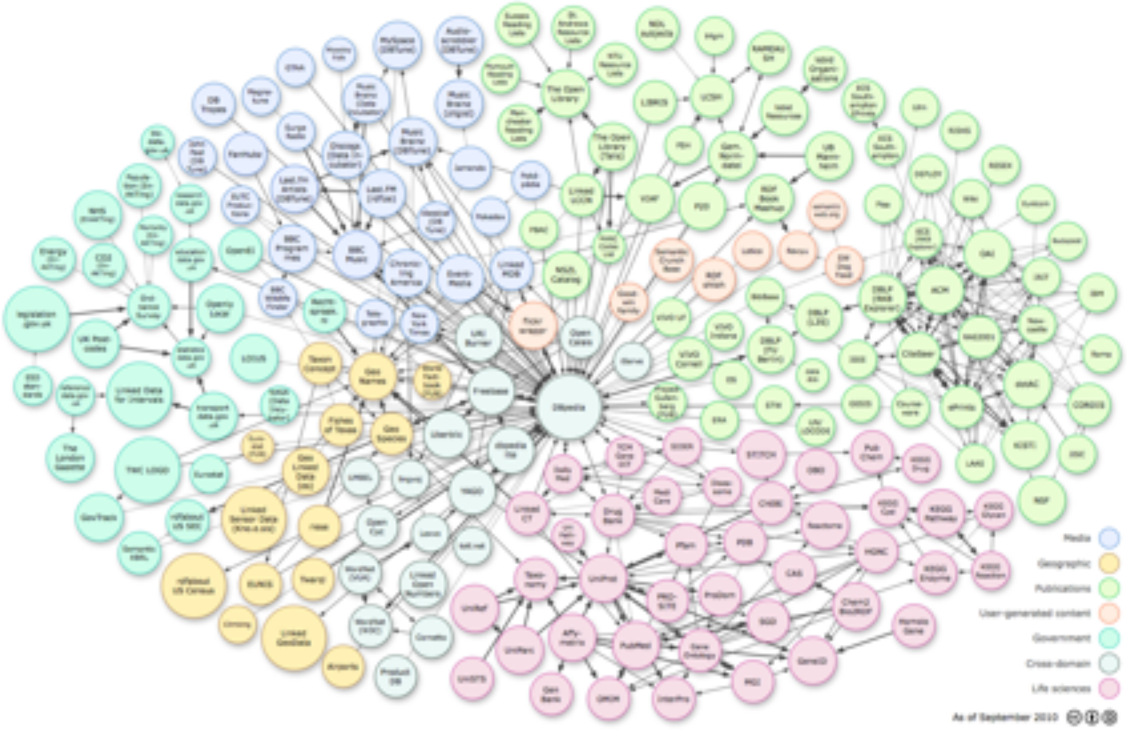
“Ready for Change?” Chapter Framework

Christopher Brewster: From Silos to Supply Chains: Linked Data and the 21st century Enterprise

1. *Picture / Visuals*



2. *Introduction*

We live in a digital age where the effective communication of knowledge is key to the survival and viability of companies and organisations around the world. As companies grow so do the challenges of making knowledge available to all actors. The need for more knowledge has fuelled and been driven by the accessibility of knowledge and data over the Internet. The immense quantity has given an illusion of abundance when the reality is that most organisations of any size suffer from significant limitations on their activity due to the difficulty of obtaining the right knowledge, to the right people at the right time. In this chapter, we consider the problem both internally in organisations, externally along the supply chain and we point to solutions which are being adopted by organisations around the world centring on ‘Semantic Technologies’. We present some of the highlights of semantic technologies and conclude with a future vision which might have alleviated some of the problems associated with the June 2011 *E. Coli* outbreak in Germany and its impact on Spanish cucumber suppliers.

3. ***Knowledge Silos***

All organisations develop knowledge silos. As organisations grow, so they create different departments, different sections and naturally people specialise or are hired with expertise in a particular area. Little by little communication between the different departments becomes more difficult. As organisations move from being startups with 5-10 people often working together in one space to becoming fully fledged companies with separate departments (development or design, manufacturing, sales, HR, finance etc.) different bodies of knowledge develop within each department. Finance may know all about the financial state of an organisation but not know or understand the new products being developed. The IT department does not understand the needs of the research department. These are typical everyday occurrences in all organisations above a certain size, exacerbated and made obvious when we consider that most organisations run very large numbers of separate databases[1]. Communication between departments becomes fossilised in certain specific channels fitting in with defined requirements (which probably made sense originally but are often too constrained today). So typically a particular form is developed to request a specific piece of information or requisition an item. If the information or item falls outside the scope originally foreseen the systems breakdown. In an organisation with an effective management culture, a telephone call is made and everything is sorted out. In an organisation with sclerotic or fossilised structures getting round such communication bottlenecks can be very time consuming and expensive.

