

Practical lurches towards semantic interoperability, and standards mash-ups in public sector data

James Bryce Clark, OASIS [1]
May 2007

[1] This submission represents personal views only. OASIS' technical program and choices are set by its members.

Abstract

A number of public administration projects successfully have used RDF. More probably should. Many are in need of higher-order knowledge representation capabilities, as well. This note describes three instances, and notes some conditions that may influence how current and future semantic methods may be applied.

Introduction: resolving questions of equivalency

With apologies to Hayek, information was born free ... but everywhere it is in silos. In governmental data exchange, as in the commercial sector, we traffic in common bits of data constantly ... names, prices, weather conditions ... the address where something is located ... the fact that someone is someone else's employer or parent or executor. We do so in huge volume. And yet, embarrassingly, much of that data is not yet universally expressed, or even re-usable.

When two correspondents face each other across the Internet, or time, each holding an apple (or datum), there are several possible communication outcomes. The distinction is relevant, in thinking about governmental uses of semantic methods.

- They may use the same language, or at least natively understand each other's meaning. Huzzah! Perhaps they do not need KR help. If only this were common.
- They may both be, in good faith, holding equivalent apples ... and trust each other reasonably well ... but unable to understand that this is the case. Or they are trying to talk about different aspects of the apple. This is a communication challenge; but it may be simple to resolve. (We will consider the case of emergency warning messages.)
- They may be holding the same thing, but unable to easily trust or process each other's description without a common point of reference. (We will return to this, in the context of automobile repair data.)
- They may be holding different things, but unclear about the equivalency ... inadvertently or deliberately. (This may have meaning in looking at the ongoing global 'Core Components' project.)

The latter three cases are customers for semantic KR methods! The last two acutely may need mediation or reconciliation of meaning, as well as granularity.

Data exchange and competition: Auto repair data in Europe

In 2002-03, a group of automobile manufacturers (OEMs) and auto repair industry representatives in the European marketplace, along with regulators from the EU Enterprise Directorate, convened and participated in an OASIS technical committee. [2]

Their objective was to define data exchange specifications for OEM data about certain vehicle repairs and parts, to make it broadly available to independent repair shops as well as the OEM's own repair facilities. (Among other things, this was thought important to maintain widely effective auto exhaust emission control in Europe, and stimulate competition.)

[2] <http://www.oasis-open.org/committees/autorepair/>

The committee defined and issued a mutually acceptable data structure. [3] However, the committee declined to approve it by final vote, due to a stated disagreement about how to bear the cost of provisioning that data. [4] Eventually, seeing no voluntary resolution of the cost sharing issue, the European Parliament passed legislation mandating its use nevertheless, in a resolution amending its Directive 72/306/EEC. [5]

[3] <http://www.oasis-open.org/committees/download.php/2412/Draft%20Committee%20Specification.pdf>

[4] See Appendix C to the draft specification.

[5] <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P6-TA-2006-0561+0+DOC+XML+V0//EN>

The OASIS Auto Repair Information TC draft specification, now made law, relies principally on W3C's Resource Description Framework (RDF). RDF permits data-exchange parties to rely on the meaning of each other's data and queries without requiring run-time consultation -- a resource that may not always be available among competitors. The specification also defines and consumes several other namespaces, including some industry specific ones (such as vehicle identification number), and some common concepts from other general schemes (such as W3C's SWAP Personal Information Markup [6] for personal address data like 'phone', 'address' and 'city'; OASIS' UBL [7] for 'price', 'currency', etc., and Dublin Core [8] for resource 'creator', 'title', 'subject' and 'date'). Finally, the specification must accommodate the addition of 'local' namespaces and taxonomies for a given subset of consumers, such as the product parameters or parts catalogs for a specific OEM.

[6] <http://www.w3.org/2000/10/swap/pim/travel.html>, or more broadly, see <http://www.schemaweb.info/schema/SchemaInfo.aspx?id=32>

[7] <http://www.oasis-open.org/committees/UBL/>

[8] <http://dublincore.org/>

The user community for this industry-specific data was well defined, known, and actively engaged in the specification's definition. A large number of potential mismatches of meaning actually were identified and discussed during the design phase of that definition exercise. Arguably, those conditions are an excellent case for agreement on a well-designed, detailed level of semantic markup.

