

A Report on Linked Data

Aman Shakya, Hideaki Takeda
 National Institute of Informatics
 2-1-2 Hitotsubashi, Chiyoda-ku,
 Tokyo, Japan 101-8430
 {shakya_aman, takeda}@nii.ac.jp

ABSTRACT

Linked data is a recent practice of publishing and interlinking structured data on the Semantic Web. This report gives a brief introduction to linked data and best practices. Then the report presents experiences from the Linked Data Planet conference, June 2008 which provided insights to industry professionals about linked data.

1. INTRODUCTION

The term Linked Data was coined by Sir Tim Berners-Lee in his Linked Data Web architecture note [1]. It refers to a style of publishing and interlinking structured data on the Web. The Wikipedia definition of Linked Data [2] is, "Linked Data is a term used to describe a method of exposing, sharing, and connecting data on the Semantic Web. The practice emphasizes Web access to data using existing Web technologies such as URIs and HTTP".

Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. The significance of Linked Data is in the fact that the value and usefulness of data increases the more it is interlinked with other data. It provides a data commons on the Web, where people and organizations can post and consume data about anything. This common data network is often called the Web of Data.

2. BASICS OF LINKED DATA

Berners-Lee outlined the following four rules of Linked Data in his design issues notes [1].

1. Use URIs to identify things that you expose to the Web as resources.
2. Use HTTP URIs so that people can locate and look up (dereference) these things.
3. Provide useful information about the resource when its URI is dereferenced.
4. Include links to other, related URIs in the exposed data as a means of improving information discovery on the Web.

Semantic Web requires that resources are identified by the universal URI set of symbols. There is a tendency for people to invent new URI schemes such as LSIDs, URN schemes, XRIs, DOIs and so on. It is highly recommended to use HTTP URI so that we can inherit all the HTTP mechanisms already in place.

The standard Web transfer protocol, HTTP, should be used to be "on the Web".

There are many significant data stores existing today. However, many times the data, if available at all, is buried in a zip archive somewhere. Linked data is about making data available in standard ways so that others can use and link to. This is essential to connect the data we have into a web. It is the unexpected re-use of information which is the value added by the web.

2.1 Two Types of Resources

It is important to distinguish between the resources found in the current document Web and the real world resources identified in the Linked Data Web. All the resources we find on the traditional document Web are *information resources*. Information resources can have representations in the form of file, byte stream, etc. All "real-world objects" that exist outside of the Web are *non-information resources* (also called '*other resources*'). There should be no confusion between identifiers for documents and identifiers for other resources.

2.2 HTTP and Content Negotiation

Content negotiation is a useful mechanism offered by HTTP by which clients can request for desired type of contents from the server. When a user agent makes an HTTP request, it sends along some HTTP headers to indicate what data formats and language it prefers. For RDF/XML, the standard serialization format of RDF, the content type is 'application/rdf+xml'. Content negotiation thus allows publishers to serve HTML versions of a web document to traditional web browsers and RDF versions to semantic web-enabled user agents. If the headers indicate that the client prefers HTML, then the server can generate an HTML representation. If the client prefers RDF, then the server can generate RDF.

Therefore, a data source often serves three URIs related to each non-information resource, for instance:

- <http://www4.wiwiw.fu-berlin.de/factbook/resource/Russia>
(URI identifying the non-information resource Russia)
- <http://www4.wiwiw.fu-berlin.de/factbook/data/Russia>
(information resource with an RDF/XML representation describing Russia)
- <http://www4.wiwiw.fu-berlin.de/factbook/page/Russia>
(information resource with an HTML representation describing Russia)

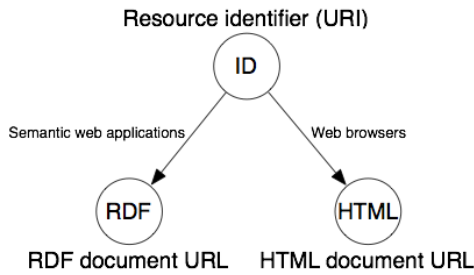


Figure 1. Three URIs served for a non-information resource.

2.3 Dereferencing HTTP URIs

URI Dereferencing is the process of looking up a URI on the Web in order to get information about the referenced resource. Information resources are dereferenced directly by returning a representation of the resource along with HTTP response code 200 OK. But non-information resources cannot be dereferenced directly. There are two approaches proposed which may be used to dereference such URIs - Hash URIs and 303 redirects.

