From Information to Sense-Making: Fetching and Querying Semantic Repositories

Tope Omitola, Ian C. Millard, Hugh Glaser, Nicholas Gibbins, and Nigel Shadbolt

Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
{tobo, icm, hg, nmg, nrs}@ecs.soton.ac.uk

Abstract. Information, its gathering, sharing, and storage, is growing at a very rapid rate. Information turned into knowledge leads to sense-making. Ontologies, and their representations in RDF, are increasingly being used to turn information into knowledge. This paper describes how to leverage the power of ontologies and semantic repositories to turn today's glut of information into sense-making. This would enable better applications to be built making users' lives easier and more effective.

1 Introduction

Modern digital devices have made the gathering, sharing, storing, and creating data to be a relatively painless and trivial exercise. This is evinced by the latest IDC Report[1] which reported that in 2009 alone, the amount of digital information produced in the world grew by 62% to nearly 800,000 petabytes, and in 2010 it will grow to 1.2 million petabytes. Data is being captured today as never before, while the amount of data generated by people doubles roughly every 1.5 years. The plummeting of costs of gathering and storing these data is having a tremendous impact on the expectations of individuals as they create and share information about themselves and about their relationships with others.

Humans have always gathered, stored, shared, and created data with friends and members of their tribe or clan – data on where to find good food, what places to avoid, etc. The cost and overhead of communication has always been very high and, before the advent of computation and the internet, the serendipity of someone coming across what you wanted to share has always been very low. But the ease of computation and the internet has made it easier to gather, create, and share information. Information is being generated at such a prodigious rate that the challenge now is **"sense-making"**, how do we curate information, version it, maintain it, index it, search it, query it, retrieve it, and re-use it, thereby helping people discover relevant content. The questions we should be asking ourselves are what shall we collect and what applications can we build

[1] http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

that will help users make **sense** of these data in order to make users' lives easier, more effective, and more interesting.

1.1 Information, Knowledge, and Sense-Making

In order to achieve adequate or useful sense-making from our information, they need to be turned into knowledge. By knowledge, we mean "information that can be sufficiently interpreted to enable action"[1]. To turn information into actionable knowledge, it is necessary to understand the connections between it and business and human processes. To manage this knowledge effectively, an important solution is the development and application of technologies useful to render information interpretable in order to enable effective action. Six challenges to effective knowledge management have been identified[1]. These are:

1. Knowledge Acquisition: Although we have a surfeit of information, the knowledge available in them are often insufficient or poorly-specified. The goal of Knowledge Acquisition is to turn information into usable knowledge. This could involve making tacit knowledge explicit, thereby identifying gaps in the knowledge already held, acquiring and integrating knowledge from multiple sources, and acquiring knowledge from unstructured media (e.g. natural language or free-flowing text).
2. Knowledge Modelling: Here, acquired knowledge is made usable for problem-solving by using model structures for its representation. These model structures, called ontologies, are specifications of the generic concepts, attributes, relations and axioms of a knowledge base or domain. They can act as a format for understanding how knowledge will be used.
3. Knowledge Publishing: The challenge here is to get the right knowledge, in the right form, in the right place, to the right person, at the right time. This will involve understanding the knowledge representation structures, the knowledge of the user(s), and of their context.
4. Knowledge Re-use: Knowledge representation are usually highly domain-specific, thereby making them more difficult to be re-purposed for different domains. A good solution to this problem will provide high leverage for effective knowledge use.
5. Knowledge Maintenance: This challenge involves making the knowledge repository functional. One of the key questions here is how does a consumer of a knowledge repository know that something has changed, and what has changed in a repository.
6. Knowledge Retrieval: In a world of distributed and federated knowledge bases, the challenge here includes finding where the relevant knowledge is, understanding its structure(s), and querying it for its values.

On the World Wide Web, ontologies are now more commonly used as the model structures to turn information into (re)usable knowledge. RDF[2] is being widely used as the representation language for ontologies, and are also the

[2] http://www.w3.org/TR/REC-rdf-syntax/

foundation for the next generation of the Web, i.e. the "Web of Data". This Web of Data is a Web of actionable information, i.e. information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of meaning in which the logical connection of terms establishes interoperability between systems[2].