Extensions to loosely organized data: the Common Altering Protocol

One of the most challenging aspects of emergency and incident management efforts, to disseminate notices and assistance in catastrophic weather, hazard or security conditions, has been the lack of technical interoperability. Historically, siloed separate communication systems (dedicated emergency networks, broadcast radio, Internet,

cellular networks, etc.) often cannot intercommunicate, or even share a single message among them.

The OASIS Standard "Common Alerting Protocol" (CAP) [9] was developed by OASIS' Emergency Management Technical Committee [10] to enable public warning information over exchange a wide variety of data networks and systems. CAP specifies a common, very light, XML-based data structure for warning messages, generically suitable for multiple transport methods. CAP remains simple, so as to remain fully compatible with existing heterogeneous legacy public warning systems.

[9] http://www.oasis-open.org/committees/download.php/15135/emergency-CAPv1.1-Corrected_DOM.pdf

[10] <http://www.oasis-open.org/committees/emergency/>

Essentially, CAP specifies a document model composed of a few simple categories of metadata that practically any alerting system can parse. Of course, they also can be combined as necessary to support more complex messages and information. The principal defined elements include:

- an <alert> element, containing basic message identified data such as time-stamping, recipients, and containers to pass other implementation-specific instructions;
- <info> elements to contain the core details about the alert event (such as category, urgency, severity, source, event codes and the like);
- <resource> elements to contain pointers and descriptions (or serializations) of relevant data sources such as images or audio files; and
- <area> elements to specify geographic application of the alert data, using specified standard geospatial reference systems.

An extremely broad variety of event-specific messages and warnings easily can be encapsulated in the CAP model.

CAP v1.0 was approved as an OASIS Standard in May of 2004, and implemented by United States National Oceanic and Atmospheric Administration (weather reporting) and the United States Geological Survey (earthquake, volcanic and landslide events). CAP v1.1 added several functions, and after final approval at OASIS, was cross-contributed to ITU-T for a joint workshop in 2006 [11] and global approval as an ITU Recommendation, anticipated in 2007.

[11] <http://www.oasis-open.org/events/ITU-T-OASISWorkshop2006/proceedings.php>,
<http://www.itu.int/ITU-T/worksem/ictspw/index.html>

Considering the kinds of diverse networks that a tsunami, hurricane or toxic spill warning must travel -- and response teams often assembled on the fly -- there will be pressure to keep CAP messages easily parsed by low-bandwidth or low-processing-power recipients. But CAP's DOM being XML, it is readily augmented. Where it is needed, methods like RDF are an obvious opportunity to extend meaning. Subsets of first-responders may have more acute precision needs, in an emergency. It is one thing to say, there is a forest fire at *this* time, at *that* 3-dimensional location. Presumably that's enough, for a ranger to evacuate an area. But as we get into provisioning of firefighting help, the need for meaning may escalate. Where is the fire's leading edge? *Is* there a leading edge? Where is the fire battalion? *Which* one -- the one that worked all night,

last night? Do they have trucks? Do you mean the kind of trucks that carry water? Or flame retardant? How *much* retardant? Etc. Where simple XML messages are used across heterogeneous networks, more detailed content can be packed into simple messages, for those who need it, without impeding the virtue of simplicity.

Assuming a multiplicity of responders to a hazardous waste emergency, all with different needs, it's essential that all of them readily can parse the basics of a warning message. But it's also possible that their more detailed data need will diverge. Fire, ambulance, police and toxic cleanup crews each may need the ability to distill, from a message, a *different* answer to the facially-identical question "where's the problem?" The ability to unpack different levels of meaning from a single message may be very useful.

Reconciling meaning: the UN/CEFACT Core Components project

In 1999, a large group of data and software vendors and users, including a significant governmental delegation, launched the electronic Business XML (ebXML) project [12], to define and then build a stack of open XML specifications capable of servicing general business data exchange needs. One of the most important, and difficult, layers of this work was the "Core Components" technical specification (CCTS) to harmonize simple data building blocks across a broad base of users. UN/CEFACT approved CCTS v2 and sent it to join the rest of ebXML [13] as part 5 of ISO Technical Specification 15000 in 2004.