2.3.1 Hash URIs

When a client wants to retrieve a hash URI, the HTTP protocol requires the fragment part to be stripped off before requesting the URI from the server. So a URI that includes a hash cannot be retrieved directly, and therefore cannot identify a web document. But we can use them to identify other, non-document resources, without creating ambiguity.

For example an organization Example Inc. could use the following URI to represent the person Alice.

`http://www.example.com/about#alice`

The information about Alice can be easily looked-up as follows

1. Form the URI of the document by truncating before the hash
2. Access the document to obtain information about #alice

`http://www.example.com/about#alice`

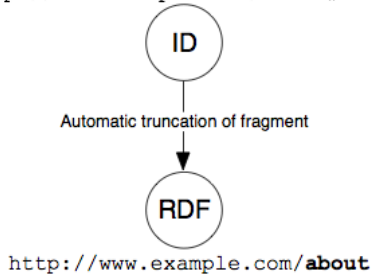


Figure 2. Dereferencing a Hash URI

Hash URIs can also be used with content negotiation as illustrated below so that both HTML and RDF representations of the resource can be served to human and machines respectively.

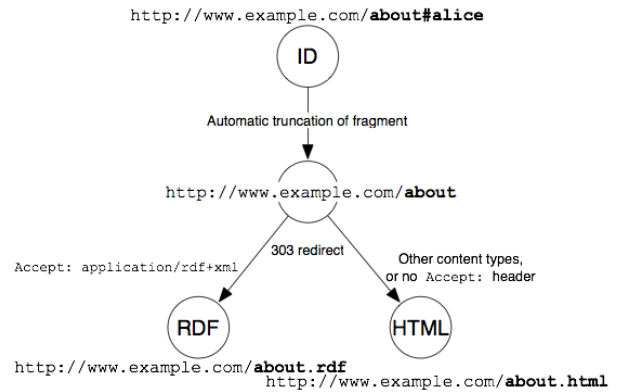


Figure 3. Using hash URI with content negotiation.

2.3.2 303 URIs

Alternatively, the server can be made to respond to the URI requests with a 303 status code and the URL of a document that describes the resource. In a second step, the client dereferences this new URI and gets a representation describing the non-information resource. It is illustrated in the following figures.

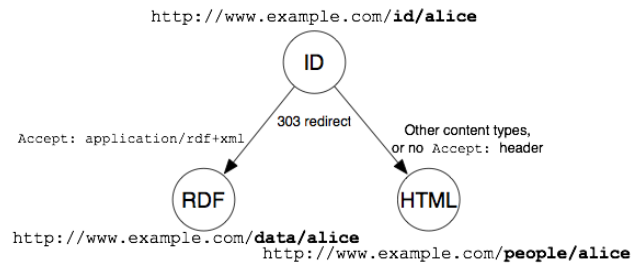


Figure 4. Handling 303 URIs.

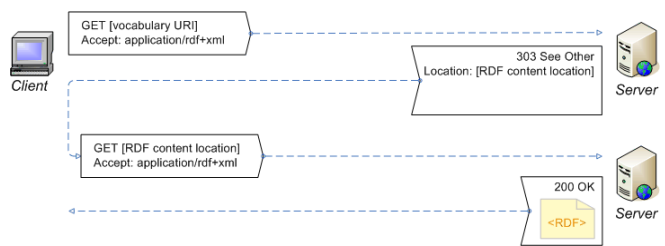


Figure 5. Dereferencing 303 URIs.

2.3.3 Choosing between Hash and 303 URIs

The advantage of hash URIs is that a family of URIs can share the same non-hash part. So a group of descriptions can be retrieved with a single request to the document describing the resources. However, the disadvantage is that a client interested only in one resource will have to load the data for all other resources. 303 URIs, on the other hand, are very flexible because the redirection target can be configured separately for each resource. But the large number of redirects may cause higher latency.

Thus, hash URIs should be preferred for rather small and stable sets of resources that evolve together. Ideal cases are RDF Schema vocabularies and OWL ontologies, where the terms are

often used together, and the number of terms is unlikely to grow much in the future. 303 URIs should be used for large sets of data that are, or may grow, beyond the point where it is practical to serve all related resources in a single document. If in doubt, it's better to use the more flexible 303 URI approach.

2.4 Cool URIs

Some guidelines have been set up to create good URIs.

Simplicity. It is better to have short memorable names. For instance *http://dbpedia.org/resource/Berlin* is better than *http://www4.wiwiss.fu-berlin.de:2020/demos/dbpedia/cgi-bin/resources.php?id=Berlin*

If possible we may clean up complex URIs by adding some URI rewriting rules to the configuration of the web server.

