What is Linked Data?

by Mike Bergman - Monday, June 23, 2008

http://www.mkbergman.com/447/what-is-linked-data/

We Offer a Definition and Some Answers to Enterprise Questions

The recent LinkedData Planet conference in NYC marked, I think, a real transition point. The conference signaled the beginning movement of the Linked Data approach from the research lab to the enterprise. As a result, there was something of a schizophrenic aspect at many different levels to the conference: business and research perspectives; realists and idealists; straight RDF and linked data RDF; even the discussions in the exhibit area versus some of the talks presented from the podium.

Like any new concept, my sense was a struggle around terminology and common language and the need to bridge different perspectives and world views. Like all human matters, communication and dialog were at the core of the attendees' attempts to bridge gaps and find common ground. Based on what I saw, much great progress occurred.

The reality, of course, is that Linked Data is still very much in its infancy, and its practice within the enterprise is just beginning. Much of what was heard at the conference was theory versus practice and use cases. That should and will change rapidly.

In an attempt to help move the dialog further, I offer a definition and Structured Dynamics' perspective to some of the questions posed in one way or another during the conference.

Linked Data Defined

Sources such as the four principles of Linked Data in Tim Berners-Lee's Design Issues: Linked Data and the introductory statements on the Linked Data Wikipedia entry approximate -- but do not completely express -- an accepted or formal or "official" definition of Linked Data *per se*. Building from these sources and attempting to be more precise, here is the definition of Linked Data we used internally:

Linked Data is a set of best practices for publishing and deploying instance and class data using the RDF data model, naming the data objects using uniform resource identifiers (URIs), and exposing the data for access via the HTTP protocol, while emphasizing data interconnections, interrelationships and context useful to both humans and machine agents.

All references to Linked Data below embrace this definition.

Some Clarifying Questions

I'm sure many other questions were raised, but listed below are some of the more prominent ones I heard in the various conference Q&A sessions and hallway discussions.

1. Does Linked Data require RDF?

Yes. Though other approaches can also model the first order predicate logic of *subject-predicate-object* at the core of the Resource Description Framework data model, RDF is the one based on the open standards of the W3C. RDF and FOL are powerful because of simplicity, ability to express complex schema and relationships, and suitability for modeling all extant data frameworks for unstructured, semi-structured and structured data.

2. Is publishing RDF sufficient to create Linked Data?

No. Linked Data represents a set of techniques applied to the RDF data model that names all objects as URIs and makes them accessible via the HTTP protocol (as well as other considerations; see the definition above and further discussion below).

Some vendors and data providers claim Linked Data support, but if their data is not accessible via HTTP using URIs for data object identification, it is not Linked Data. Fortunately, it is relatively straightforward to convert non-compliant RDF to Linked Data.

3. How does one publish or deploy Linked Data?

There are some excellent references for how to publish Linked Data. Examples include a tutorial, How to Publish Linked Data on the Web, and a white paper, Deploying Linked Data, using the example of OpenLink's Virtuoso software. There are also recommended approaches and ways to use URI identifiers, such as the W3C's working draft, Cool URIs for the Semantic Web.

However, there are not yet published guidelines for also how to meet the Zitgist definition above where there is also an emphasis on class and context matching. A number of companies and consultants, including Zitgist, presently provide such assistance.

The key principles, however, are to make links aggressively between data items with appropriate semantics (properties or relations; that is, the *predicate* edges between the *subject* and *object* nodes of the triple) using URIs for the object identifiers, all being exposed and accessible via the HTTP Web protocol.

4. Is Linked Data just another term or branding for the Semantic Web?

Absolutely not, though this is a source of some confusion at present.

The Semantic Web is probably best understood as a vision or goal where semantically rich annotation of data is used by machine agents to make connections, find information or do things automatically in the

background on behalf of humans. We are on a path toward this vision or goal, but under this interpretation the Semantic Web is more of a process than a state. By understanding that the Semantic Web is a vision or goal we can see why a label such as 'Web 3.0' is perhaps simplistic and incomplete.

Linked Data is a set of practices somewhere in the early middle of the spectrum from the initial Web of documents to this vision of the Semantic Web. (See my earlier post at bottom for a diagram of this spectrum.)

Linked Data is here today, doable today, and pragmatic today. Meaningful semantic connections can be made and there are many other manifest benefits (see below) with Linked Data, but automatic reasoning in the background or autonomic behavior is not yet one of them.

Strictly speaking, then, Linked Data represents doable best practices today within the context both of Web access and of this yet unrealized longer-term vision of the Semantic Web.

5. Does Linked Data only apply to instance data?

Definitely not, though early practice has been interpreted by some as such.