[12] <http://www.ebxml.org>

[13] <http://www.iso.org/iso/en/commcentre/pressreleases/archives/2004/Ref904.html>

CCTS built on the general data element harmonization goals of ISO/IEC TS 11179 [14], which recommended that recurring elements of commercial metadata be organized in a structure of definitions, identifiers and categories, generally using object-oriented methods and shared registries. As in the original ebXML project plan, CCTS "components" are intended to be (a) maintained and harmonized as a growing, global common set of resources, from which messages can be composed, and (b) deployed from cooperative repositories. CCTS defines some specific elemental data building blocks (such as parties, addresses and quantities); a logical model for the meaning of each data component; and methods for producing expressions of the logical data, such as into W3C's XML Schema or Schematron. CEFACt gives each basic construct of data has a normative textual description, and a set of elements expressed in the modeling method UML [15].

[14] <http://metadata-stds.org/11179/>

[15] <http://www.uml.org/>

CEFACT's CCTS methodology is intended to build a growing, globally-useful "Core Component Library" (UN/CCL). The CC Library anticipates contributions of business messages and data components in current use. Early submissions since 2002 already have been woven into several early releases of the UN/CCL. But arguably, submissions and processing of new components are the most challenging feature of the Library. That methodology specifies that the reconciliation of contributions to the current approved pool is conducted, essentially, by hand by a committee. A designated CEFACt

"harmonization committee" adjudicates the uniqueness (or duplicative character) of each new submission, and conforms it where needed to the CCTS model requirements.

Consider what happens when a putatively new data entity is submitted. A number of different apples-to-apples scenarios are possible. The submitter of an "apple" component might offer that data element to CEFACT, arguing that it is not adequately serviced by the Library's existing "apple". Perhaps it is a special apple. Perhaps it's a pomme de terre, and not a pomme at all. Or, skeptically, perhaps a submitter simply wishes to assert a difference, so as to be relieved of the burden of conforming to the pre-existing approved model for apples.

CCTS provides a detailed model-to-model interrogation method; but inevitably it depends on judgment and juried evaluation. This makes the submission and processing of new contributed components potentially very time consuming, and even subjective. Knowledge representation, particularly in resolving conflicting taxonomies or viewpoints, may have much to offer a manual process of this sort. More deterministic methods of mediating between definitions (of "location", "payment" or "security level", for example) might simplify some potential collisions between existing messages that putatively represent the same phenomena. Some recent discussions already have started about possible relationships between the CEFACT UML model's objects and instances, and RDF triples.

Conclusion

Data specifications generally, and in public administration acutely, often weave together multiple disparate elements for coherent re-use. [16]. "Not invented here" rarely seems to apply to routine governmental data, outside of high-security fields. Public agencies often must assemble their information from multiple and sometimes unanticipated sources. The ability to synthesize and combine mash-ups of available data, and taxonomies, rather than impose one top-down view or one exclusive set of required methods, appears especially important.

[16] Another public sector "mash-up" example is the OASIS TaxXML TC's 2005 report to the Organisation for Economic Co-operation and Development (OECD), which considers combinations of XBRL [<http://www.xbrl.org/>], UBL, OAGI, the OASIS CIQ address specifications [<http://www.oasis-open.org/committees/ciq/>] and several tax-specific OECD recommendations. See http://www.oasis-open.org/committees/download.php/14242/OASIS_XML_Position_Paper_for_Tax_Administrations_v2-01.pdf. In our field, these combinatorial approaches seem to be becoming the norm for government users.

Ready augmentation of meaning, and the ability to serve and reconcile multiple different views, are proving a necessary part of such a heterogeneous system.

OASIS' own e-Government-focused committees and constituencies have instructed us to place high value on extensibility, flexibility, and interoperability between standards and standards organizations. Our users expect our larger community of work product to work together well. Creative and alternative combinations of standards -- and the ability to exchange deeper meaning across various spontaneous consumers -- are essential capabilities for building powerful specifications for public administration.