Stability. We should try to keep the URIs stable and persistent so that we do not have to change the links in the future resulting in broken links. It is better to keep implementation-specific bits and pieces such as .php and .asp out of the URIs. We may want to change technologies in future.

Manageability. We should keep our URIs manageable. Define your URIs in an HTTP namespace under your control. Often some kind of primary key is needed inside the URIs, to make sure that each one is unique.

2.5 RDF description for a URI

It is recommended that the following information be returned when the URI for a resource is dereferenced.

1. **The description** - all triples from the dataset that have the resource's URI as the subject.
2. **Backlinks** - all triples from the dataset that have the resource's URI as the object.
3. **Related descriptions** - additional information about related resources that may be of interest in typical usage scenarios.
4. **Metadata** - such as a URI identifying the author and licensing information.

The data source should at least provide RDF descriptions as RDF/XML. Other formats like Turtle or TriX may also be provided.

3. LINKING DATA

Links among structured data elements are made using RDF links or predicates. Usually, the application domain will determine which RDF properties are used as predicates. We should reuse terms from well-known vocabularies wherever possible. Some of the recommended well-known vocabularies are given below.

- Friend-of-a-Friend (FOAF)
- Dublin Core (DC)
- Semantically-Interlinked Online Communities (SIOC)
- Description of a Project (DOAP)
- Simple Knowledge Organization System (SKOS)

- Music Ontology
- Review Vocabulary
- Creative Commons (CC)

A more extensive list of well-known vocabularies is maintained by the W3C SWEO Linking Open Data community project [5]. For URI references of geographic places, research areas, general topics, artists, books or CDs, we should consider using URIs from data sources within the W3C SWEO Linking Open Data community project [5], for instance Geonames, DBpedia, Musicbrainz, dbtune or the RDF Book Mashup.

Defining New Terms. New terms should only be defined if they do not exist in the well-known vocabularies. Some guidelines are as follows.

- Do not define new vocabularies from scratch
- Provide for both humans and machines (for eg, add `rdfs:comments`, `rdfs:label` for terms and properties.)
- State all important information explicitly. For eg, state all ranges and domains explicitly.
- Do not create over-constrained, brittle models; leave some flexibility for growth. Therefore, unless you know exactly what you are doing, use RDF-Schema to define vocabularies instead of OWL.

We should use the RDF Data Model to have linked data. However, some features of RDF are best avoided in the linked data context.

- The use of blank nodes is discouraged.
- The use of RDF reification is also discouraged as the semantics of reification are unclear and as reified statements are rather cumbersome to SPARQL.
- It is better not to use RDF collections or RDF containers as they do not work well with SPARQL.

3.1 URI aliases

URI aliases are common on the Web of Data. It is common practice that information providers set `owl:sameAs` links to URI aliases they know about. An `owl:sameAs` link indicates that two URI references actually refer to the same thing. Such RDF links can be set manually or they can be generated by automated linking algorithms.

3.1.1 Automatic Data Linking

We can use an automated record linkage algorithm to generate RDF links between data sources. Record Linkage is a well-known problem in the databases community. However, there is still a lack of good, easy-to-use tools to auto-generate RDF links. Therefore it is common practice to implement dataset-specific record linkage algorithms to generate RDF links.

Pattern-based Algorithms. If certain identifiers (for eg, ISBN numbers) are used as part of HTTP URIs, it is possible to use simple pattern-based algorithms to generate RDF links between the resources.

More complex property-based Algorithms have also been successfully demonstrated which use multiple features. For eg,

-Interlinking DBpedia and Geonames.

-Interlinking Jamendo and MusicBrainz.

3.2 Serving Information as Linked Data

3.2.1 Serving Static RDF Files

The easiest way to serve linked data is to put static RDF files in the server. Hash URIs have to be used in that case as 303 redirect URIs are not possible.

3.2.2 Serving Relational Databases

There are some tools available to expose data in relational databases as linked data. The D2R Server is such a tool for serving Linked Data views on relational databases. We only have to provide the declarative mapping between the schemata of the database and the target RDF terms. D2R Server also provides a SPARQL endpoint for the database.

Alternatively, some other software available for producing linked data from databases are

- OpenLink Virtuoso
- Triplify (a small plugin for different Web applications)

3.2.3 Serving other Types of Information

A lot of structured data is also available in other formats like CSV, Microsoft Excel, or BibTEX. There are some tools available to convert them into RDF. For eg, ConverterToRdf, RDFizers, etc.

