Standards for Open Source Information Retrieval

Wray Buntine
Helsinki Institute of
Information Technology
University of Helsinki, Finland
buntine@hiit.fi

Michael P. Taylor
Index Data Aps.
London, England
mike@indexdata.dk

Francois Lagunas
Exalead
Paris, France
lagunas@exalead.com

ABSTRACT

Standards are important because they make a field more open to small and medium businesses and to academic players. We review a number of standards that apply to information retrieval and web search, and discuss the role that they play. We also discuss some areas where there is potential for the development of standards, where for instance information retrieval would benefit, and where standards development appears feasible.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION

The various standards for communications on the web, such as HTML and HTTP, SMTP, FTP, MIME, URI's and soforth, are the key infrastructure that made the phenomenal commercial growth of the web possible (arguably many other factors exist including including hardware and telecommunications, academic incubation, etc.). These standards allowed small, medium and large businesses to join in the development of products and services. A similar phenomenon occurred before this with IBM PCs and their standard hardware architecture. The web has proven to be a true revolution in business whose effects are still being seen (for instance web advertising recently overtook some forms of print in total expenditure). The importance of standards in creating an open playing field should not be underestimated.

With this motivation in mind, in this paper we review some of the major standards available to the information retrieval (IR) and web search communities. This serves two purposes: first, to survey the field and understand the current offerings; and second, to form a basis for discussion on the penetration of these standards and of new areas where standards might well be used.

Our approach to standards w.r.t. IR are presented in the second section. Then, in the third section, a number of fairly well established standards are presented and discussed. Information extraction is discussed separately because we believe it is important for the future of IR to extend its semantic capabilities, yet no standards currently exist. Finally, some more speculative proposals are considered.

2. ON STANDARDS

This section outlines some aspects of standards that are best clarified before covering the standards themselves.

2.1 What is a standard?

The standards we include in this discussion are taken from those listed at our website, `OpenSourceSearch.ORG`, that have been pointed out to us over time, or have been used by our own group. Notice that we exclude from the discussion standards in the area of traditional web activities, web services, web commerce and semantic web. The former are rather orthogonal to the issues of IR, and may be combined as needed. The latter, the semantic web, will play a growing role in IR, a role still being understood while semantic web standards themselves are in flux.

Moreover, we use the term standards rather loosely. Officially, standards should have been approved by some national or international body such as ISO, NIST, IEEE, etc. We also use the term to include proposed standards, *de facto* standards, protocols that have been published, perhaps by a commercial organisation, and have become in the colloquial sense, "standard."

2.2 Standards not covered

Finally, a special mention needs to be made of XML even though it falls in the category of a general and somewhat orthogonal standard. Embedding annotation (for instance, named entities) and document structure (for instance sections and subtitles) is generally done with XML. Moreover, it is the basis for semantic web standards such as the Resource Description Framework (RDF). When XML is used, languages such as XSL and XQuery become available. While XML is not an efficient format for communication, we look forward to advances from the Efficient XML Interchange (EXI) initiative. Thus, while XML and tools are useful and form an underlying technology, they are not critical standards to affect the direction of open source information retrieval.

Another standard here that should be mentioned is JXTA[1]: "a set of open protocols that allow any connected device on the network ranging from cell phones and wireless PDAs to PCs and servers to communicate and collaborate in a P2P manner." While it is a well known P2P standard, it is not generally used in distributed or P2P IR work because it is not well targeted for it.

2.3 A framework

In discussing standards, we will use a common summarising framework based around the important issues:

[1] `http://www.jxta.org/`

Target: What is the intended target for the standard. i.e., which group of developers or what functions are being highlighted. Described with the target should be a compelling reason as to why the standard should take on.

Advantages: What key advantages will the adoption of the standard offer.

Barriers: What are the potential barriers to the standard being adopted.

