Opening Walled Gardens:
RDF / Linked Data as the
Universal Exchange Language of Healthcare

David Booth PhD[1], Rafael Richards MD, MS[2], Michel Dumontier PhD[3], Conor Dowling[4]

A Response to the Office of the National Coordinator for Health Information Technology, HIT Policy Committee's Request for Comment Regarding the Stage 3 Definition of Meaningful Use of Electronic Health Records (EHRs)[5]

These comments apply specifically to IEWG-101, IEWG-103 and SGRP-204A, but also address the overall approach taken for the exchange of healthcare information, following the spirit of the PCAST Report[6] and its recommendation for a universal exchange language for healthcare.

Background

To enable the most effective use of exchanged healthcare information -- both for direct patient care and for secondary use such as quality-of-care measurement and research -- it is essential that the information be represented in a machine processable[7] language. Being machine processable allows the information not only to be correctly understood and processed automatically by the recipient, but it also allows the information to be automatically integrated with information obtained from other sources.

However, although necessary, being machine processable is not sufficient when thousands of independent information producers and consumers are involved. In such contexts lifecycle issues of versioning, incremental adoption and incremental standardization become critical, and the decision of which machine processable language(s) to adopt has a big impact on the overall cost and effectiveness of the result.

Those who are unfamiliar with Semantic Web technology are likely to assume that vocabularies must be standardized *before* the public would benefit from the exchange of health information in a universal, machine processable language. This is not the case, for three key reasons:

---

[1] Senior Software Architect. http://dbooth.org/ Email: david@dbooth.org

[2] Assistant Professor, Division of Health Science Informatics, Johns Hopkins University; W3C Healthcare and Lifesciences Invited Expert; HL7 Working Group, Anesthesia and Critical Care. Email: rafaelrichards@jhu.edu

[3] Associate Professor of Bioinformatics, Carleton University. Email:michel.dumontier@gmail.com

[4] President, Caregraf Inc.; www.caregraf.org Email:conor-dowling@caregraf.com

[5] http://www.healthit.gov/sites/default/files/draft_stage3_rfc_07_nov_12.pdf

[6] http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf

[7] For a definition of "machine processable" see: http://opengovdata.io/2012-02/page/5-1-2/principle-data-format-matters

- If a universal, machine processable language is adopted, and it is self-describing such that term definitions can be easily obtained automatically, then *some* information consumers can and will make effective use of received information even before vocabularies have been standardized. Indeed, healthcare providers who are consuming exchanged information have a built-in market incentive to distinguish themselves from their competitors by doing so. (On the other hand, healthcare providers have little incentive to produce such information, and that is why it is essential that Meaningful Use requirements mandate its availability.)
- The exchange and use of not-yet-standardized information paves the way toward standardization, by providing real-life implementation experience with the various competing standardization choices.
- By mandating the use of a machine processable language, information producers have a built-in incentive to supply information according to the vocabularies that are most likely to be standardized -- to minimize their own change when standards are adopted -- thus accelerating the standardization process.

We believe that existing and draft Meaningful Use requirements have not adequately considered the possibility of mandating a universal exchange language for healthcare information. Specifically, we believe that a schema-flexible language (specifically RDF[8]) can address the needs of healthcare information exchange better than more schema-rigid languages such as XML that are currently under consideration. Adoption of RDF as a universal information exchange language for healthcare would significantly reduce the overall cost and increase the overall effectiveness of electronic healthcare information exchange.

Benefits of RDF

RDF is the foundational information exchange language for the Semantic Web[9]. It is not the only schema-flexible information exchange language in existence, but it is the best currently available for addressing large-scale information integration and vocabulary evolution problems such as the problem we face in healthcare informatics, for several reasons:

- **RDF is syntax independent.** This helps put the focus on semantics instead of syntax. Contrast with XML.
- **RDF plays well with other information representation languages.** Because RDF is syntax independent, existing formats can be readily mapped to RDF.
- **RDF is schema flexible.** Information can be represented in RDF before vocabularies have been standardized. Whenever vocabularies become standardized, they can be semantically linked with existing RDF data, allowing automated conversion between vocabularies. This allows innovation to proceed rapidly -- no committee bottlenecks to slow down innovation or adoption -- while still benefiting from standards when they are defined. (Contrast with XML.) Standardization helps interoperability, but takes time to achieve.