Pubby is a tool that can be used to add Linked Data interfaces to SPARQL endpoints. The DBpedia project uses Pubby in front of the SPARQL endpoint over an OpenLink Virtuoso repository.

3.2.4 Implementing Wrappers around existing Applications or Web APIs

Some examples of this are

- The RDF Book Mashup – It requests data about the book, its author as well as reviews and sales offers from the Amazon API and the Google Base API.
- SIOC Exporters for WordPress, Drupal, phpBB
- Virtuoso Sponger

4. EXPLORING LINKED DATA

Linked data can be crawled by search engines following the RDF links. Additionally, we can have the following to increase the chance of our data getting indexed.

- Ping the Semantic Web
- HTML Link Auto-Discovery: As shown in the figure below we may link to RDF descriptions using a `<link>` element in the HTML header.

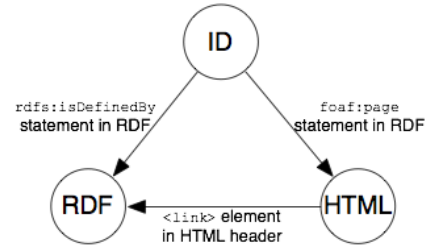


Figure 6. HTML link auto-discovery.

- Semantic Web Crawling: a Sitemap Extension
- Dataset List on the ESW Wiki: We can add out new data sources to the ESW Wiki datasets list [5] in the category Linked Data and SPARQL endpoint list.

We may also set several RDF links from our FOAF profile to the URIs of central resources.

Linked Data Browsers. There are a couple of generic Linked Data browsers available today. Some are listed below.

- Tabulator
- OpenLink RDF Browser
- Disco
- Zitgist Data Viewer, etc

5. THE LINKED DATA PLANET

The Linked Data Planet Conference and Expo, spring 2008, June 17-18 was held at the Roosevelt Hotel, New York City, USA [8]. The conference was mainly aimed at providing industry professionals with insights into the technologies that will enable them to

- connect data contained in silos within organizations in a meaningful way
- extract and correlate data from web sites and databases for analyzing trends and decision support, customer and vendor relationship management and social networking.

The evolution of the current Web of “linked documents” to a Web of “linked data” is steadily gaining mindshare among developers, architects, systems integrations, users, and the more than 200 software companies developing semantic web-oriented solutions. Notable examples on the Web today include, DBpedia, the Zoominfo search engine, the Bambora travel recommendation site, a number of social networking sites, numerous semantic web technology-based services, various linked data browsers, SPARQL query language and protocol-compliant data servers and data management systems, and a growing number of web sites exposing machine-readable data using microformats, RDFa, and GRDDL.

The LinkedData Planet audience mainly included system architects, enterprise architects, software developers, consultants and technical managers, mostly seeking to learn about linking data sources and technologies to get more value from their data.

5.1 Conference Organization

The conference was co-chaired by Bob DuCharme, solutions architect/author, Innodata Isogen and Ken North, author, consultant, Ken North Computing, LLC. The sponsors and partners were a number of organizations working with Semantic Web technologies.

Silver Sponsors

- Talis Group
- Ontotext Lab
- Franz Inc.
- OpenLink Software Inc.
- Calais, Reuters

Association and Analyst Partners

- World Wide Web Consortium
- Semantic Web Company
- OASIS (Organization for the Advancement of Structured Information Standards)
- New York Semantic Web
- NY XML SIG

The event was covered by a long list of media partners. The active involvement of so many organizations demonstrated the growing interest about linked data.

5.2 Exhibitors

The event showcased a number of exhibitors including

- Calais, Powered by Thomson Reuters
- Franz Inc. (AllegroGraph)
- Netrics
- New York XML Special Interest Group
- Ontotext Lab
- OpenLink Software, Inc.
- Relevant Digital
- TopQuadrant Inc.
- Zitgist LLC

Calais is a Thomson Reuters initiative that supports the interoperability of content and development of rich semantic applications. Calais 2.0 was demonstrated at the event. The Calais Web service enables publishers, bloggers and sites to automatically metatag the people, places, facts and events in their content with the help of natural language processing technologies. OpenCalais is a suite of various tools that will take plain-text and automatically embed metadata (such as

Microformats) into it. They have a Wordpress plugin called Tagaroo which reads your Wordpress blog entry as you type it, and automatically suggests tags. It also pulls down Flickr images based on these tags so they can be included with the entry.

