GÖRÜŞLER / OPINIONS

Intelligent Information Retrieval Needs Smart Tools and Supporting Standards

Akıllı Bilgi Erişim, Akıllı Araçlar ve Onları Destekleyen Standartlar İster

Alan GILCHRIST[*]

Abstract

Though Google is an effective and popular search engine for locating information on the World Wide Web, it has tended to have the effect of blinding people to the enormous and growing problem of accessing and sharing digital information, much of which is not even available on the Web. If we are to effectively tackle this problem we will need a range of information processing standards and tools. To illustrate this belief, brief mention is made of the new British Standards and emerging International Standards dealing with structured vocabularies and interoperability; and of some systems which are leading the way in their application and development.

Keywords: *World-wide-web, Google, Information retrieval, Structured vocabulary, Standards*

Öz

"World-wide-web"de Google, etkin ve popüler bir arama motorudur. Fakat birçoğu Web'de bile olmayan bilgilere erişme ve onları paylaşma açılarından ortaya çıkan büyük ve gittikçe büyüyen sorunun varlığıyla, kişileri körleştirme etkisine sahip olmaya başlamıştır. Bu sorunla etkin bir şekilde baş etmek istiyorsak, geniş bir bilgi süreçleme standardına ve araçlarına ihtiyacımız vardır. Bu anlayışı örneksemek üzere, çalışmada, kendi arasında işlemlenebilen ve yapısallaştırılmış sözcük dağarına ilişkin yeni Britanya Standartları'na ve ortaya konulmaya başlanan Uluslararası Standartlara ve ayrıca uygulama ve geliştirme yolları açık olan bazı sistemlere kısa değinmeler yapılmaktadır.

Anahtar sözcükler: *World-wide-web, "Google", Bilgi erişim, Yapısallaştırılmış sözlük, Standartlar*

[*] Cura Consortium and Metataxis, UK.

Introduction

The "Father of Libraries" is said to have been founded by King Sargon of Akkad some 4500 years ago. The World Wide Web was first proposed by Tim Berners-Lee in March 1989, some 20 years ago. Small wonder, then, that we are still trying to get to grips with this most extraordinary technological and social development, particularly as it is still growing and developing so rapidly. And while we are trying to understand this revolution, libraries themselves are under threat, particularly when viewed merely as physical spaces warehousing material that *might* be requested. With this challenge in mind, the British Library commissioned a report on what was termed "The Google Generation" (CIBER, 2008): how to define it, what were its characteristics, and how did it impact on traditional libraries. First of all, the study concluded that the Google generation was a myth in that people of all ages were accessing the World Wide Web (www) not just the younger generation, and secondly that whereas people had relatively accurate mental images of the libraries they use, it was impossible for them to have any meaningful image for the Web. Brindley, the Chief Executive of the British Library sees this as a form of information (il)literacy, saying "Although young people demonstrate an apparent ease and familiarity with computers, they rely heavily on search engines, view rather than read, and do not possess the critical and analytical skills to assess the information that they find on the Web. These behavioural traits are also increasingly becoming the norm for all age-groups, from younger pupils and undergraduates through to professors...Most people including serious scholars tend to think that 'most' material is available on the Web – this search engine, two clicks mentality, will not serve us well as the basis for a digital future" (Brindley, 2009). Undoubtedly, libraries will have to adapt radically to the new digital environment, but this does not mean throwing away the traditional intelligence that libraries have deployed over the ages, but contributing that knowledge and experience to the development of the Web, and that implies challenging the current mentality of many, possibly most, users of the Web. Libraries and the Web must co-exist synergistically.

