From: **Eric Prud'hommeaux** eric@w3.org
Subject: challenges of distributed data
Date: 11 January 2019 at 06:21
To: group-data-ws-pc@w3.org
Cc: Andra Waagmeester andra@micel.io

[[
People think that is a pain because it is complicated.
The truth is even worse.
RDF is painfully simplistic but it allows you to work with real-world data and problems that are horribly complicated.
]]
-- Dan Brickley and Libby Miller

It seems one goal of this workshop is to enable the use of various graph databases for public data stores. Here we discuss some lessons learned from years of Semantic Web projects.

The goal of RDF/SemanticWeb/Linked Data, etc is allow folks to use, combine and extend each others' data with a minimum of coordination. This is largely done through costly coding disciplines which deliver unambiguity and minimal diffusion. As graph data is used for more decentralized applications, similar costs are incurred. An example is wikidata, which has unique URLs ensuring unambiguity, a social process for vetting new terms and entities intended to minimize diffusion, and a uniform model through which all assertions are made. Resource-oriented architectures like Wikidata and FHIR (RESTful clinical records) are pretty easy to expose in RDF because they are already addressing the hardest problems.

Demo Half-life

One of the biggest challenges for the Linked Open Data cloud has been balancing innovation against maintaining usefully query-able data. While sites like Uniprot, with a specific mandate for data stewardship, consistently offer widely-useful and trusted data, much of the LOD cloud has no paid stewards and no rigorous data governance. Pharmaceuticals who wanted to address this created Open PHACTS specifically to add much-needed stewardship to make sure that today's code works tomorrow, or even ten minutes from now. Others turn to professional curators like SciBite to ensure a stable knowledge platform. The problem of data quality and stability is a social one; someone has to be paid to care about the data and delegated authority to take appropriate action. The if-you-build-it-they-will-come tack rarely pays off without some serious attention to reliability from the user perspective.

Social and Machine Contracts

While wikidata as a whole mandates no continuity governance, specific communities can establish their own standards of data practice. An example of this is the GeneWiki project which uses Wikidata as a hub for life science linked data and is performing consistency checks and exploring version and migration as enforced by schema mappings. This contract is described in [1], which clarifies the stability of the schemas and their underlying data structures. It is our suggestion that any publicly-available data repository follow such a practice.

[1] https://github.com/SuLab/Genewiki-ShEx#stability-of-the-shape-expressions

Eric Prud'hommeaux (Micelio, Janeiro)
Andra Waagmeester (Micelio)
--
-eric

office: +1.617.258.5741 32-G528, MIT, Cambridge, MA 02144 USA
mobile: +1.617.599.3509

(eric@w3.org)
Feel free to forward this message to any list for any purpose other than email address distribution.