Examples of such knowledge silos exist all around us. A well known aeronautics company was organised around three traditional departments - design, manufacturing and sales. While design designed, manufacturing would redesign and then manufacturer, and then finally sales would sell but sometimes also redesign and manufacture as well. No department would receive appropriate feedback from another as to the suitability of the process they were responsible for.

In the life sciences, one of the major challenges is the large number of laboratories a pharmaceutical company has around the world, and the many different concurrent areas of research being undertaken. Historically, there would occur frequent duplication of effort, or else research which could have given insights, was missed because of lack of communication between research teams[2]. This lack of communication has continued

[1] Microsoft SQL Server clients have on average over 1700 databases per company (http://www.scribd.com/doc/47442610/oracle-vs-sql-server)

[2] The film *Extraordinary Measures* about the search for a drug that would cure the fatal infant disease Pompe describes this clearly.

to be true until very recently because in spite of an impressive collection of public and shared databases, both of research literature (Pubmed) and of genetic and molecular results (SwissProt, Uniprot, etc.), there was no way to undertake federated queries across data sets.

A very similar phenomenon occurs with the problem of finding who the experts are in an organisation. Experts live in their specific sub-department and few people know they are experts in a specific area of knowledge beyond their immediate collaborators. It is thus often very hard to find out who knows what in an organisation of any significant size. As organisations grow, the problem becomes ever greater. As a great deal of knowledge is undefined, procedural or otherwise tacit, this magnifies the challenge.

There are two aspects to this situation. One concerns the mere absence of communication or opportunities for communication. For example, if one is having trouble with the marketing department promoting one's new product (idea, project, event, etc.), it does not follow that one would ask the colleague next door how to solve the problem -- as one is typically unaware that they have the necessary skill, experience, contacts etc. How could one imagine that the laboratory in Copenhagen is also working on molecule XYZ but for its use in cardiovascular disease rather than my interest in diabetes? Most organisations try to address this concern with 'away days', or 'off site' meetings or other forms of informal gatherings. There is plenty of evidence that informal gatherings lead to serendipitous discovery of solutions to long standing problems. However, this approach is very hit and miss.

The second challenge concerns commonality of language. In large complex manufacturing companies, this is very evident in the different labels, tags or names applied for the same object depending on whether it is engineering, sales or finance which is concerned. The same is true in many other areas. Breast cancer specialists describe the same mammogram with different terminology in different hospitals. The reverse is also true; life science researchers refer by the name *gene* to different entities including *DNA sequence, protein, RNA sequence, an allele*. This means in practice that even if communication channels exist, knowledge is not shared because it comes under a different heading, a different category, a different label.

So far, we have considered the problems mostly from the perspective of communication and knowledge sharing within a given organisation. The problems described are compounded when we consider complex supply chains.

4. ***Supply Chains***

Modern industrial society is highly dependent on complex supply chains which cross national boundaries, multiple human languages and a myriad different industrial processes and human cultures. Supply chains are prime examples of structures which suffer from knowledge loss. In principle it is an extraordinary achievement that supply chains work at all given the potential for ‘chinese whispers’ i.e. for the loss of information as it is transferred across or along the chain. There are many inherent difficulties in sharing knowledge along the supply chain. There are many different actors, who all have different priorities, different business models. These actors often come from very different cultures and even when the language is common, breakdowns can occur and have catastrophic results. Simple example: the Mars Orbiter broke up while orbiting Mars in 1999 due to outsourced software having been written using metric measuring units rather than imperial units that NASA used at the time.

The issue just as within a single organisation is not only that different actors have different objectives and languages but also that at a purely technical level they are using different databases, with different schemas, different ways of representing their corner of the universe. And very often relevant data is not preserved in a manner that allows subsequent processing and analysis e.g. by recording data in unstructured formats such as PDF or MS Word. For example, many fault reports in manufacturing companies are written in MS Words, and then saved as PDF documents in a database. A large proportion of universities provide descriptions of their taught courses as pdf documents. In both cases, relatively structured data is made inaccessible due to technical choices.