Yes, Google *is* a powerful search engine offered by an extraordinarily successful company, but it is not yet, and probably never will be, the answer to all search problems for all databases, intranets and websites. A recent search on the words "information retrieval" produced "about" 8,220,000 hits in. 21 of a second. Fantastic! But looking at the first page of hits we get the list below – all, admittedly, having something to do with information retrieval, but somewhat arbitrarily:

The Wikipedia article on information retrieval (IR)

An entire book on IR (from 1979)

Advertisement for a book on IR

Journal article on improving IR with tag clouds

Announcement of a Conference on IR

A university syllabus for an IR course

An "Address Not Found"

Advertisement for an IR journal

Advertisement for another IR journal

...and who reads beyond the second or third page of hits, let alone 822,000 pages? This is not meant to be a criticism of Google, which provides a splendid service for finding much useful and interesting material on the World Wide Web, a retrieval service that is based on a largely traditional search engine (as far as one can tell), and enhanced by its linking algorithm (operating rather like a citation index); the whole supported by truly massive computing power. To repeat, this is not a criticism of Google, but it is important to maintain a proper perspective before being dazzled by its magic.

Consider the following:

1. A futurologist with the Cisco Corporation predicts that by 2010 the Web will be doubling every 10 hours, and in ten years time, every 10 seconds (Financial Times, 2008, p.8).
2. A study by the University of Berkeley in California has calculated that the "Deep Web" contains some 91,000 terabytes, while the "Surface Web", which is easily accessed by search engines, contains a mere 167 terabytes (Deep Web, 2010).

...and a commonly heard observation that...

3. Google does not operate so effectively at the level of the enterprise intranet where its statistical algorithms have less "evidence" to work with.

The first two points underline the scale of the problem to be faced in the future of the Web, one which those working on Semantic Web technologies are hoping to solve, while the third point suggests that many individual systems – the databases, intranets and websites mentioned above – are struggling with their own smaller but significant retrieval problems, and struggling seems to be the operative word. The CEO of the French search engine company Sinequa is reported as saying "search has been so difficult to deploy with so many problems with relevancy and security management that IT people have lost hope and tend to stop projects before they start". If there are such problems at the enterprise level, how is it possible that "'most' material is available on the Web"? One answer has been proposed by Weinberger in his book with the catchy title *Everything is miscellaneous* (Weinberger, 2007). In his enthusiasm for the Internet and the power of social networking (directed in part at knowledge discovery), Weinberger perhaps goes a little too far in his jaded view of traditional library knowledge organization. He argues that there are "three orders of order"; the first where physical objects are ordered (e.g. shelf arrangement). The second where surrogates are ordered (e.g. the card catalogue), but that we are now faced with having to order the bits into which content has been digitized. Weinberger then says "The power of the miscellaneous comes directly from the fact that in the third order, everything is connected and therefore everything is metadata."In other words, he claims that we are now breaking away from hierarchical order to the infinite linking to be seen on the Internet, linking between resources with whatever labels anyone wants to place on electronic resources. Here, the word metadata has the widest possible meaning as does the word resource. Thus, the fact that A enjoyed a book by Orhan Pamuk can be used by the book supplier Amazon, applying accumulated similar metadata, to suggest to B that he or she might also like the book by Orhan Pamuk; or a photograph of Istanbul on the website Flickr, tagged by its owner as June 1999 might attract others who happened to be in Istanbul in that month. It must not be forgotten that the estimated number of Internet users at June 30, 2009, was 1.67 billion (Internet, 2010). The potential range of user types and queries suggested by this figure is mind-blowing, and as with the apparent conflict noted above between libraries and the Web, it is clear that we need both, and that we need all possible forms of knowledge organization if we are going to harvest the riches of the Web, as well as many other resources that are not currently available or easily accessible through the Internet.