One of the stimulating, but controversial, keynotes of the conference was from Dr. Anant Jhingran of IBM, who made the strong and absolutely correct observation that Linked Data requires the interplay and intersection of *people*, *instances* and *schema*. From his vantage, early exposed Linked Data has been dominated by instance data from sources such as Wikipedia and have lacked the schema (class) relationships that enterprises are based upon. The people aspect in terms of connections, collaboration and joint buy-in is also the means for establishing trust and authority to the data.

In Zitgist's terminology, class-level mappings 'explode the domain' and produce information benefits similar to Metcalfe's Law as a function of the degree of class linkages [1]. While this network effect is well known to the community, it has not yet been shown much in current Linked Data sets. As Anant pointed out, schemas define enterprise processes and knowledge structures. Demonstrating schema (class) relationships is the next appropriate task for the Linked Data community.

6. What role do "ontologies" play with Linked Data?

In an RDF context, "ontologies" are the vocabularies and structures that capture the schema structures noted above. Ontologies embody the class and instance definitions and the predicate (property) relations that enable legacy schemas and data to be transformed into Linked Data graphs.

Though many public RDF vocabularies and ontologies presently exist, and should be re-used where possible and where the semantics match the existing legacy information, enterprises will require specific ontologies reflective of their own data and information relationships.

Despite the newness or intimidation perhaps associated with the "ontology" term, ontologies are no more complex -- indeed, are simpler and more powerful -- than the standard relational schema familiar to enterprises. If you'd like, simply substitute *schema* for *ontology* and you will be saying the same thing in an RDF context.

7. Is Linked Data a centralized or federated approach?

Neither, really, though the rationale and justification for Linked Data is grounded in federating widely disparate sources of data that can also vary widely in existing formalism and structure.

Because Linked Data is a set of techniques and best practices for expressing, exposing and publishing data, it can easily be applied to either centralized or federated circumstances.

However, the real world where any and all potentially relevant data can be interconnected is by definition a varied, distributed, and therefore federated world. Because of its universal RDF data model and Web-based techniques for data expression and access, Linked Data is the perfect vehicle, finally, for data integration and interoperability without boundaries.

8. How does one maintain context when federating Linked Data?

The simple case is where two data sources refer to the exact same entity or instance (individual) with the same identity. The standard `sameAs` predicate is used to assert the equivalence in such cases.

The more important case is where the data sources are *about* similar subjects or concepts, in which case a structure of well-defined reference classes is employed. Furthermore, if these classes can themselves be expressed in a graph structure capturing the relationships amongst the concepts, we now have some fixed points in the conceptual information space for relating and tieing together disparate data. Still further, such a conceptual structure also provides the means to relate the people, places, things, organizations, events, etc., of the individual instances of the world to one another as well.

Any reference structure that is composed of concept classes that are properly related to each other may provide this referential "glue" or "backbone".

One such structure provided in open source by Zitgist is the 21,000 subject concept node structure of UMBEL, itself derived from the Cyc knowledge base. In any event, such broad reference structures may often be accompanied by more specific domain conceptual ontologies to provide focused domain-specific context.

9. Does data need to be "open" to qualify as Linked Data?

No, absolutely not.

While, to date, it is the case that Linked Data has been demonstrated using public Web data and many desire to expose more through the open data movement, there is nothing preventing private, proprietary or subscription data from being Linked Data.

The Linking Open Data (LOD) group formed about 18 months ago to showcase Linked Data techniques began with open data. As a parallel concept to sever the idea that it only applies to open data, François-Paul Servant has specifically identified Linking Enterprise Data (and see also the accompanying slides).

For example, with Linked Data (and not the more restrictive LOD sense), two or more enterprises or

private parties can legitimately exchange private Linked Data over a private network using HTTP. As another example, Linked Data may be exchanged on an intranet between different departments, etc.

So long as the principles of URI naming, HTTP access, and linking predicates where possible are maintained, the approach qualifies as Linked Data.

10. Can legacy data be expressed as Linked Data?

Absolutely yes, without reservation. Indeed, non-transactional legacy data perhaps *should* be expressed as Linked Data in order to gain its manifest benefits. See #14 below.

11. Can enterprise and open or public data be intermixed as Linked Data?

Of course. Since Linked Data can be applied to any data formalism, source or schema, it is perfectly suited to integrating data from inside and outside the firewall, open or private.

12. How does one query or access Linked Data?

The basic query language for Linked Data is SPARQL (pronounced "sparkle"), which bears close resemblance to SQL only applicable to an RDF data graph. The actual datastores applied to RDF may also add a fourth aspect to the tuple for graph namespaces, which can bring access and scale efficiencies. In these cases, the system is known as a "quad store". Additional techniques may be added to data filtering prior to the SPARQL query for further efficiencies.