Increasingly, private corporations and governments are realising the potential of encoding knowledge as RDF. Facebook[3] recently announced their Open Graph Protocol[4] which allows a Facebook user to integrate other non-Facebook web pages into the user's social graph. Open Graph Protocol uses machine-processable semi-structured data to mark up web pages. Various governments, in order to improve the delivery of services to their citizens, are opening up their data and publishing these data in semi-structured format, many of them in RDF, to improve the delivery of goods and services. The United States government has set up data.gov[5] to release public data. The UK Government, keen to unlock the benefits of economic and social gain of public sector information (PSI) reuse, has set up data.gov.uk[6]. All these efforts will enable their citizens to ask questions, such as: "Where can I find a good school, hospital, investment advisor, or a good employer?" They will also lead to improvements of the delivery of services to the public.

Adequate steps need to be taken to learn how to publish and consume (public) semi-structured data.

2 A Case Study of Publishing and Consuming Public Data

Omitola et. al. [4] carried out a case study showing how United Kingdom geographical data, from the UK's Ordnance Survey, can be used as a set of "join points" to mesh public data for crime, mortality rates, and hospital waiting times. Meshing was defined as the ability to naturally merge together a dynamic set of information sources, and a join-point as a point of reference shared by all datasets. They investigated the use of disparate sets of data in an effort to better understand the challenges of their integration using Semantic Web approaches. Part of this investigation involved ascertaining the datasets that were available, their formats, and converting them into (re)usable formats. They faced a number of challenges which included:

1. Sourcing the datasets: Since many of the datasets of interest were in comma-separated-value (CSV) format, a process of converting the datasets to RDF had to be undertaken,
2. Modelling the datasets: The sources of the datasets contained little or no explicit semantic description of the data, [4] had to provide schema definitions for these data,

[3] http://www.facebook.com
[4] http://developers.facebook.com/docs/opengraph
[5] http://www.data.gov
[6] http://data.gov.uk

3. Aligning of datasets: The datasets were harvested from disparate public bodies, the problem of unintentional coreference, i.e. different names referring to the same thing, had to be solved,
4. The challenge of re-use and consumption of the data also needed to be solved.

A number of solutions used to solve some of the challenges were introduced and described in the paper. These included, inter alia,

1. the use of the SCOVO ontology [5] to model the multi-dimensional aspects of the datasets, such as time, entity types, etc,
2. the design decisions and the efforts needed to convert the (mostly) CSV-encoded datasets into RDF,
3. the selection of the appropriate "join point" used to integrate the datasets,
4. the usage of a co-reference service to resolve co-references, and
5. the use of Exhibit[7] to develop the client application used to consume the data.

3 Ontologies, Linked Data, and Linked Data Cloud

The continued growth in the adoption and usage of ontologies and semi-structured data in government and industries is bringing about the growth in datasets published in linked data format, and a growing interest in connecting these datasets together. Linked Data is a style of publishing data on the Web that emphasises data reuse and connections between related data sources. This growth and interest can be seen in the Linked Data community[8] which aims at making data freely available to everyone and to extend the Web with a data commons by publishing various open data sets as RDF and by setting RDF links between data items from different data sources. Figure 1 shows a linked data cloud of the data sets that have been published and interlinked by the community so far. Collectively, the data sets consist of over 13.1 billion RDF triples, which are interlinked by around 142 million RDF links (as of July 2009).

With these datasets in different knowledge bases and data stores, there is a paradigm shift occurring. This shift is an important one. We are moving away from the paradigm of "given a set of data, what technique(s) can I use on this dataset and gain insights" to the paradigm of "given a problem, what is the best dataset I can get to solve the problem or answer the questions". The kinds of questions we may want to ask are:

1. We could use data from a geographical database such as Geonames[9], combine this with a social network to determine how much people's locations and the distances between them affect their chance of being friends. If we combine this with census data, we could ask questions to determine if population size or population demographics affect this chance of making new friends (to ascertain if people from small towns are more open to new friendships, etc).

[7] http://simile.mit.edu/wiki/Exhibit/API
[8] http://linkeddata.org/
[9] http://www.geonames.org/