Zitgist is a company that provides Linked Data products and services. Michael Bergman, the CEO of Zitgist was available to explain and discuss the latest products and services of the company. Some of its products related to linked data are zLinks, Zitgist DataViewer and Zitgist Linked Data Platform (zLDP). A recent project of Zitgist is **UMBEL**, which stands for Upper Mapping and Binding Exchange Layer. UMBEL has two purposes:

1. to provide a lightweight structure of subject concepts as a reference to what Web content or data "is about"; and
2. to define a variety of binding protocols for different Web data formats to map to this "backbone."

The constituent building blocks of UMBEL are WordNet, OpenCyc, Wikipedia and YAGO. UMBEL also provides a number of web services based on this top level ontology.

OpenLink Software is a company founded by Kingsley Idehen who is one of the most prominent Linked Data protagonists. The products of this company include the Virtuoso Universal Server. OpenLink Virtuoso is a platform for deploying Linked Data.

Franz Inc. exhibited the well-known AllegroGraph scalable Semantic Web framework. The framework now includes social network analysis as well.

5.3 Keynotes

The conference had an extraordinary list of keynote speakers. The keynote speakers included Kingsley Idehen, Ian Davis, Sir Tim Berners-Lee, Atanas Kiryakov and Anant Jhingran. The keynote speeches were as follows.

Creating, Deploying, and Exploiting Linked Data

-Kingsley Idehen, President and CEO, OpenLink Software Inc.

Idehen effectively tried to 'demystify' the concept of Linked Data in this first keynote. He started by talking about well-established technologies like ODBC for handling database sources. Then, he mapped the concepts to Linked Data by showing analogy between the technologies. The basic idea in both cases is to access data by reference or pointers. He argued that DSN (Data Source Name) identifies the data source. However, URIs name individual data records so that we can interlink data records. He also highlighted the advantage and importance of using HTTP for URIs. He covered the issues, technologies, and approaches to platform-independent and standards-based conceptual views of heterogeneous logical data sources across the enterprise. He discussed issues relating to production and exploitation of enterprise linked data in front of or behind the corporate firewall. He also pointed out that the line between the enterprise and individuals continue to blur regarding their data.

Sponsor Keynote: The Semantic Web as a Blue Ocean Opportunity

-Ian Davis, Chief Technology Officer and Director, Talis Group

In this keynote speech, Ian Davis emphasized that enterprises and organizations can achieve more through sharing their data and collaborating than being closed and isolated in islands. A radical transformation can take place from islands of data to densely interconnected data spaces, bringing huge value to organizations with the scale and algorithms to exploit the value in the connections. He highlighted that the Linked Data Web offers an unbound blue ocean of global commercial opportunities for enterprises and entrepreneurs.

Web of Data

-Sir Tim Berners-Lee, Director, World Wide Web Consortium

Sir Tim Berners-Lee, the inventor of the Web, pointed out that efforts for the Semantic Web are going in the right direction with the Linked Data movement stating it as “the Semantic Web done right”. The Linked Data principles, techniques and best practices form the basic substrate for the Semantic Web. He emphasized that for wider success of the Linked Data movement we should lobby government organizations, commercial information providers, etc to open their data in the Linked Data web. There is a lot of valuable data locked up in different organizations which could have been better utilized and further enhanced by joining into a single Web of data. He pointed out that to better ensure the participation of organizations in this movement to have them provide data in standard linked data formats we should never try to replace the existing well-established database solutions. Rather we should exploit the scalability and stability of the existing systems and import the wealth of underlying data into Linked Data format by additional layer of tools and mechanisms. On the other hand, he also highlighted the major challenges for Linked Data. Creating standard vocabularies for linking data types of data and having everyone agree upon the standards is a tremendous job. Ontology development is still a difficult process but is important for defining the semantics of data being linked. Developing human interfaces to the Linked Data is another important challenge. We should build proper usable interfaces to produce and consume linked data for beneficial purposes. He argues that motivation for contributing to the linked data web would be in the same line as the motivation for contributing to the current web. Federated querying of distributed linked data is also a difficult issue. Scalability on the web scale and inference are other challenges ahead.

Web 2.0, Enterprise 2.0 and Information Management

- Dr. Anant Jhingran, VP and CTO, Information Management Division, IBM

In this talk, he discussed how linked data serves the information needs of Web 2.0/3.0 and Enterprise 2.0 applications. He pointed out that currently linked data is mostly at the instance level only. However, enterprises frequently link data at schema level, and people link data at social level. So we should consider 3 spheres for having linked data, namely, instances, schema and people. He also highlighted that there needs to be a virtuous cycle of linked

data and value creation which in turn produces more linked data and more value out of it. Simply building up linked data web does not necessarily add value. He feels that the vision still lacks compelling use cases. Moreover, there are still issues like quality of information, ownership etc that need to be addressed.