We were considering trying to present a systems or architectural framework for search to explain the various standards and their roles within information retrieval and search. This, however, proved elusive. Suffice it to say, that all standards resolve around an interface between key actors in a search system. Sometimes those actors represent separate businesses or users, and sometimes they represent separate software systems or components.

3. EXISTING STANDARDS

In this section, actual standards and current or emerging *de facto* standards are presented.

3.1 Query and retrieval: CQL and SRU

In the mid-to-late 1990s, the HTTP protocol that underlies the World Wide Web began to be used for more complex operations. Various digital collections were made available for searching by means of web interfaces, including library catalogues, museum collections, staff directories, aggregated article abstracts, and of course databases of web-pages, such as Yahoo, AltaVista and latterly Google. Most of these search UIs worked by using HTML forms to submit an HTTP GET request to a server – in other words, to generate a URL that contained the query together with auxiliary information such as the number of records to retrieve, whether to restrict results to those in a particular language, etc. Despite the growing use of AJAX techniques, the simple HTTP GET URL is still by far the most widely used technique for searching on the web.

In the absence of a standard for encoding rich queries into HTTP URLs, each of these services created their own conventions: for example, when searching for English-language PDF documents that contain the word "dinosaur", Yahoo uses the "va" parameter for the keyword, "vl" for the language (with the value "lang_en") and "vf" for the file format. Meanwhile, AltaVista uses "aqa" for the keyword, "kls" which has the boolean value "1" for English-language results only, and "filetype" for the format. Google is different again, using "as_q", "hl" (this time with the value "en") and "as_filetype". This inconsistency has not been an issue for service providers with no ambitions beyond providing a human-facing Web interface to their search engines, but makes it impossible to write a generic searching client that works across many different engines.

In response to this need, several standardisation processes have proposed wildly different solutions, which vary primarily in how they trade power against simplicity. At one end are the SOAP-based Web Services, which potentially achieve impressive results, but are hobbled by the complexity of the protocol and by numerous different implementations that do not properly interoperate. At the other end of the spectrum is OpenSearch, which has as its primary goal a low barrier to implementation, but which suffers from a correspondingly low level of precision in its queries. Perhaps the best trade-off between these extremes is SRU (Search/Retrieve via URL[2]), a candidate standard sponsored by the Library of Congress and developed jointly by an informal consortium of both public and private organisations from the commercial and academic sectors.

SRU was developed by librarians and engineers with many years' practical experience of the older information retrieval standard ANSI/NISO Z39.50. Thus it benefits from decades of experience of how to create specifications that facilitate semantic interoperability as well as the syntactic interoperability addressed by the other initiatives. The principal lesson of Z39.50 has been that semantic interoperability is both much more difficult and ultimately much more important than syntactic: that the same query can be broadcast to a hundred services is of little value if they interpret it in a hundred different ways.

Thus, from the beginning, SRU has been designed to enable queries to express precise semantics, using the related candidate standard CQL (Common Query Language[3]). CQL provides the means for individual query terms to be related to particular indexes (e.g. "author=farlow and subject=dinosaurs"), which allows a simple interface to semantically tagged content. For instance, a CQL search interface to MedLine can be configured to allow search fields such as Gene, Protein and Species. CQL indexes are taken from *context sets*, analogous to XML namespaces, an arrangement that allows domain specialists to create sets of indexes appropriate for searching in their domain, and provides the basis for simple semantic-based search. CQL context sets can also contain relational modifiers, boolean modifiers, date comparison, and refinement for sorting.

Although SRU and CQL capture most of the power of Z39.50, and add much that is new, their simplicity presents a very low barrier to implementation, and many free toolkits are available to facilitate the development of both clients and servers[4]. SRU has been adopted by the NISO Metasearch initiative as the basis for its searching profile, and SRU installations and implementations include those of the Library of Congress, Nature Publishing Group, Talis Information Systems and the Alvis project[5].

Summarising the key issues:

Target: search and information retrieval over the web, but primarily intended as a user-friendly replacement for the older Z39.50 protocol.