Where Weinberger is undoubtedly right is where he recognizes the power of metadata, but this simple-sounding eight letter word needs to be defined and examined. As a start, a clear distinction should be made between the free-wheeling anarchic tagging seen in social networking and the disciplined and defined tagging used in more formal systems; though there is no intrinsic reason why the two should not co-exist happily for certain applications. There is much good work in progress in the establishment and application of metadata standards, aspects of which are well presented in a book by Zeng and Qin (2008). These two authors have compiled an excellent reference work which presents overviews of current standards for general purposes: Dublin Core (DC) and the Metadata Object Description Schema (MODS); as well as others devoted to cultural objects, visual resources and rights management. Following chapters take the reader through the "building blocks", covering *Elements,* some of which are the traditional features to be found in the Anglo-American Cataloguing Rules (AACR2)*, (e.g. Author, Publisher, Date of publication), while other *Elements* used for documentation rather than library purposes, include such features as Audience, Type of document (e.g. Report, Press release, Contract). Specific *Elements* are then collected into *Element sets* that together can be used in the description of resources of a particular type or purpose. *Element sets* can be complete standards such as the Dublin Core, or a selection of them, or an extension, taking care in all cases to preserve compatibility with other organizations where appropriate. Where compatibility is not an issue, an organization

* Set to be superseded by Resource Description and Access

may choose to define its own *Elements* and *Element set*. So far, only the framework has been established and it is obvious that each *Element* must be defined by some form of authority list (known in metadata jargon as a *Value space;* the range of which for an *Element set,* with accompanying rules, is known as an *Encoding scheme*). The *Value space* may be a simple syntax rule such as the convention that formats a date as DD/MM/YYYY, or a list such as *RFC 4646 – Tags for identifying languages* (RFC-Ref, 2010). More complex is the huge range of structured vocabularies including the well-known *Medical Subject Headings (MeSH)* or the *Dewey Decimal Classification,* which may be *de facto* standards themselves, others being compiled according to standard processes which are now described.

In the U.K., work has been completed on a fundamental revision of the old British Standards 5723 and 6723, which dealt with the compilation of monolingual and multilingual thesauri respectively (and which were later developed into the equivalent International Standards 2788 and 5964). All five parts of the new Standard have now been published, though the fifth has been issued as a Discussion Document rather than a Standard. [British Standards Organisation, 2005a; British Standards Organisation, 2005b; British Standards Organisation, 2007a; British Standards Organisation, 2007b; British Standards Organisation, 2008). These Standards mark a significant shift into the electronic world of the 21st Century, and attempt to address issues that confront a far wider audience than the old BS5723. Consequently, parts 4 and 5 deal with interoperability between vocabularies, particularly the knotty and labour-intensive activity of mapping; and the electronic exchange of vocabularies between different information systems. The thesaurus, which has evolved over the intervening years since BS 5723 was first published in 1979, is still seen as a basic and solid tool in the spectrum of structured vocabularies, but the new Standard also covers the vocabulary aspects of classifications, subject headings lists, business classifications for file plans, taxonomies, ontologies and semantic authority lists. Monolingual and multilingual thesauri are considered as being variations of basic principles, and treated accordingly. Work then started in 2008 on developing BS 8723 into an International Standard, and a Working Group comprising information specialists from the U.K., U.S.A., Canada, Australia, France, Germany, Spain and Denmark have been working hard to ensure that the results are truly international and treat problems of multilingual vocabularies accurately and comprehensively. The new International Standard has been given the title *ISO 25964: Thesauri and interoperability with other vocabularies,* and the first part: *Thesauri for information retrieval* has been formally circulated for discussion. Work is already under way on the second part, entitled *Interoperability with other vocabularies.*