Templated SPARQL queries and other techniques can lead to very efficient and rapid deployment of various Web services and reports, two techniques often applied by Zitgist and other vendors. For example, all Zitgist DataViewer views and UMBEL Web services are expressed using such SPARQL templates.

This SPARQL templating approach may also be combined with the use of templating standards such as Fresnel to bind instance data to display templates.

13. How is access control or security maintained around Linked Data?

In Zitgist's view, access control or security occurs at the layer of the HTTP access and protocols, and not at the Linked Data layer. Thus, the same policies and procedures that have been developed for general Web access and security are applicable to Linked Data.

However, standard data level or Web server access and security *can* be enhanced by the choice of the system hosting the data. Zitgist, for example, uses OpenLink's Virtuoso universal server that has proven and robust security mechanisms. Additionally, it is possible to express security and access policies using RDF ontologies as well. These potentials are largely independent of Linked Data techniques.

The key point is that there is nothing unique or inherent to Linked Data with respect to access or control or security that is not inherent with standard Web access. If a given link points to a data object from a source that has limited or controlled access, its results will not appear in the final results graph for those

users subject to access restrictions.

14. What are the enterprise benefits of Linked Data? (Why adopt it?)

For more than 30 years -- since the widespread adoption of electronic information systems by enterprises -- the Holy Grail has been complete, integrated access to all data. With Linked Data, that promise is now at hand. Here are some of the key enterprise benefits to Linked Data, which provide the rationales for adoption:

- Via the RDF model, equal applicability to unstructured, semi-structured, and structured data and content
- Elimination of internal data "silos"
- Integration of internal and external data
- Easy interlinkage of enterprise, industry-standard, open public and public subscription data
- Complete data modeling of any legacy schema
- Flexible and easy updates and changes to existing schema
- An end to the need to re-architect legacy schema resulting from changes to the business or M & A
- Report creation and data display based on templates and queries, not IT departments
- Data access, analysis and manipulation pushed out to the user level, and, generally
- The ability of internal Linked Data stores to be maintained by existing DBA procedures and assets.

15. What are early applications or uses of Linked Data?

Linked Data is well suited to traditional knowledge base or knowledge management applications. Its near-term application to transactional or material process applications is less apparent.

Of special use is the value-added from connecting existing internal and external content via the network effect from the linkages [1].

A Hearty Thanks

Johnnie Linked Data is starting to grow up. Our little semantic Web toddler is moving beyond *ga-ga-goo-goo* to saying his first real sentences. Language acquisition will come rapidly, and, like what all of us have seen with our own children, they will grow up faster than we can imagine.

There were so many at this meeting that had impact and meaning to this exciting transition point that I won't list specific names at risk of leaving other names off. Those of you who made so many great observations or stayed up late interacting with passion know who you are. Let me simply say: Thanks!

The LinkedData Planet conference has shown, to me, that enterprises are extremely interested in what our community has developed and now proven. They are asking hard questions and will be difficult task masters, but we need to listen and respond. The attendees were a selective and high-quality group, understanding of their own needs and looking for answers. We did an OK job of providing those answers, but we can do much, much better.

I reflect on these few days now knowing something I did not truly know before: the market is here and it is real. The researchers who have brought us to this point will continue to have much to research. But, those of us desirous of providing real pragmatic value and getting paid for it, can confidently move forward knowing both the markets and the value are real. Linked Data is not magic, but when done with quality and in context, it delivers value worth paying for.

To all of the fellow speakers and exhibitors, to all of the engaged attendees, and to the Juperitermedia organizers and Bob DuCharme and Ken North as conference chairs, let me add my heartfelt thanks for a job well done.

Next Steps and Next Conference

The next LinkedData Planet conference and expo will be October 16-17, 2008, at the Santa Clara Hyatt in Santa Clara, California. The agenda has not been announced, but hopefully we will see a continuing enterprise perspective and some emerging use cases.

Zitgist as a company will continue to release and describe its enterprise products and services, and I will continue to blog on Linked Data matters of specific interest to the enterprise. Pending topics include converting legacy data to Linked Data, converting relational data and schema to Linked Data, placing context to Linked Data, and many others. We think you will like the various announcements as they arise. ;)

Zitgist is also toying with the use of a distinctive icon to indicate the availability of Linked Data conforming to the principles embodied in the questions above. (The color choice is an adoption of the semantic Web logo from the W3C.) The use of a distinctive icon is similar to what RSS feeds or microformats have done to alert users to their specific formats. Drop me a line and let us know what you think of this idea.

---

[1] Metcalfe's law states that the value of a telecommunications network is proportional to the square of the number of users of the system(n^2), where the linkages between users (nodes) exist by definition. For information bases, the data objects are the nodes. Linked Data works to add the connections between the nodes. We can thus modify the original sense to become Zitgist's Law: the value of a Linked Data network is proportional to the square of the number of links between the data objects.
