Linking the clouds

Richard Wallis
Technology Evangelist, Talis
Tel: 0870 400 5000
E-mail: Richard.Wallis@talis.com

I often observe how events, announcements and thoughts that seem to be unconnected on the surface conspire to produce a theme of the moment, and it has happened to me again with 'cloud computing'.

This theme kicked off for me a few months back whilst I was explaining the hosted software-as-a-service architecture that underpins the stable of new application services we are rolling out at Talis. I have been using the term 'cloud computing' for a while as another way of describing this—meaning more than just hosting—and recently it has started to connect with audiences. Maybe this is because there is now a general background noise of mainstream organisations announcing their commitment to cloud-based services: the Telegraph Media Group's decision, for example, to phase out Microsoft Office and Exchange for 1,400 of its users in favour of Google applications, or the regular stream of universities announcing their moves to cloud providers of e-mail. The latest entrant from the world of libraries into cloud computing is OCLC (Online Computer Library Center, Inc.), who recently launched web-scale circulation, acquisitions and license-management services alongside their WorldCat operation.

Then one fine morning last year, I was driving to work letting the usual waves of political infighting and current affairs emanating from BBC Radio 4's *Today* programme wash over me when I heard the words 'cloud' and 'computing' used in a context that had nothing to do with predicting the effects of global warming. A technology buzz-phrase has definitely made it when it appears on national drive-time radio.

There has been simple hosting of computers on behalf of customers for many years, but cloud computing is much more than that. Firstly, there is not a one-to-one relationship between computers and the customers of the applications they run. A good analogy here is the electricity-supply industry. You would not expect to visit a power station and find a vast array of individual generators feeding each and every customer, but that is how traditional computer-system hosting could be represented. As power stations were developed to satisfy the needs of thousands of customers at a time, the computer systems in the cloud are designed to run a single service that handles all the transactions of many customers. As you can imagine, producing applications that can handle the varying loads from many customers is a whole different scale of problem to the single application, for a single customer, on a single system that has been the previous norm. As with the power-generation industry, things do not stop with a single system. If your local power station stops working your lights do not go out. This is because the load is spread across a network of stations interconnected via the electricity grid. In the same way, cloud-computing systems are designed to spread their load across many computers, often located in many data centres.

The obvious benefit to moving to cloud-based services is that you do not have to purchase, house and maintain your own hardware. Suppliers of the service also have the benefit of having only one version of software to upgrade when they introduce new features. This allows them to roll out upgrades on a very regular basis, providing users with a constantly evolving service. For example, some of the Talis services I referred to earlier are regularly upgraded as often as once a month.

The trouble with in-vogue technology buzz-phrases is that they get used and abused by people and organisations to promote their view of the world, and cloud computing is not immune from this. Some simple traditional hosting solutions are being labelled as cloud computing even though they only exhibit a few of its attributes and benefits. When a solution is described to you as being 'cloud computing', it is worth checking what is actually meant by the label.

The other, not unconnected, theme of the moment is the 'semantic web'. It is not unconnected because it depends on data and services being delivered from the 'cloud' that we all interact with, the web.

The semantic web as a concept gained a bad press in the early days because most of the work behind it was taking place in research departments and the standards that underpin it were still in flux. It became caricatured as a utopian dream of a fully catalogued internet that would never come to fruition. The last piece of the semantic web standards jigsaw, SPARQL (**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage), dropped into place only last year, since when there has been an explosion of grass-roots community and commercial activity around publishing data using these standards. The pragmatic use of the standards to deliver useful benefits is now often referred to as 'linked data'.

The 'Linked open data project' is the most obvious example of this explosion.[1] Many openly shared data sets from diverse sources such as Wikipedia, the CIA, the Swedish National Library, PubMed, Thomson Reuters and the BBC are identified from here. These data sets are openly published on the web, in a similar way to how we are used to seeing web pages/documents openly published. The linked data web is not a replacement for, but just another evolution of, the web we have been experiencing for the last few years. In the same way that a link in a document on one web server can reference one on a different server, a link in one data set can reference an element of data in a set across the web on another server. Unlike document links, these semantic links carry meanings, such as 'location of' or 'published by'. The major benefit of the Resource Description Framework (RDF)—the format used to encode data in a linked data store—is that a link between one element and another is held as a Universal Resource Indicator (URI), the *http://* format of the web. This means that such a link can equally point to a set of data on the same system or to one on the other side of the planet.

'What is the real benefit?', you may ask. For example, a simple site storing place names could be greatly enhanced with live information from Wikipedia entries and location information from Geonames, in a way that the managers of that site could never justify the effort of compiling on their own. Simply linking data from one set to another, although powerful, still requires some management effort, though. SPARQL, the query language of the semantic web, enables such links to be identified on the fly. Taking my simple example, a site could query both DBPedia and Geonames, or even the CIA Factbook, for relevant information to include in a response to a user query at the time of the request, with the obvious benefits of data currency and absence of management overheads.

Library data has started to appear as linked data. In addition to the Swedish National Library publishing its catalogue data this way, the Library of Congress has published its subject headings data in this form.[2] It has been joined by OCLC, which had published Dewey summaries as linked data.[3] With just those two data sets available, imagine how much easier it would be to build into a display full descriptions of a book's classification without having to rely upon cataloguing effort! It is early days yet. Nevertheless, you can already see the dots appearing that, when joined up, will enable the sharing of data from library and non-library sources, to provide a much richer experience for library users.

The drawing together of the benefits emerging from these themes – and combining with other initiatives such as open data, being encouraged by the practice of government[4] and others – has the potential to totally change the landscape of our computing infrastructure and the way we deliver (and think about delivering) services to users. Not to mention the explosion in the devices those users are using …but that is a whole other theme.

References

1 http://linkeddata.org
2 http://id.loc.gov/authorities/search/
3 http://dewey.info/
4 http://data.gov.uk