IS THERE A LIBRARY-SIZED HOLE IN THE INTERNET?

David Weinberger is senior researcher at Harvard's Berkman Center for Internet & Society, and has been instrumental in the development of ideas about the impact of the web. Shortly before his recent keynote presentation at OCLC's EMEA Regional Council Meeting in Florence, he spoke with **Sarah Bartlett**

You rose to prominence as an internet thought leader, with pioneering texts such as *The Cluetrain Manifesto* and *Everything is Miscellaneous*. What led you into the world of libraries?

In *Everything is Miscellaneous* I explored the way the internet is redefining our ideas about how we organise things and ideas, and the move from physical to digital and networked library resources is a prime example of that. As a result of *Everything is Miscellaneous*, I was offered a position as co-director of the Harvard Library Innovation Lab.

Besides format changes, what is the most significant impact of the web on libraries today?

Library knowledge – the content; the metadata; what librarians and the community know about items held – is being lost to the web. This represents an immense amount of culture. The most basic components of the web are links, but if you want to talk about a book, what do you link to? There is no clear answer.

They might turn to Wikipedia, but only around 70,000 books actually have a page on Wikipedia, so people rely on commercial sites like Amazon. We aren't even meeting the most basic requirement, linking, much less having a way to refer to the history of the work, how it's affected people and culture.

Facebook holds huge volumes of information about its users and their lives, but we have no equivalent for what libraries know. That is a huge hole in the internet, and it has at least two negative consequences. Firstly, as library information becomes harder to find, it becomes less relevant. Secondly, libraries themselves become marginalised. The culture that libraries represent becomes invisible on the internet, and the perceived value of libraries diminishes.

Could libraries have done anything differently?

In the face of considerable challenges, libraries have done very good things, but it's going to take more. Libraries are providing open access to a closed world, and that is a tremendous service, but they are severely constrained by copyright laws that were not designed for a networked age. They also have limited budgets. But they were early to digitise their catalogues in the 1960s, and they were also quick to get on the web as portals. And for at least five years, there has been a blooming of library innovation, which is going down the right track.

And what is the right track?

Assuming that content remains locked up, then I think the right track is to make library information both public and interoperable where possible. Libraries can best achieve this by sharing their data, making it interoperable

'It's a huge technical challenge; many brilliant people are working on it'

so the rest of the world can mash it up with other data. There has been a huge amount of development across the library world in this area. It's a huge technical challenge; many brilliant people are working on it, and I think there's real progress.

Most of the data will be metadata; libraries can't publish the content of a book but they can publish information about the book. So that's bibliographic data and anything else the library can make available without violating copyright or breaching user privacy, including anonymised and aggregated usage data. They can also encourage their communities to talk more publicly about their interactions with the library. Libraries know a huge amount

about the cultural objects we've entrusted with them. We have thousands of years invested in our culture, and we lose it at our peril. It's not just about the books, and not even just the librarians; I also mean the communities that libraries serve. Through a library graph, libraries could collectively make available every scrap of information they have about those cultural items.

What do you mean by 'library graph'?

A library graph, which is a concept rather than a physical entity at this point, is a means by which the library world can publish its knowledge in an extensible and highly useful way. It draws on linked data, which breaks information down to its smallest molecules, forming simple statements that connect two things by a relationship and point to relevant web locations. Those statements interconnect within and across disciplines, to form a graph. Information previously stored in siloes now spreads right across the internet, and every time someone adds a new statement, the heap gets smarter.

Let's take the book as an example. When people talk about *Hamlet* or *Moby Dick*, they are generally not talking about specific editions, but there are times when editions and translations need to be distinguishable. Representing this complexity in terms of content, structure and ideas is a challenge that graphs can meet. With graphs we can traverse the entire space of ideas and information about the objects that libraries hold, discovering and capturing every conceivable type of relationship.

So people will add to the library graph, maybe linking to datasets elsewhere?

Yes. The graph would open up library knowledge to every other website, every field, every person, in the form of machine-readable

data. The danger is that if we don't do this, if we only use the web as a portal into the library, the culture that libraries represent may be overwhelmed by a more accessible and superficially attractive culture.

Will this machine-readable text eventually marginalise the web's human-readable content?

Preserving the human cultural endeavour is the most important thing. Machines can help us find content or even summarise it, but even that's not essential. Machine-readable metadata helps us find what we value, but ultimately the aim is to get words, pictures and sounds in front of humans.

How does the library graph help users to access content?

Firstly, the graph makes it easier for users to browse information in interesting ways. Just as Facebook surfaces relationships between people we know, so we might unearth new connections on the library graph. Secondly, the graph will make it very easy for other websites to talk about the works we've entrusted to libraries. Thirdly, and very importantly, users will be able to build their own services on the library graph, for example

creating recommendation engines that provide just the right degree of difference.

What do you mean by 'just the right degree of difference'?

I'm concerned about how we get people to encounter ideas that are just different enough from what is familiar to them. It's a very small degree of difference that works. If ideas are too different – like some deeply challenging revision of history, for example – then people will not be persuaded and will tend to dismiss them. Only with the right amount of difference will people appreciate something new and perhaps be changed by it. This is a real librarian skill: 'Here's something that's similar to *Gone Girl*; it's by this author, Patricia Highsmith from the 1950s, and you might try that.' With 'you might try that' the librarian is expressing just the right degree of difference.

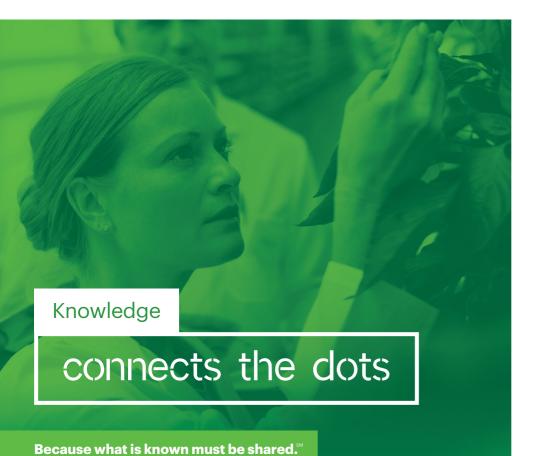### Will people be able to play an active role in the library graph?

Most of the textual analysis needed to generate the library graph will be algorithmic, but there are other ways of building the library graph. In the digital age, people interact with works of culture in a more public way,

by blogging or reviewing library resources online for example, and we can share these community contributions.

If people could use an open e-reader that supported the new open annotation standard, then their notes might become part of a continually-enriched corpus of text. By incorporating marginalia in metadata, we could reconstruct what a professor of history found interesting in a set of works. We could mash that up with other historians and thinkers.

How do you see the future of the library graph?

The future is unclear because we don't know what's going to happen with copyright; we don't know which business models will prevail; I certainly don't expect every library to cut back on the services they provide in order to engage in linked data projects. Libraries are already trying to do too much with too little. The movement towards linked data and the library graph is likely to come from those libraries that have the resources needed, and from funding groups that want to invest. ■

Sarah Bartlett is a freelance writer specialising in technology and the information industry.