Advantages: Builds on the experience of the Z39.50 community from digital libraries.

Barriers: Information retrieval and digital libraries are not

[2] `http://www.loc.gov/sru/`
[3] `http://www.loc.gov/cql`
[4] e.g., YAZ, `http://indexdata.com/yaz/`
[5] `http://www.alvis.info`

strongly overlapping communities, where, for instance, evaluation of search results and the nature of content is quite different.

3.2 Metadata publishing

Two approaches provide metadata about resources.

3.2.1 OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting[6] (OAI-PMH) is a flexible standard for allowing the publication and harvesting of *metadata about resources*. Examples of metadata are the well known Dublin Core (DC) used to described primarily publication data about a document. OAI-PMH servers are expected to provide DC metadata at a minimum, and may include other metadata such as digital rights. The resources described are intended to be books, physical media, web services and web content.

The OAI-PMH protocol provides commands for enquiries about available tags (i.e., tagged subsets which sub-groups can be retrieved by), and available metadata formats, and commands for retrieval of metadata selecting by date or by tags. The protocol operates via HTTP and returns results in XML with entries for the header and the metadata.

The primary users of this technology are libraries and digital libraries and publishing organisations within large institutions such as universities. Major search engines have embraced the protocol to interoperate with digital libraries.

Because the protocol allows arbitrary metadata formats, for instance, XSL transformations can be applied when publishing content, it can be used for more general XML publishing and harvesting tasks where meta-data content is a natural extension of the Dublin Core. For instance, semantic annotations, in-link information and categorisations could be included.

Summarising the key issues:

Target: Publication of meta-data about digital (and non-digital) resources. Intended as a means to support distributed systems.

Advantages: Bulk access to sets of metadata. Is used by search engines as a means of accessing some digital libraries.

Barriers: Distributed systems for IR have not been successful.

3.2.2 RSS

The Really Simple Syndication (RSS 2.0) standard provides a way to syndicate the content of a website in a push manner. RSS has some Dublin Core style fields and is widely supported. One extension for multimedia is the Media RSS proposal by Yahoo that may become a *de facto* standard for multimedia. It contains descriptors for bit-rate, sampling and so forth to properly describe audio and video content. Proper discovery of multimedia and its properties is a known problem for crawlers.

[6] `http://www.openarchives.org/OAI/openarchivesprotocol.html` `A9.com`.

The following subsection on OpenSearch also makes use of RSS. Note that in contrast to OAI-PMH which provides metadata in batches, RSS provides metadata about individual resources in a timely fashion, as they become available.

Summarising the key issues:

Target: Timely notification about digital resources. Targets internet news and blogging community.

Advantages: Solves the problem of crawling actively updated and produced content.

Barriers: Has seen rapid growth in use, and extensions are being proposed to extend its applicability to multimedia and search.

3.3 Search syndication: OpenSearch™

OpenSearch™ is a "set of simple formats for the sharing of search results." It consists of an XML resource definition for a search engine's capabilities, a results format, and a simple query language embedded in URLs.

OpenSearch is a standard proposed by Amazon to aggregate efforts around its search engine, A9.COM. It is an open format the uses a Creative Commons license. Like the SRU/CQL protocol it can operate via HTTP and returns results packaged in XML, in this case an extension of RSS 2.0, the syndication format.

The underlying idea is that every search service should be able to accept a standardized input format, and provide its results in another standardized format. This way, it is very easy to build services such as meta-search on top of basic search services. Search providers offer a description file that is an XML specification of their capabilities. This has terms such as `AdultContent` and `SyndicationRight` to describe functionality. A related idea is the meta-search standard STARTS [4], that did see use, perhaps because it did not have the support of a company such as `A9.com`.

The query syntax of OpenSearch, compared to CQL with its rich pedigree, is very simple: a list of keywords. Thus, this does not allow the flexibility of simple semantic-based search as in CQL. Moreover, the results format, an extension of RSS, does not provide a mechanism for providing more general aggregate information related to the query such as related categories or named entities.