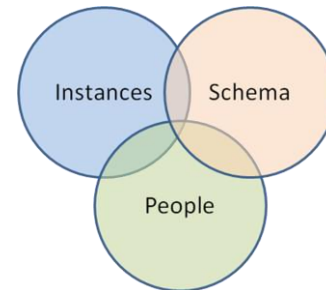


Figure 7. Linking by instance, schema and people.

Sponsor Keynote: Scalable Semantics User's Guide

- Atanas Kiryakov, Head, Ontotext Lab (Sirma Group)

In his keynote, Atanas Kiryakov mostly discussed about the scalability issues of Semantic Web technologies in large scale applications. Scalability is an important concern among enterprises and gradually being tackled by recent cutting edge technologies.

5.4 Sessions

There were several interesting parallel sessions. Some are mentioned in the following text.

How to efficiently publish and locate linked data

Richard Cyganiak, Engineer Researcher, Digital Enterprise Research Institute (DERI)

In this tutorial, the speaker demonstrated the Sindice API and presented simple examples of applications which can be created by interfacing to it. Sindice is a Semantic Web crawler and search engine developed by DERI. The API enables us to query the Semantic Web of linked data from our own applications. Some examples were shown. Details of the Semantic Sitemap were explained. Semantic Sitemap is a new proposed standard to effectively locate linked data so that huge amounts of data can be used by Semantic Web crawlers and clients. Actually, it is just an extension to the existing Sitemap protocol. It is written as an XML document.

Linked Data: The Real Web 2.0

Uche Ogbuji, Partner, Zepheira

The speaker pointed out that Linking Open Data (LOD) is simple, readily adaptable by Web developers, and complements many other popular Web trends. Open data is the real substance of Web 2.0, and not flashy AJAX effects. He stressed that Linked Data makes data more widely used by making its components easier to discover, more valuable, and easier for people to reuse. DBpedia can be used as a nucleus to explain Linked Data, its

simplicity and potential. He also talked about “Linked Enterprise Data”, or LED, which concerns deploying Linked Data within different enterprises by wrapping over existing information system rather than replacing them. Linked Data need not necessarily be public. For example, with Linked Data, two or more enterprises or private parties can legitimately exchange private Linked Data over a private network using HTTP.

Linked Data Workshop

The Linked Data workshop was in the form of a panel discussion with Bob DuCharme, Solutions Architect and Author / Conference Chair, Innodata Isogen as the moderator and the following panelists.

- Dr. Melliyal Annamalai, Principal Product Manager, Oracle
- Michael Bergman, CEO, Zitgist LLC
- Stefanos Damianakis, President & CEO, Netrics
- Uche Ogbuji, Partner, Zepheira
- Nikita Ogievetsky, Vice President, Morgan Stanley
- Walter Perry, Managing Director, Fiduciary Automation
- Dr. Andy Seaborne, Research Scientist, Hewlett-Packard Research Laboratories

There were mainly discussions on the issues about Linked Data architecture and system development highlighting important areas such as data access control, using distributed data vs. aggregated data, etc. The discussions were based around questions submitted by the audience.

The Fellowship of the Web: The Two Towers

Dr. James A. Hendler, Tetherless World Senior Constellation Professor, Rensselaer Polytechnic Institute

In this talk, Jim Hendler highlighted two approaches to ontologies and the Semantic Web. One direction sees Ontology (with capital O) as formal representation with heavy weight semantics while another direction keeps ontology (with small o) less formal with lightweight semantics only. With heavy semantics powerful reasoning can be done and successful applications have been demonstrated in enterprise scales. However, such systems cannot tolerate any inconsistency. On the other hand, with lightweight ontologies not much reasoning can be done. However, there is far less risk of inconsistencies because only little ontological agreements are in place. With little semantics, applications can scale to the web. He described how these different models can be used to link data in different ways showed different kinds of Web applications, from Enterprise Data Integration to Web 3.0 startups, and the different kinds of techniques needed for these different approaches. In addition, there were also some discussion about RDFS and OWL.

From DBpedia to OntoWiki - Emergent Data and Semantics from Social Collaboration

Sherman Monroe, CEO, Monrai Technologies

The speaker discussed technologies exploiting emergent semantic representations from social collaboration and contributions, in the light of the example applications like DBpedia and OntoWiki. He also introduced a new software called Cypher which exploits natural language processing technologies to bootstrap and enhance semantic technologies.