If we look back historically to long supply chains like the Silk Road bringing silk and spices to Europe in the Middle Ages, effective supply chains have always required and provided for a transfer of knowledge in both directions (upstream and downstream). But this can be slow. It has sometimes taken decades or centuries for technologies and know-how to spread. Thus the Romans were convinced that silk came from trees.

We now live in a digital age where effective communication of knowledge is essential and possible. We cannot, indeed we should not, wait upon mechanisms for the exchange of information that depend on word of mouth or serendipitous encounters.

In traditional supply chains, each actor only has to talk to their immediate supplier and customer. Hence the metaphor of a chain with clear links with a trajectory that flows from raw material to end user. This is no longer the case. Not only are chains far more complex, with multiple interconnected players, but also there are many pressures to increase the flow of data and knowledge along the whole supply chain. There are:

- regulatory pressures for quality or health reasons (Can you as a retailer guarantee the toy manufacturer has not used lead paint?),
- consumer pressures for ethical or environmental reasons (Is your football manufactured using child labour?),
- purely commercial pressures where actors need greater information about the origins or final destinations of products (Can you guarantee that that Hermes bag is not reaching the local flea market?).

Furthermore actors across the supply chain need more data and knowledge for their own commercial purposes. Upstream actors (e.g. food producers) need to understand more effectively the behaviour, needs and choices of downstream actors (e.g. food consumers). Until now all knowledge has flowed downstream i.e. people have wanted to know who produced a specific product, but few producers have really known who consumed their product. Not much knowledge flows upstream. Yet increasingly pressure is rising to integrate data and knowledge given the additional need for rapid responses and greater flexibility in the supply chain. In the food supply chain, for example, there are constant unpredictable disruptive events (crop failures, floods etc.) and sudden changes in product demands, which need much more agile responsiveness on all actors involved in the supply web. With change spreading ever more rapidly and the increased transparency of our global communication system, availability of correct and appropriate information is more essential than ever for effective operations but also for commercial survival.

Integrating data from different sources in different formats is a substantial challenge and this is where Semantic Technologies are playing an increasingly important role.

5. *Semantic Technologies and Data Integration*

The original World Wide Web, the one most of us still use today, whether it is the public incarnation or on a company internal intranet, consists of a vast collection of documents linked together - some of which may be generated automatically from underlying databases. We navigate this structure either by using search engines like Google or by following links from one document to another. This web is designed for human consumption. Machines to a large extent cannot process all that vast collection of information out there apart from providing an index. Complex questions cannot be answered except through substantial human effort. Even though many web pages are generated from information in databases, the underlying data is largely inaccessible for reuse by a wider audience.

In our document centric World Wide Web world, we are able to ask questions such as the following:

- Show me all documents with the words ‘product recall’ or ’10 megapixel digital camera’
- Show me all documents with the words ‘E. Coli’ and ‘cucumber’ in them.

We cannot, however, ask questions like:

- Show me all published articles containing the words ‘business process’ written by faculty in this institution who teach operations management to MBA students.
- Show me reports written in the last six months by members of the new widget development team.

Similarly with databases, one can request a report showing:

- A list of all sales made between 1 May and 30 June.
- All employees employed by the admin department
- All items of expenses charged to Tom Brown

Sadly we cannot ask for reports that:

- List all dentists within 10km who are members of the Royal College of Surgeons and have a free slot next Friday.

- List all parts that have had failures used in engine XYZ designed by Tom Brown and sold to client ABC.
- Identify all bands from the city of Manchester whose music type is rock and whose date of formation is before 2003.

To summarise:

- There are too many documents (creating infosmog)
- Documents are designed for humans to read, not for machines to process
- Where data exists, it lies across many databases *within* the organisation
- Data also often lies across many databases *across* many organisations

To address these limitations, the concept of the Semantic Web was proposed by Tim Berners-Lee, Jim Hendler and Ora Lassila, at the end of the 1990s. The fundamental idea is to create a 'web of data', of machine readable data in parallel to the 'web of documents' which already exists. Berners-Lee's vision is a world where vast quantities of data are published both by the public and internally in organisations, where the format follows certain standard rules, and uses a variety of standard vocabularies or 'ontologies', and where different, disparate data can be interlinked to answer questions and to connect both people and knowledge in a manner that was impossible before.

