

Diabologic: DBpedia

by Frank Dolinar

It started out, in 1989, as an informal attempt to provide document sharing for the scientists at CERN (<http://public.web.cern.ch/public/>), the European Organization for Nuclear Research. Twenty years ago, documents existed on computers with different operating systems and proprietary file formats. It took a lot of tinkering to share properly formatted documents among researchers. Short of that, the solutions were to send the documents in plain text – which lost the formatting of the equations or the diagrams and thereby was more than a little self-defeating – or to print out the document and send the paper to a colleague.

Neither approach provided the researchers with the ability to: 1) share both the concepts and data in the original researcher's format; and 2) deliver documents in a timely fashion. Something had to change.

In March 1989, Tim Berners-Lee wrote a document titled "Information Management: a Proposal", and got permission from his management to give his proposal a try. It took a year to design and implement the necessary protocols and to write the first web browser. Thus, in 1990, the World Wide Web was born.

The web has become *the* pervasive and disruptive technology of the last two decades. Little did we know how dramatically it would change our lives. It has revolutionized communications and computing; changed the environment of our personal, business, government, and entertainment interactions; and has reshaped the economy.

But that's hardly the end of the story. Last month (February 2009), Tim Berners-Lee, who is credited as the inventor of the World Wide Web, gave a presentation at this year's TED Conference (http://www.ted.com/talks/tim_berniers_lee_on_the_next_web.html). He suggests that the future web will (indeed must) make more data accessible to all of us – lots more data.

If you find yourself awash in more data than you know what to do with, you might wonder why you would want more? The simple answer is relationships. To make connections among bits of data that will actually help find answers to complex questions, you need more data than is typically available to an individual, or a business, or perhaps even a government. Yet once this data is available, it becomes possible to make such connections and find your way from the question to potentially useful data.

It may not lead immediately to an answer, but with enough data and the right connections, you can more quickly determine where to start your search and how to proceed. With enough data you can begin to see the patterns. Seeing patterns often can simplify the process.

What data should get shared? All public databases, including government, industry, publications, research results, etc. (I would be significantly concerned about sharing personal financial or medical data, but I get the impression from Berners-Lee's presentation that this is not his intent.)

We already have some success and sophistication of web-based sharing. Without it, the links I've provided in this document wouldn't be useful, or even possible.

However, Berners-Lee calls our attention to a new priority that he calls "Raw Data Now". In a February 11, 2009 article for Compute Magazine, John Koenig says of Berners-Lee,

"Speaking at the TED Conference, he made his point with a simple premise, "Data, you cannot naturally use by itself." In other words, for it to be useful, we should combine and share the data we have collected with other data. ... Not only should we share our own data, but we should demand that governments and businesses share the data they prepare as well, he says. This is the new objective for the World Wide Web. As he points out, "data drives a huge amount of what happens in our lives... because somebody takes the data and does something with it." To Berners-Lee, it is from this sharing of data that advances in science will come."

To make such sharing both effective and valuable, Berners-Lee provides three points of instruction:

- 1) a URL (i.e. a browser link) should point to the data;
- 2) anyone accessing the URL should get data back; and
- 3) relationships in the data should point to additional URLs with data.

These rules are a much simpler departure from the past 10 years of discussion of the so-called "semantic web", which has received little acceptance because of its level of abstraction. It's a long way from the current state of the web to the time when we will have software access to a web of data.

If you want to see what's possible using this approach when you have enough data, view Hans Rosling's talk (http://www.ted.com/index.php/talks/hans_rosling_reveals_new_insights_on_poverty.html).

Then look at the DBpedia website (<http://dbpedia.org/About>), where they've taken Berners-Lee's ideas to heart and are developing tools that will make it possible to deal with all that data that we want, need, and will soon be available to us.