The primary role of OpenSearch is to provide search syndication interoperability for search engines and aggregators. For this task, it has a well designed though simple schema. A remarkable number of search engines have enlisted with the capability. Support for it is built into some open source systems such as Nutch and DataparkSearch Engine. However, there is no clear result aggregation strategy, so it is unclear how well the standard can be used as a basis for distributed or federated search in any general sense.

Summarising the key issues:

Target: Smaller search engines, to allow syndication on

Advantages: Allows syndication of content.

Barriers: Lacks a more general strategy for results aggregation.

3.4 Site documents: Google Sitemaps

Sitemaps is a scheme Google is testing to inform and direct crawlers about available pages (for instance, in the hidden web or under difficult URLs) and about frequency of update. Google ties this with advantages to the website maintainer in terms of unique feedback about the site's search characteristics.

Sitemaps it not a *de facto* standard, but nevertheless it has been very well received by the community, especially due to the rise of dynamic web pages and frequently updated pages (e.g., blogs that update daily). Thus, this may become a *de facto* standard. Unlike RSS and OAI-PMH which advertise and aggregate metadata, Sitemaps merely provide, in a passive way, information to the crawler and thus represent a lower overhead for websites developers.

Summarising the key issues:

Target: Simplify crawling of sites, using Google site tools as a lure to get website maintainers involved.

Advantages: Makes crawling simpler.

Barriers: Seeing rapid adoption.

4. INFORMATION EXTRACTION

Recent developments in question answering systems, retrieval in XML, and semantic-based search share the common goal of offering more structured content to a user based on some underlying semantics recognised in the textual content, and possibly pre-tagged in the documents. A critical step for this in some applications is the use of information extraction (IE) or some other natural language processing to semantically tag documents. In some cases, such as the Wikipedia, basic tagging may already exist in the content, but in general some form of information extraction is required.

This general area, embedding more semantic information in content to support richer retrieval, is undergoing rapid development. Unlike indexing in current IR systems, there is no agreement on the right general architecture to employ this additional semantic information. Named entities and relations might be extracted and placed in a database, or some custom processing of hierarchical term spaces might be embedded in a retrieval engine. Full inference systems such as the open source Sesame system[7] for RDF are not currently practical for large document collections.

To support the information extraction step, in ALVIS we have developed an XML linguistic annotation format that supports this task [1], based on an emerging annotation format of the TC37SC4/TEI workgroup.

In most IR systems, linguistic processing is usually performed immediately prior to indexing time, but this restricts the processing to crude methods such as Porter stemming. In ALVIS we have also adopted an open, extensible architecture for document processing that allows components to be developed independently for different tasks in the document pipeline. One advantage of this approach is that XSL can be used for efficient extraction and conversion of content when document processing services with different needs communicate.

Other platforms for information extraction in the broader information access context include GATE [2] and UIMA [3], both mainly based on the Tipster format, and both implemented as Java systems which programs plug in to. Our architecture instead uses XML as the binding mechanism.

5. OTHER DEVELOPMENTS

In other areas, there is the potential for systems and standards to support open source information retrieval and search. here we have considered some of the major components of an information retrieval or search engine system, w.r.t. their suitability for standardisation.

5.1 Distributed search

Many paradigms exist for distributed search and information retrieval including Federated search engines with query routing (effectively acting like a meta-search system), and various forms of peer-to-peer. Peer-to-peer search works now in multimedia applications where the searchable content is short title strings. Federated search works in the digital library context where simple and effective results ranking strategies exist such as by date, location, etc.

For general information retrieval applications in a non-trivial distributed manner, there is no accepted methodology at present. If and when a practical methodology does emerge, standards should follow.