Semantic Web technology consists of a set of technologies each of which are relatively simple but which in combination prove to be extremely powerful. There are several layers which build on each other and of these the most central are[3]:

1. Objects in the world, whether physical or abstract, have unique identifiers similar to the "name" of a web page i.e. the long address beginning "http:// ...". These identifiers (technically called "Uniform Resource Identifier") can be created by anyone so they do not depend upon a centralised database. This is an adaptation of the traditional web pages addresses (URLs) and is equivalent to proper names for things and people but slightly more systematic and intended to be machine readable.
2. Statements about the world are made using 'triples' such as *London - IsCapitalOf - UnitedKingdom* where each element in the triple would be identified by a unique identifier (URI). This is equivalent to the everyday idea of simple sentences or statements such as "London is the capital of the United Kingdom". Triples have "subjects", "objects" and "predicates".
3. The triples are expressed in a formal model (the "Resource Description Framework" or RDF) in a machine readable syntax (of which there are a number, the most commonly used being RDF XML). RDF provides mechanisms to says some things are of certain types. Closely related is the RDF Schema which provides the formal rules so as to define classes and express the concept of "subjects", "objects" and "predicates" which make up a triple. Together RDF and RDFS allow one to express simple taxonomies about the world.

[3] There are other layers of the technology about which further details can be found in the suggested readings at the end of this chapter.

4. Ontologies or formal vocabularies allow more complex statements about the world. One can define formally classes and properties of classes. Ontologies are usually expressed in RDFS or the more formally rich “Web Ontology Language” (OWL). Ontologies allows logical conclusions to be drawn (inference) ranging from the relatively simple (if A isMarriedTo B, then B isMarriedTo A) to the more complex (if A isMemberOf ProjectB, and C isMemberOf ProjectB, then A knows C). There are a great many ontologies which have been established and little by little a considerable number are becoming recognised as *de facto* ‘standards’. One of the most widely used such ontologies, and one of the most commercially significant, is the “Good Relations” ontology developed by Martin Hepp for the needs of e-commerce.
5. SPARQL is the query language which enables queries to be posed to collections of triples. This is key in facilitating the integration of data from multiple sources.

This stack of technologies together with a number of others allow individuals and organisations to publish either internally or publicly sets of data. These data sets can be accessed across the internet (just as one would with a web page), and then integrated with other data whether in house or from yet other external sources. All this was technically possible quite rapidly after Tim Berners-Lee’s original proposals, but for a certain period of time there was only limited data available with relatively restricted access. This all changed with an explosion of “linked data” since 2007.

6. *Open Knowledge: Linked Data*

Semantic technologies cannot not have an impact on business and society, on the world in general, unless there is sufficient uptake and adoption. A sufficient number of people need to be using the technologies and above all publishing data so that the technology is useful. To further this end Berners-Lee proposed a set of principles to facilitate the publishing of data on the web so that a global data space would be created, the so called “Linked Data principles”. Linked Data is data published on the web that is “is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets[4]”. Ordinary links in web pages allow different websites to be connected. Links in Linked Data allow formal statements to be made that link arbitrary arbitrary things in the world.

Following the declaration of the principles, there has been an explosion of data available on the web. The Linking Open Data project brought researchers, universities and small companies together, and these were quickly followed by large organisations such as the BBC, Thompson Reuters and the Library of Congress. One of the key developments was the transformation of the publicly edited Wikipedia into a structured set of data called Dbpedia (http://dbpedia.org/). This has acted as a central hub to which many other data sets link, a dizzying array of different types of data mostly freely available for organisations and people to use. Data has been added concerning geographical locations, scientific publications, music, programmes on television and

[4] Bizer et al. 2009

radio, all kinds of life science data including proteins, genes, metabolic pathways, drugs and clinical trials, political and historical data and statistical and census data.

This complex interlinked ‘cloud of data’ provides an immensely rich resource for all types of organisations to integrate with their own data, as well as providing examples for others to follow. Although the data on the ‘linked data cloud’ is mostly open for use by anyone, there are plenty or organisations which offer their data for non-commercial use for free, and charge for commercial usage (e.g. musicbrainz). The open knowledge provided in the ‘cloud of data’ has had substantial impacts in persuading organisation, especially governments, to make more data publicly available in machine readable formats. The most important initiatives in this regard have been efforts by the UK and US governments to make large quantities of government data available to the wider public for both social and business purposes. This has resulted in the ‘data.gov’ and ‘data.gov.uk’ websites where data has been made available not just in standard database formats (cvs) but also as linked data formats (in RDF) enabling its use and integration in all kinds of projects both commercial and non-commercial.