Using Machine Learning to Discover and Understand Structured and Unstructured Data

Dr. William Cohen, Associate Research Professor, Carnegie Mellon University Machine Learning Department

This tutorial presented how modern machine learning methods can be used to extract structured data from unstructured and semi-structured contents. The data thus obtained is accurate enough for many interesting operations like data-mining, structured queries, collaborative recommendation, classification, and set expansion. The effectiveness of these technologies was presented with experimental results.

How to Publish Linked Data on the Web

Tom Heath, Researcher, Talis Information Ltd

This was a detailed tutorial for Linked Data publishers. The session covered all the basics of linked data, principles and best practices covered in the early sections of this report.

NY Semantic Web Meetup – Panel discussion: “The Semantic Web is open for business. Are you Ready?”

The panel basically discussed around important matters essential to get the Semantic Web into the mainstream. Industry and businesses people need to see the value in adopting standards and having common knowledge base. However, the benefit for businesses to do so is still not very clear. Companies have advantage in maintaining data closed or providing access to it only through their platform. However, this is slowly changing with the growing importance of open data and emergence of data sharing platforms like Open Social. To have these technologies more widely adopted, more useful applications are needed for consuming and using the Linked Data. There are examples of applications in specific domains like life science and solving particular enterprise problems mainly of data integration. However, we need more applications. We also need more standards and ontologies which would be widely adopted. Having consensus among people is one of the most difficult thing in this. While it may be easier to have standards within enterprises, agreeing about common standards for external data is difficult. Another issue is the lack of user friendly tools that would be widely used by people. Natural language processing and artificial intelligence can help but we are yet far from perfect.

Other Sessions

There were many other lively sessions in the conference

- Starting with SPARQL : making RDF shine
- Best Practice in Semantic Systems Development
- Building a Practical Semantic Framework: The role of taxonomies and controlled vocabularies in data integration
- TripBlox : Shared Travel Information, Microformats, Ideas and Intent
- Semantic Discovery for Enterprises and Consumers using Service Oriented Architectures
- Enabling Semantic Applications Through Calais
- DITA, Semantics, Content Management, Dynamic Documents and Linked Data - A Marriage Made in Heaven?
- Integrating Relational Data Into the Semantic Web
- Applying Semantic Web Technologies to Enterprise Solutions
- Improved Services Through Behavior and Activity Recognition
- Leveraging Semantic Technology for Infrastructure Mediation
- The Social Internet, Promise or Plague in Education?
- Semantic Technology in the Real World: Challenges and Opportunities
- NY XML SIG Meeting

6. IMPRESSIONS FROM THE CONFERENCE

The extravagant event with good participation from industry professionals and entrepreneurs strengthened the claim that the Linked Data movement has certainly taken off. This would prove to be a very effective direction towards the practical realization and applications of the Semantic Web vision. Linked Data has been demonstrated as a quite well established technology by now. The basic principles and guidelines have been clearly laid out. Tutorials with practical examples and concrete implementations have served well in establishing Linked Data. A number of technological frameworks and tools are already available and in use to create and deploy Linked Data. Linked Data has become a basic substrate for the Semantic Web which can be fully implemented with current technologies and is really happening.

The Linked Data movement has gained significant momentum with distinguished leaders in academics and industry. There is already a wide participation and many more organizations are joining this effort for creating the vast Linked Data commons. The shared global database provides immense potential for businesses and entrepreneurs to monetize. Linked Data is

opening up an uncontested blue ocean of opportunities. The growing interest from the business community indicates that people are starting to realize this potential. However, this powerful technology still needs to be advertised and explained to wider range of industry professionals, government organizations and content providers. Events like this serve well in this goal. Perhaps this justifies the next Linked Data Planet, fall 2008 which will be held October 16-17, 2008! Santa Clara Hyatt, Santa Clara, CA, USA!

Unlike the big vision of the Semantic Web, Linked Data is much simpler, easier to explain and not difficult to realize. A lot of Linked Data can be produced with little effort from the developers and content providers. This is because Linked Data builds upon existing well established HTTP technology. The Linked Data community is taking the right direction by exploiting existing legacy database systems instead of trying to replace them with premature semantic technologies. Huge volumes of valuable and good quality data exist in database silos in organizations. By convincing the stakeholders to open up the data silos into a global Linked Data Web a huge Web of Data can be realized soon. This could serve well in breaking the notorious chicken and egg cycle of the Semantic Web.

