposite.

By this definition it can be argued that the LOD effort - based on bottom up dictation of technical principles like dereferenceable URLs for entities, RDF, OWL, "follow your nose" and similar - has failed. Outside of academically funded projects or sporadic "labs" of companies so vaste to be able to fund relatively reality detached individuals, the LOD methodologies have not proven to generate interest. A measure of this is that there is no evidence that people that were otherwise employed on other technologies (E.g. regular data integration, DBs, BI, Analysts etc) are spontaneously using the LOD stack in any significant measure [1].

In this paper we argue that there is still enormous value to unlock in facilitating industry in sharing valuable data but research seeking to unlock this value should learn from the past and follow a pragmatic, flexible, business value driven approach.

The four steps to a new L(O?)D methodology (and cloud picture).

Step 1: The "D communities" Cloud - and not only on the web

In the real world, there are many data sharing communities at large, i.e. any community that shares structured data even if not openly.

These are numerous and fundamentally important (multi trillion dollar industries) as well as growing pushed by regulations and the sheer desire to optimize business processes. In healthcare and life science, for examples, standards like SDTM (Study Data Tabulation Model) [2] are being mandated by regulatory organizations like FDA for product submissions. Similar examples of standardizations and "data communities" exist and have existed for decades, at various stages of deployments, for hundreds of other specific use cases and needs.

We could see a new research starting from a simple survey of these communities to create a novel LOD Cloud with only the "D part": *listing communities where there is strong, self sustained, active, business valuable structured data exchanges*. One could even imagine drawing this initial "D Communities " map to look similar to the current LOD map: with bubble sized with the number of participants or the estimated value of that data sharing community and clustered by topic.

A big difference between the previous LOD cloud and the new one would be sort of "ontologically" centered (centered about the use of interoperable dictionary mechanism) vs centered around the use of common identifiers.

Also, we notice that the new D Community cloud might, but more likely might not, be connected necessarily to the Web mechanics of URLs URIs even HTTP. In the new D Community diagram, a community of companies that have to submit an individual regulatory report each month in a predefined machine readable format would form a perfectly valid bubble.

We argue that while this detachment from the Web mechanism might depart painfully from the hope of universal automation, the result would embraces the real world via much wider inclusion and much wider possible impact. We argue furthermore that we should still be calling it "semantic web" research as it would be research on the impact of "knowledge representation technologies" (semantics) in distributed communities ("on the web"). So it would be a "semantic web" exchange even if made by actors who send structured reports to a regulator via email or online submission facilitated by the research in question.

Step 2: Finding the first Bridge

If we are to believe the intuition that "Linked Data" is more valuable than disconnected data, then we argue that the easiest path to prove this would be for research to start from the above "D

Communities" cloud. The objective of this stage of the research would be to find initial target split communities in which we believe a "linkage" initiative would *succeed*.

One would initially ask: are there then 2 communities - from the D list of large valuable - that would benefit from cross pollination and interoperability?

We argue that this step should be done with rigor and skepticism. This selection should be based strongly on interviews to industry and community participants at each level - from business to the compliance and IT elements of the community. Questions might include:

- What value could there be if you also had data from the community X
- If it's so valuable why do you think it hasn't been done before (technical? Legal? Low value vs effort? Business model? Do people get easily around that problem? Very infrequent use case?)

Step 3: Building the bridge, toward the first LD communities

We argue that it is only after a deep study of the D communities, on the "use case" gap between them and perceived "value" in the linking that one could (and should!) attempt at build a bridge

A natural research objective at this point would be: what would be the simplest way to get that interoperability going sufficiently to unlock that value?

We argue that at this point RDF should seriously be questioned as a proposed method for standardization and bridging. The research should aim at the simplest way, the one that has the lowest path of resistence across the 2 D communities. E.g. the answer might be simply "standardizing CSV headers and table names" (E.g. as in SDTM [1]) or using excel sheets, or standardizing APIs formats and responses. Measuring *success* or *failure* according the the previous definition would be then critical at this point. If success is attained, we would at this point have the first LD Cloud, *even if the technical means for the individual links are not per se compatible among each other*.

Step 4: wider interconnection and standardizations: a novel pragmatic L(O?)D Cloud

Assuming positive results on 3, it is at this point one could realistically believe that a L(O?)D cloud could be attained. The efforts would then be to try to find a "repeatable recipe" that can be shared so that - at least for groups of D communities - the interoperability is achieved outside 1 to 1. What could be the technical mechanism to make the 1 to 1 scale to N to N is to be devised

but we feel that sharing best practices, simplicity and allowing for still a degree of manual work and use of commonly used data format (xml/json/csv) might deliver many of the benefits at a fraction of the conceptual cost. We believe that even with the above degree of flexibility, the resulting cloud picture might never end up being fully connected. It would be a "patchy", yet "at the end of the day functional for use cases that matter" LOD cloud where a set of communities have considerably lowered the cost of interoperability, unlocked value in cross community use cases via sharing working examples, success stories and simple recipes.

About the O part of the LOD cloud, we believe Open is just a hinder to the real objective of the Semantic Web and Knowledge Representation research, which should be interoperability. Openness is a "nice" attribute in the very most of cases (but obviously not all see GDPR) but should not be a defining attribute and should be probably taken out of prominence in the new semantic web research.

Conclusion: searching for a catalysts, hoping for a revolution

The Linked Data/ Semantic web initiative hoped for an interoperability revolution of the same (or bigger) magnitude than how the Web revolutionized human legible information sharing.
This has not happened. Arguably, this was caused by a combination of researcher bias toward openness, bias caused by the previous success of "standardization" (e.g. HTTP/HTML) and unwillingness to consider business value driven dynamics and evolutions in big data technologies and pragmatic ways people produce and exchange data.

We believe that the new LD/Semantic Web research should not try to find the "revolution" in a new technical format or (web) standard alone. It should instead be technically flexible and focused on acting as minimally invasive "catalyst" to achieve faster the value that eventually would have anyway emerged by industry initiative.

But this effort could as well be revolutionary in the end: by sharing recipes that work, having business D communities as testimonials, the new L(O?)D effort and map would likely be very inspiring let alone valuable, and this time have a truly effective network effect

References

[1] : A More Decentralized Vision for Linked Data, Axel Polleres, Maulik R. Kamdar, Javier D. Fernandez, Tania Tudorache, Mark A. Musen (submitted DeSemWeb 2018, via OpenReview)
[2] : SDTM https://www.cdisc.org/standards/foundational/sdtm