5.2 Crawling

Two crawl-related standards discussed previously were SiteMaps and OAI-PMH. The former eases the crawlers work at a site, and the later provides a means for crawlers to collect and redistribute resource meta-data in bulk. Crawling is a task that can be distributed more easily, on the Grub[8] system has on open source client-side for a distributed crawler, with an open protocol.

5.3 Personalisation

Personalisation is a task that would be well supported by an open standard with a transparent, secure, and trustworthy implementation. While one strategy is just to use Google for all one's desktop applications, stepping outside the monoculture requires personalisation be available to many different web applications and to roaming users. Moreover, a suite of common support tools for analysis need to be provided so that diverse systems can integrate personalisation.

5.4 Results ranking

Results ranking has two critical factors that make it particularly amenable to a standards based approach. The first factor is transparency. Offering access to ranking schemes

[7] `http://www.openrdf.org`

[8] `http:/www.grub.org`

and justifying ranking at runtime are commonly proposed as advantages of open source search.

However, transparency also leaves the potential for rank manipulation by, for instance, rank spamming methods. The second factor favoring open ranking is the potential to develop means of supporting the circumvention of this same rank spamming. This is an ideal community based task if appropriate trustworthy controls are enabled.

5.5 Static ranking

Static ranking is the ranking of documents independent of any query. PageRank™ is a well known such ranking. It allows documents to be ordered within the collection to support, for instance, efficient results ranking. Static ranking can also be used on topic specific collections where different documents in the collection are more or less related to the topic of the collection.

5.6 Document storage

Why not standardise the storage of documents? As a very first step, documents could be stored in XHTML. The major problem is that HTML on the web is quite badly formed. Standard open source tools such as W3C's Tidy are still improving, and a common engineering approach is not to parse the HTML but instead to process it for word extraction, etc., basic tasks where parse trees are not required and robust tools exist to do partial parses on the fly.

5.7 Digital rights management

Digital rights management (DRM) is an essential feature to be integrated into search to broaden the pool of content, especially in the area of multi-media. This can already be seen in academic services such as Scirus[9], where some results are commercial content. Organisations such as the BBC and Deutsch Welle have large digital libraries which need protection if they are to be made available to the general public.

The Creative Commons provides a basis here for licensing, but it is not a digital rights management system. DRM standards need to be adopted by the large commercial organisations with the vested interests here.

6. CONCLUSION

A number of well developed standards, proposed standards and arguably *de facto* standards exist in the community, including CQL, SRU, OAI-PMH, Sitemaps, and OpenSearch.

CQL is best known in the digital library community, with traditions such as name spaces, fields, and data types such as dates. A lot of its functionality is shared by the specific query protocols adopted by IR systems such as Terrier and Lemur.

Sitemaps and OAI-PMH provide well thought out protocols for particular aspects of crawl and harvest. Both provide opportunity for use in the open source community beyond their original intension, with suitable extensions.

[9] `http://www.scirus.com`

Standards for tagging the results of information extraction (IE) are expected to see good use in open source IR because IE is a pipelined task that invariably requires a range of different tools, and the one tool can see common use on different IR systems. These two communities can interact well together through the use of standards.

6.1 Acknowledgments

The work is supported by the EU by the ALVIS STREP.

7. REFERENCES

[1] S. Aubin, J. Derivière, T. Hamon, A. Nazarenko, T. Poibeau, and D. Weissenbacher. A robust linguistic infrastructure for efficient web content ananalysis: the alvis project. In *Digital Semantic Content across Cultures*, Paris, 2006.

[2] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Anniv. Meeting of the Assoc. for Comp. Linguistics*, 2002.

[3] D. Ferrucci and A. Lally. UIMA: an architecture approach to unstructured information processing in a corporate research environment. *Natural Language Engineering*, 10:327–348, 2004.

[4] L. Gravano, K.-C. Chang, H. Garcia-Molina, and A. Paepcke. Starts: Stanford proposal for internet meta-searching (experience paper). In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 207–218. ACM Press, 1997.