While the private sector (with the exception of some areas such as the pharmaceuticals industry) has been lagging behind in the deployment of effective information management, the public sector, particularly in the U.S.A., has been notably active. The National Library of Medicine continues to develop the ambitious initiative called the Unified Medical Language System, a huge “Metathesaurus”, based on the Medical Subject Headings List (MeSH) and SNOMED CT, which is slowly and painstakingly mapping many smaller vocabularies into the system (UMLS, 2010). The National Science Digital Library (NSDL) is providing a single point of access to a range of scientific and mathematical databases, a service relying on the back-room mapping of different vocabularies (NSDL, 2010). These exercises in sharing and interoperability rely on the intelligent use of metadata (such as the Dublin Core metadata set), structured vocabularies (such as MeSH and the U.N. Food and Agriculture Organization’s Agrovoc Thesaurus) and authority lists (such as the International Standard Organization’s ISO 3166, listing country codes). But there is a huge range of such metadata element lists, structured vocabularies and authority lists, so that there is now a number of emerging initiatives whose aim is to make such tools available electronically through central registries. One such, described as “providing services to developers and consumers” is offered by the National Science Digital Library as an extension of its service mentioned above (NSDL Registry, 2010). As a further initiative the NSDL is a member of a collaborative project aimed at pooling resources into an even larger registry called the “Extended Metadata Registry Project” (XMDR, 2010). This is supported by nine public sector bodies, including the National Science Foundation, the Department of Defense and the National Cancer Institute. There is also one firm from the private sector, an international informatics company.

All of these initiatives are using and developing information handling standards, such as XML (the eXtended Markup Language), RDF (Resource Description Language) and SKOS (Simple Knowledge Organization System – used for the electronic transfer of thesauri). All of these are standards (albeit variations on other areas of application as in the case of XML). All of this is vital work, much of it promoted and supported by W3C, the World Wide Web Consortium dedicated to the establishment of standards that can open up the Web and make the Semantic Web a reality.

As a search engine for the Web, Google and other search engines will have vital roles to play and will continue to develop, but the ability of the Web to provide answers to questions is still a long way off. In the meantime, much effort must be expended on providing intuitive access to the resources that may hold answers or lead to answers on other sites (an obvious example is a search on Google for a topic such as ‘Black holes’, leading to retrieval of the Wikipedia site with definitions, discourse and references). But more complex paths must be created, and this will require (1) standards such as those briefly described above, (2) a wide range of structured vocabularies and authority lists (3) mappings between these vocabularies where useful, and (4) a higher degree of information literacy in the user population. These four requirements can be achieved only through human endeavour and the will to succeed.

References

Brindley, L. J. (2009). Challenges for great libraries in the age of the digital native. *Information Services and Use,* 29, 3-12.

British Standards Organisation. (2005a). *BS 8723-1 structured vocabularies for information retrieval part 1: definitions, symbols and abbreviations.* London: BSI.

British Standards Organisation. (2005b). *BS 8723-2 structured vocabularies for information retrieval part 2: thesauri*. London: BSI.

British Standards Organisation. (2007a). *BS 8723-3 structured vocabularies for information retrieval part 3: vocabularies other than thesauri*. London: BSI.

British Standards Organisation. (2007b). *BS 8723-4 structured vocabularies for information retrieval part 4: interoperability between vocabularies.* London: BSI.

British Standards Organisation.(2008). *DD 8723-5 structured vocabularies for information retrieval part 5: exchange formats and protocols of interoperability*. London: BSI.

CIBER. (2008). *Information behaviour of the researcher of the future.* London: University College London.

Deep Web (2010). Retrieved on January 7, 2010 from http://en.wikipedia.org/wiki/Deep_Web

Financial Times. (2008). September 17, 2008, p. 8.

Internet (2010). Retrieved on January 8, 2010 from http://en.wikipedia.org/wiki/Internet

NSDL (2010). Retrieved on February 7, 2010 from http://nsdl.org

NSDL Registry (2010). Retrieved on February 7, 2010 from http://metadataregistry.org

RFC-Ref. (2010). Retrieved on January 8, 2010 from http://rfc-ref.org/RFC-TEXTS/4646/index.html

UMLS. (2010). Retrieved on February 7, 2010 from www.nlm.nih.gov/research/umls

Weinberger, D. (2007). *Everything is miscellaneous.* New York: Henry Holt and Co.

XMDR (2010). Retrieved on February 7, 2010 from https://hpcrd.lbl.gov/SDM/XMDR/overview.html

Zeng, M. L. & Qin, J.(2008). *Metadata*. London: Facet Publishing.