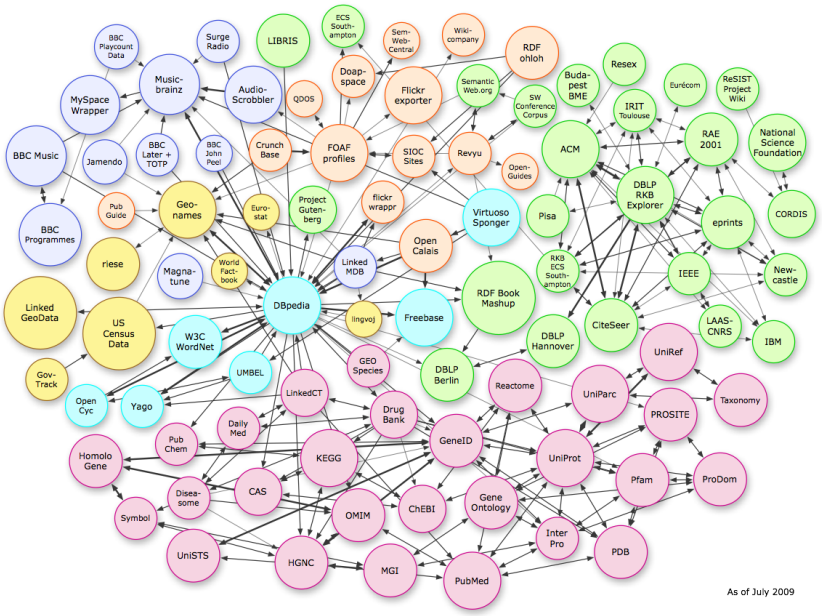

Fig. 1. The Linked Data Cloud

2. We could use CO_2 emmisions data of all the councils of the United Kingdom and check it with the different income levels of each council and with data of the average educational attainments of each council to determine if there is a correlation between income and awareness of environmental issues.

The outputs from answering questions such as the ones above may be used as inputs to an inductive reasoner or "human" expert for required validations and/or necessary recommendations.

In order to answer the aforementioned queries, we need to make it easy to find the relevant datasets and datastores in the first instance. We need to build a platform that makes it easy for data to find data.

4 Interlinking Datasets and Co-referencing Service

For data to find data, we need a dataset discovery mechanism and, after discovering relevant datasets, the selection of the best-suited ones. To make this possible, the Linked Data community has come up with **voiD**[10][6] , a "Vocabulary of Interlinked Datasets".

[10] http://void.rkbexplorer.com/

4.1 voiD

voiD is an RDF based schema used to describe the content and the interlinking between datasets. There are two core classes at the heart of voiD:

1. A dataset (*void:Dataset*), i.e. a collection of data, which is:
   - published and maintained by a single provider,
   - available as RDF,
   - accessible, for example, through dereferenceable HTTP URIs or a SPARQL[11] endpoint
2. The interlinking modelled by a linkset (*void:Linkset*). A linkset in voiD is a subclass of a dataset, used for describing the interlinking relationship between datasets. In each interlinking triple, the subject is a resource hosted in one dataset and the object is a resource hosted in another dataset. This modelling enables a flexible and powerful way to state the interlinking between two datasets, such as how many links there exist, the kind of links, and who made these statements.

Figure 2 depicts[12] the modelling of the interlinking in voiD.

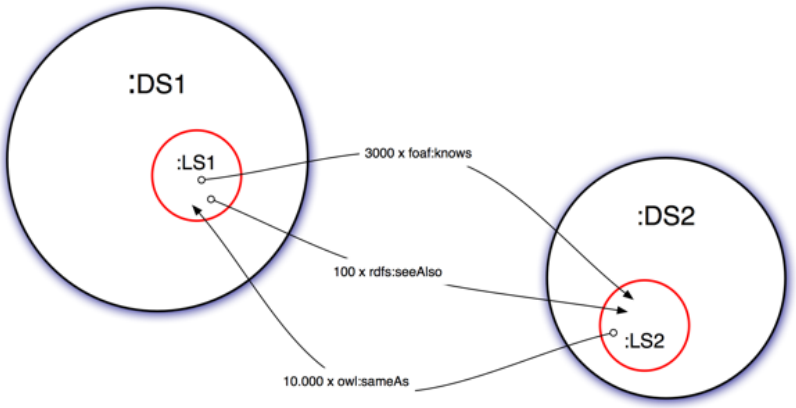

Fig. 2. Model of voiD Interlinking

4.2 Co-referencing

The acquisition of knowledge from heterogeneous sources carries with it the problem of unintentional coreference, i.e., different sources may refer to the same

[11] http://www.w3.org/TR/rdf-sparql-query/
[12] from http://semanticweb.org/wiki/VoiD

entities by different means. If we want to use this knowledge effectively, we need to be able to determine which entities are co-referent, and to collapse these multiple instances into a standard representation. **sameAs**[13][3] has been developed as a web service for the collection and publication of information about co-referent entities. For example, if one wants to look at co-referent entities for the UK's "county of cambridgeshire", this can be called via the following URI: *http://sameas.org/html?q=county+of+cambridgeshire*. Doing this, one can see that the "county of cambrigeshire" is defined in Geonames as *http://www.geonames.org/2653940/* and in DBpedia[14] as *http://dbpedia.org/page/Cambridgeshire*.