[8] http://www.w3.org/RDF/

[9] http://www.w3.org/standards/semanticweb/

- **RDF is schema agnostic.** Multiple vocabularies and data models can co-exist peacefully within the same dataset, interlinked. (Contrast XML.) This is important when uniform standards have not yet been adopted because it allows agnostic data representation: existing data representations can continue to be supported while new data representations are added to the mix. RDF excels at integrating information from different vocabularies and data models.
- **RDF data is self-describing.** Concepts are identified by URIs that can be used to easily locate term definitions or other related information, thus making data easier to interpret correctly. (This can be partially achieved in XML, but not as easily as in RDF.)
- **RDF is easy to generate.** Existing relational, hierarchical or other data representations can be easily converted to RDF for information exchange and integration. Tools and techniques are readily available.
- **RDF is proven technology.** RDF was designed specifically for Web-scale information integration, as a foundational technology for the Semantic Web. It has been used for numerous large scale information integration problems. Many US government datasets are now publicly available in RDF, as are many non-governmental datasets. RDF adoption has steadily increased since its first standardization in 1999, especially in biology and life sciences. All term sets in the UMLS -- over 100 term sets -- have been converted to RDF.
- **RDF provides semantic interoperability**. Each data element, as described above, is self-describing. Each data element is independently linked to its defining ontology or terminology such as SNOMED, LOINC, or RxNORM, all of which are also in RDF.
- **RDF is data-atomic**. Each data element can be independently queried, changed, or deleted independent of any other data element. Data does not depend on its physical arrangement, and cannot "break" the database by its deletion.
- **RDF is scalable**. RDF is designed for web-scale applications.
- **RDF enables inferencing and knowledge discovery.** Data represented in RDF can be reasoned over to discover new knowledge.
- **RDF is secure.** Existing security and privacy mechanisms can used with RDF just as with XML or any other information representation language.
- **RDF is an open, non-proprietary, international standard.** RDF and related standards such as OWL and SPARQL were developed by the World-Wide Web Consortium (W3C) for unencumbered international use.

What can RDF do better that other proposed exchange languages?

Meaningful Use currently mandates a patchwork of idiosyncratic formats, such as HL7, CCD/ C32, CCR, NCPDP SCRIPT, C-CDA and QRDA. While such formats provide a degree of machine processability, in comparison, RDF offers significant advantages:

- RDF offers a simple, uniform language for all healthcare information. This decreases the overall complexity of processing and integrating healthcare information.
- RDF provides substantially better lifecycle characteristics, which is critical given that thousands of providers will act as information producers and consumers.

- RDF data, following Linked Data principles[10], is more self describing.

Recommendations

Given the above considerations, we recommend the following:

1. ***The HIT Policy Committee should seriously consider adopting RDF as a uniform, universal exchange language for healthcare***, using Linked Data principles, in order to reduce the overall cost and increase the overall effectiveness of electronic healthcare information exchange.

2. ***Meaningful Use should require that all EHR data be made available***, with full fidelity and granularity, in a machine processable language -- both data elements whose vocabularies have been standardized and those that have not (yet). As explained above, this would enable more effective use of electronic healthcare information and accelerate the vocabulary standardization processes.

Does exchanging healthcare data (as RDF) support the HITPC guiding principles?

This section addresses some specific questions posed in the RFC.

Q: Does RDF support new models of care (e.g., team-based, outcomes-oriented, population management)? A: Yes. A major strength of RDF is its flexibility for a wide range of uses.

- ***Address national health priorities (e.g., NQS, Million Hearts)*** A: Yes. Mandating a uniform exchange language will simplify the task of supporting new uses and applications of healthcare information.
- ***Have broad applicability (since MU is a floor) to***
    - ***o provider specialties (e.g., primary care, specialty care)*** A: Yes. A strength of RDF is its ability to absorb new data models and vocabularies.
    - ***o patient health needs*** A: Yes, by facilitating information integration.
    - ***o areas of the country*** A: Yes. RDF is an international standard -- not biased toward any particular area of the country.
- ***Promote advancement -- Not "topped out" or not already driven by market forces*** A: Yes. RDF would substantially advance the cause of semantic exchange and integration of healthcare information.

[10] Summary of Linked Data principles: http://linkeddatabook.com/editions/1.0/#htoc9

- ***Be achievable – e.g. there are mature standards widely adopted or could be widely adopted by 2016*** A: Yes. RDF was first standardized in 1999. Its adoption has increased steadily since then. It has been used by many major information integration projects, and has especially gained traction in biomedical research. Hundreds of US government datasets have been published as RDF, and many other non-government datasets have as well.
- ***Reflect reasonableness/feasibility of products or organizational capacity*** A: Yes. RDF is easy to generate from relational databases, hierarchical and object databases, XML documents, spreadsheets, and other data representations, and tools are available as well.
- ***Prefer to have standards available if not widely adopted*** A: Yes. RDF and related specifications such as SPARQL and OWL are established, international standards defined by the World Wide Web Consortium (W3C).