While the development of the linked data cloud is an extraordinary achievement, in reality it is only one step, one part of the wider application of semantic technologies. We now turn to some examples of their application in practice.

7. *Examples in practice*

The obvious question upon the arrival of new technologies, new ways of doing things is whether there has been any uptake of these innovations. Like all innovations which demand a different way of thinking about the world, uptake has initially been slow but has seen a tremendous growth in the period 2007-2011. The first users of any scale were pharmaceutical companies who had the necessary resources but also recognised economic imperative of increasing the flow of knowledge both internally in their organisations and in interacting externally with the very large number of publicly developed databases.

- Domain: Health/Pharmaceutical
  - Ely Lily uses SW technology for prioritising drug discovery targets.
  - PharmaSURVEYOR uses SW technology to compose safer drug regimes for patients, to limit drug side effects and interactions. This uses ontologies to specify medical conditions and integrates with data from multiple databases.
- Domain: Space Exploration
  - NASA uses SW technology for expertise location. Information about employees’ work history, affiliation, skills and teams they have worked in is collected by NASA in multiple databases. Identifying the right individual/skill/work history combination has been revolutionised by the POPS system (developed by Clark and Parsia LLC) .
- Domain: Media and Web publishing
  - The BBC uses SW and LD technologies to link all its programmes, presenters and artists so that web site users can easily find what other programmes person ‘David

Attenborough’ participates in. Furthermore, all music data is imported directly from the crowd-sourced Musicbrainz website as RDF. Similarly, BBC Nature has URIs for every species it is interested in and aggregates data from a number of external sources (Wikipedia, WWF’s Wildfinder, IUCN, etc.). The most recent use of this technology is the sports domain, applying what they call “dynamic semantic publishing” which uses “linked data technology to automate the aggregation, publishing and re-purposing of interrelated content objects according to an ontological domain-modelled information architecture”[5]. This has been used for the World Cup in 2010 and will be used for the London Olympics 2012.

- Domain: Government services
  - As mentioned previously, the UK government has released a huge amount of data in standard formats on its data.gov.uk website. This was followed by the release of a large amount of geographical data from the Ordinance Survey, for free and in RDF format. These initiatives have spawned a mini industry of applications using the data for creative and commercial ends. For example applications based on this data include one that compares all cars and their fuel consumption costs http://uk-car-fuel-emissions.findthebest.com/) and another that enables school choices to be based on public data (http://schoolscout.co.uk/pg/customsearch/search)
  - The Amsterdam Fire Department uses SW and LD technologies to manage fire fighting incidents, by pulling information from a a number of different data sources in partner organisations.
- Domain: Web security
  - Garlik.com uses SW technologies to integrate vast quantities of data about people and the data available about them on the web so as to provide services to protect people’s privacy and online security. Semantic technologies enable the integration and reasoning over a very large body of data.
- Domain: e-Commerce
  - The Good Relations Ontology developed by Martin Hepp has enabled a paradigm shift in the way e-commerce websites present their products. Websites using the Good Relations Ontology to mark up their products allow Google, Yahoo and other search engines to display their products in a much more effective manner. Bestbuy.com had a 30% increase in website traffic and a corresponding increase in sales.

The importance of these examples (and there are many others) is exemplified by the fact that Google, Bing and Yahoo recently collaborated to establish a new semantic standard (available at schema.org) for website markup so as to further facilitate the integration of information across the web at a *data* level rather than merely at level of web pages read by human beings as has been the case up till now.

8. ***The Case of the Cucumber***

[5] Rayfield, J.P. Dynamic Semantic Publishing, Sport and the Olympics at the BBC, SemTech London 2011, and http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_news_websites_content_mana.html

In late May, early June 2011, there occurred a major *E. Coli* outbreak in the Hamburg area of Germany. Nearly 40 people died and over 3000 people were infected some with severe symptoms. The outbreak was initially blamed on organic cucumbers, tomatoes and lettuce imported from Spain. These were blamed for carrying the *E. Coli* virus. The consequences were disastrous for Spanish farmers who lost tens of millions of Euros in crops that no one wanted to buy. The outbreak was then tracked down to contaminated bean sprouts produced locally in a farm in Germany. But that too proved incorrect. At the time of writing, the ultimate source of the virus has not been conclusively identified. The management of epidemic outbreaks like these depends on the rapid collection of data from a variety of sources - patients/victims, retailers, transporters, wholesalers and food producers to name only the obvious. In effect everyone on the food supply chain is implicated as a possible source of the problem.