Since we now have a mechanism to ascertain (un)intentional coreference between entities (the sameAs service), and to discover datasets (voiD), we need to develop a platform that can be used to query and fetch the relevant data amongst these federated datasets.

5 Federated Fetching and Querying

5.1 Federated Fetch and Contextualised Directories

Directories are locator services, they return references after being provided with query terms. Provision of references turns the next series of activities to "fetch", and on the WWW, this is equivalent to "federated fetch". For example, terms searches on the Google[15] search engine, after going to a special directory created by Google, return references to the real documents. This type of directory is an example of **contextualised directories**. Contextualised directories enable data to find data. An extension of contextualised directories are semantically reconciled directories. Semantically reconciled directories exploit synonyms (or coreferences) to return pointers to different words that mean the same thing. An example of this is the *sameAs* web service mentioned above. Further on from these are "semantically reconciled and relationship-aware directories" that provide higher degree of context by allowing users to discover additional data. One can query for any single set of terms and locate a bundle and, by using the relationship-awareness property of the directory, learn how that bundle relates to other bundles. A combination of *voiD* and *sameAs* gives us semantically reconciled and relationship-aware directory which we can use to find relevant data and datasets.

5.2 Geography of dataset relationships

Space is one of the principal media through which structure and form are expressed, and spatial organisation produce complex geometries of relationships and structure. A key endeavour is to determine these geometries in order to

[13] http://sameas.org/
[14] http://dbpedia.org/
[15] www.google.com

aid understanding and navigation. To determine these spatial geometries, one solution that is being explored is to leverage the power of voiD and sameAs to perform topology analysis of the datasets in the Linked Data cloud. The kinds of analysis that can be done include:

1. Degree centrality: This measures the extent to which a node or a dataset connects to all other nodes or datasets in the linked data cloud,
2. Small world: This measures the average minimum path between the nodes in different datasets,
3. Clustering co-efficient/factor: This measures the probability of two datasets being neighbours of one another.

This topology relationship will act as a meta-network layer (maintaining state of networks of relationships) which can also be continually updated to cope with the changing dynamics of the datasets.

We envisage to leverage the power given us by voiD, sameAs, and the Meta-Network layer to perform more optimised query plans and as path indices to speed up query processing. This will lead to the ability to efficiently discover datasets and select the most suitable ones.

6 Conclusion

In this paper, we described efforts being made in the (Semantic) Web community to turn today's information glut into sense-making. The central element of these efforts is the representation of information in semi-structured format, (e.g. using RDF), called ontologies. We described a case study of how this was achieved using government (public) data, the attendant challenges, and the devised solutions. Such data are now being aggregated into disparate datasets and linked together forming a linked data cloud. We described services that are in use to perform the linkage (voiD) and to ascertain similarity between data concepts (sameAs). We outlined current work that is ongoing which will use voiD and sameAs to build a federated fetching and querying platform to efficiently discover relevant datasets for (Semantic) web applications to use.

7 Acknowledgements

This work was supported by the EnAKTing project, funded by EPSRC project number EP/G008493/1. We thank Manuel Salvadores and Gary Wills for their contributions.

References

1. Nigel Shadbolt and Kieron O'Hara: *An Overview of the Aims, Ambitions and Assumptions of the Advanced Knowledge Technologies Interdisciplinary Research Collaboration*. Advanced Knowledge Technologies Selected Papers, 2003

2. Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee: *The Semantic Web Revisited*. Advanced Knowledge Technologies Selected Papers, 2006-2007
3. Hugh Glaser, Afraz Jaffri, and Ian C. Millard: *Managing Co-reference on the Semantic Web*. LDOW 2009, April 2009.
4. Tope Omitola, Christos L. Koumenides, Igor O. Popov, Yang Yang, Manuel Salvadores, Martin Szomszor, Tim Berners-Lee, Nicholas Gibbins, Wendy Hall, mc schraefel, and Nigel Shadbolt: *Put in your postcode, out comes the data: A Case Study*. 7th Extended Semantic Web Conference, May 2010, Greece.
5. M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers: *SCOVO: Using Statistics on the Web of Data*. In European Semantic Web Conference 2009 (ESWC 2009).
6. Richard Cyganiak, Holger Stenzhorn, Renaud Delbru, Stefan Decker, and Giovanni Tummarello: *Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web*. in The Semantic Web: Research and Applications, pub. Lecture Notes in Computer Science, Volume 5021, 2008.