However, in spite of exciting possibilities and initial success, the Linked Data movement still has severe challenges ahead. Many of these challenges also emerged out during discussions and talks in the Linked Data Planet. Some of them are as follows.

Human interface to Linked Data. Although there are technologies to convert huge amount of data into Linked Data the intuitive human interface to this Linked Data and the Web of Data is still at premature stage. There are certainly some Linked Data browsers and visualizers. However, they are still not practical for the ordinary users and have not proven to be very useful. Human interfaces should also be developed to contribute to the Linked Data Web for the real paradigm shift from document Web to data Web.

Useful applications of Linked Data. Most people seem to agree that there are still not enough applications to use the growing Web of Linked Data which really proves the value of creating it. Innovative applications should take up this challenge and exploit the apparent potential of the huge Linked Data commons. Enterprises can exploit this technology to achieve data integration and interoperability more easily and effectively at lower cost. Enterprises also have good technical benefits of Linked Data and Semantic Web technologies which could justify their adoption. Some of these are integration of internal data in the enterprise and external data, convenient data modeling of any legacy schema, flexible and easy updates and changes to existing schema, etc. However, the practical application scenarios are still less apparent.

Concerns of Data Ownership, Privacy, Security, Provenance and Trust. A barrier that is keeping organizations from joining the Linked Open Data initiative are concerns about ownership, privacy, security, provenance and trust. Though Linked Data does not necessarily require data to be public, there seems to be less understanding of this. Licensing and provenance metadata are being added by services deploying Linked Data and this would improve the trust in utilizing the Linked Data too.

However, more work needs to be done and guidelines and policies need to be outlined to address these concerns satisfactorily.

Semantics and Ontology creation. Linked Data only serves as a connected data web (basically at the instance level). However, defining the semantics of the links and creating standard ontologies and vocabularies to establish the semantics is still a challenging area. Collaborative techniques seem to be improving and promising in this area but there is still a long way to go. More powerful inferencing can be included to get best out of the semantics encoded in Linked Data. There are currently a number of vocabularies and ontologies. However, we need more to cover wider range of data and the existing ones need to be better adopted, reused and maintained.

Scalability. Scalability is another big challenge for large scale applications. The scalability of Semantic Web technologies seems to be improving with commercial enterprise scale products. However, scaling up to the scale of such huge network that the Linked Data Web will be is still a big challenge. More scalable frameworks need to be developed and tested. Triple stores are still not so scalable. It seems better to exploit the existing enterprise solutions and just view the data as Linked Data. SPARQL query processors also need to be optimized to serve complex queries. Reasoning, at this stage, is definitely not scalable for the data Web.

Record Linkage problem. Another big challenge for Linked Data and the Semantic Web is record linkage or identifying equivalence between URIs about the same things. Once URIs about the same things have been identified they can be linked with the owl:sameAs property. Unfortunately, this is still a difficult problem and little work has been done besides some heuristics to link data from specific data sources.

Schema and people level linking. As pointed out earlier, currently the Linked Data Web is basically linking instance data. However, enterprises mostly link data by integrating at schema level. On the other hand, social aspect also needs to be considered to practically connect up data that would be relevant to people.

Linked Data Analysis. The analysis of the growing Linked Data Web would also be an interesting, challenging and useful area. Mining this complex Web of data may present unanticipated results and demonstrate the value of this network. Modern machine learning techniques and network science may help us in this area but many unexplored challenges and possibilities remain ahead.

7. CONCLUSION

The recent Linked Data initiative has proven to be quite successful and the momentum is growing with more campaigns and participation. It is quite well defined feasible technology and is really creating a Web of data. Due to this, Linked Data is gradually being embraced by industry and business communities. This would be very significant movement for the practical realization of the Semantic Web. However, many challenges are still ahead and Semantic Web researchers have to look for ways to solve these issues.

8. REFERENCES

- [1] Berners-Lee, T. 2006. Design Issues: Linked Data. Last change: \$Date: 2007/05/02 14:30:56 \$.
<http://www.w3.org/DesignIssues/LinkedData.html>
- [2] Linked Data, From Wikipedia.
http://en.wikipedia.org/wiki/Linked_Data
- [3] <http://linkeddata.org/>
- [4] <http://esw.w3.org/topic/LinkedData>
- [5] <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>
- [6] Bizer, C., Cyganiak, R., Heath, T. 2007. How to Publish Linked Data on the Web. <http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
- [7] Cool URIs for the Semantic Web.
<http://www.w3.org/TR/2007/WD-cooluris-20071217/>
- [8] <http://www.linkeddataplanet.com/>