We can envisage an alternative scenario to that which played out. Let us imagine that all actors along the supply chain were part of the 'web of data', of the Semantic Web. After all the data was there in one form or another. It is only the fact that the data silos, the databases which cannot talk to each other prevented a much more sophisticated and rapid handling of the outbreak. We will ignore issues of privacy and data ownership for the present in the following scenario. The complex queries that needed to be posed included:

1. Which cucumbers did persons P_0 ....P_n buy, from which retailers, and where did the retailers obtain them from, using what logistics chain, and from which farm did they originate .....?
2. and what were the growing conditions?
3. and who do P_0 ... P_n have personal contact with?

This data could have been accessed, perhaps we could even postulate that it would have been accessed had the appropriate political will existed. What we do know is that:

- The Person P_0 ... P_n were the victims. We know this from the hospital records.
- The victims purchases could have, at least to some extent, been identified from their banking activity i.e. one could have linked bank cards to purchase receipts to retailers and time and location of purchase.
- Cucumber source (retailer purchasing/ordering datasets) could have been identified.
- Journey of cucumber (truck and logistics data + sensors) could have been established.
- The original farms (based on truck collection, purchase/sales receipts) could have been identified.
- Circle of contacts of P_0 ... P_n (social media, Facebook, Twitter, phone records) could have been mapped.

The technologies already exist to facilitate much of this chain of (potentially) linked data. Vocabularies/Ontologies exist to represent individuals (Foaf), health records and to integrate these with clinical results, disease and treatments (e.g. Translational Medicine Ontology), financial transactions (XBRL, Finance Ontology), geographical and logistics data (Geonames Ontology, Logistics Ontology - several others), social networks (Foaf, SIOC Ontology), and more. All of these and many other relevant formal vocabularies (ontologies) exist which would allow data from different actors to be linked together, to enable data to be integrated and for federated queries to be posed so necessary if we are to meaningfully share knowledge.

The challenge is to actually deploy these technologies in a manner that is beneficial to wider society and business. Like all technologies there will be early adopters who reap early benefits and late adopters who will benefit far less from the power of these tools. Above all, we must recognise that there is inevitable resistance when new technologies arise which have the potential to radically transform the way people collaborate and do business[6].

9. *Ready for Change: Semantic Technologies*

The question arises as to whether a given organisation should adopt these technologies. The more appropriate question is whether an company or organisation can afford not to. Given the ever growing complexity of our society and the mechanisms by which we collaborate, share knowledge and the implications of failure not to do so, **NOW** is not early enough to retool our knowledge sharing technologies.

If your organisation has more than a dozen employees, if you depend on knowledge and data as a key competitive advantage in the operation of your business, if you have trouble finding the expert in your company for a particular question or task, if you need to track the performance of products across divisions, if .... you live in the digital age, then semantic technologies are a must for the survival and future viability of your company let alone its growth.

10.*Further Reading*

On Silos

- The Silo Effect and Other Productivity Killers http://www.onpreinit.com/2009/10/business-silo-effect-it-software.html
- Bundred, S 2006, 'Solutions to Silos: Joining Up Knowledge', Public Money & Management, 26, 2, pp. 125-130

On Semantic Technologies and Linked Data

- Berners-Lee, T.; Hendler, J. & Lassila, O., The Semantic Web, Scientific American, 2001, 30-37
- Bizer, C.; Heath, T. & Berners-Lee, T., Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems, 2009
- An extensive list of case studies can be found here: http://www.w3.org/2001/sw/sweo/public/UseCases/ and here: http://logd.tw.rpi.edu/

[6] This fictional *E. Coli* scenario is very similar to actually implemented work on influenza cf. Semantic Web Methodologies Provide Access to FLU Data Without Getting You Sick of Searching, Timothy Lebo and Joanne Luciano, AMIA 2011

- Gardner, S. P. Ontologies and semantic data integration, Drug Discovery Today, 2005, 10, 1001 - 1007

On the E. Coli outbreak

- Der Spiegel (in English) http://www.spiegel.de/international/europe/0,1518,768534,00.html
-