INTEROPERABILITY ADAPTORS FOR DISTRIBUTED INFORMATION SEARCH ON THE WEB

RUBEN TOUS; JAIME DELGADO

Departament de Tecnologia, Universitat Pompeu Fabra (UPF)
Pg. Circumval·lació, 8. E-08003 Barcelona, Spain
e-mail: {ruben.tous, jaime.delgado}@upf.edu

The experience of low-barrier interoperability approaches like the Open Archives Initiative in conjunction with the success of Web Services technology encourage to think on new mechanisms to search and retrieve Web contents. We present a strategy that will allow heterogeneous Web resources to be searched and retrieved in the same way contents from digital libraries or other areas with well defined search and retrieval standards are. Simplicity, semantic-independence and specialisation are the key features of an approach that targets also the new metadata that is being disseminated as a result of the Semantic Web initiative.

Keywords: Information Interoperability, Information Search, Metadata, Adaptor Pattern, XML, Web

INTRODUCTION

While the promised land of information interoperability, the Semantic Web, is being woven, and without no guarantees about the time it will take to overcome the difficulties related to such an ambitious task, other less ambitious initiatives are defining protocols and languages to improve the way information is searched and retrieved. After the success of XML, it comes the frustration of viewing the impossibility to achieve universal agreements about metadata models, that implies a considerable constraint to the design and development of new search systems, capable of taking profit from the new information attached to digital resources. Nevertheless, initiatives like the Open Archives Initiative [1], GILS [2], SRW [3], and others, are demonstrating that it is possible to offer interesting search and retrieval capabilities over a huge set of resources by taking profit from the new technologies. Because heterogeneity is guaranteed for a long time, it is better to co-live with it rather than to ignore it, and that is the line we have been following in our research work.

This article exposes a strategy to facilitate mechanisms that allow Web resources to be searched and retrieved in the same way contents from digital libraries or other areas with well defined search and retrieval standards are. The achievement of this goal means to face the challenge of information interoperability, that has numerous facets including naming, metadata formats, document models and access protocols. This also means to take in consideration information extensibility, or community specificity. However, as noted in [4], interoperability strategies generally increase in cost with an increase in functionality, and maybe this point is the key difference between the success of the different initiatives.

The paper begins with a review of some different trends in the field of information search and retrieval on the Web and their evolution, to distill the conclusions that will serve to justify the necessity of a new approach. This part does not pretend to be a complete state-of-the-art of Web information search and retrieval technologies, but it is based in part in such

kind of works like [5] or [6]. Once analysed the current situation and extracted some conclusions the paper explores a complete different scenario, the field of digital libraries, where the course for interoperability began long before the Web, and that should serve as a reference when thinking in better solutions for the Web context. Next, we define and justify the concept of Interoperability Adaptors, that strongly relies on Web Services technology, and that aims to solve some of the problems related to the current approaches to Web distributed search. Finally we make some comments about some related works and expose our conclusions.

DISTRIBUTED INFORMATION SEARCH AND RETRIEVAL ON THE WEB

MOTIVATION. LIMITATIONS OF CENTRALISED SEARCH

The centralised approach to searching and retrieving information from the Web consists on downloading the maximum possible number of indexable text contents and then analyse them to generate the indexes that will be used to resolve the queries. Probably the best known example of this is Google [7], a Web search engine famous for using a smart and scalable strategy to rank the documents, based on the Web hyper-link structure. Despite the centralised approach has reached a great success (see Fig. 1), it has some limitations: a) the impossibility to cover the totality of the public indexable documents [8], b) the low precision provided by a keywords-based interface c) scalability, d) the impossibility to access dynamic contents.

The first point just says that there is no system that can face the challenge to download all the 'static' text contents of the Web. The study of [8] determines that the bigger search engines cover less more than 1/3 of the indexable Web. The second point says that simply specifying a sequence of keywords separated by blanks to discriminate documents among billions seems apparently not a very precise technique. Third point, scalability, stands not only for the problem to face the fast growing of the number of documents accessible through the Web, but also for the increasing number of crawlers on the net. Imagine the bandwidth consumed by thousands of concurrent robots trying to download and process all the indexable Web. It seems not a very efficient solution to replicate all this information that, furthermore, needs to be updated frequently. It is clear that here the classical trade-off between moving information and moving processes is being corrupted. Finally the last point remarks the fact that conventional crawler-based search engines are not capable of accessing the Web pages that are generated on-the-fly and that potentially may contain relevant information from databases or other sources. This is often called the 'deep web' [9], the 'hidden web' [10] or also the 'invisible web' [11]. Nobody knows really the amount of information accessible through the Web and not visible to conventional crawlers, but [10] estimates that it could be around 500 times greater than what can be reached by traditional crawling.

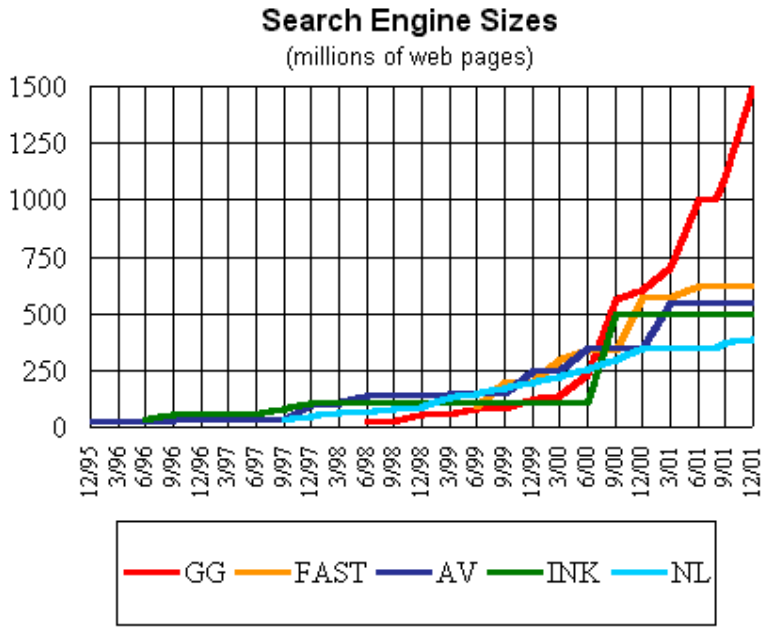

FIGURE 1 – SEARCH ENGINES EVOLUTION

METASEARCH. A ROUGH APPROACH TO DISTRIBUTED SEARCH

Metasearch (see [12] or [13] for a detailed definition) is a distributed search approach in the Web environment. Metasearch engines query a set of conventional crawler-based search engines that cover certain subsets of the public indexable Web. They also have the ability to access the hidden web, because they capture the results of the target sources on-the-fly, that allows them to harvest dynamically generated content. One important drawback of these systems, without considering legal issues (some conventional search engines explicitly forbid metasearch, for e.g. in Google's terms of service page we can find "You may not send automated queries of any sort to Google's system without express permission in advance from Google"), is that their work often does neither rely over an agreement with their underlying sources nor over specific interchange protocols. This forces them to face the problem to manage query forms and results pages designed for human consumption and written in HTML. We have had the experience to deal with some of these problems in some of our research works (see [14] or [15]) where we developed an advanced metasearch engine for spanish online newspapers. During this work we detected the necessity to define standardised machine-friendly mechanisms to access search applications connected to the Web.

THE FUTURE. THE SEMANTIC WEB

In the future it is supposed that Web contents will be not only for human consumption but also machine-understandable. This means that there will be metadata formalised in a standard way (RDF [16]) that will help software applications to understand the meaning (the semantic) of the information linked to the Web. This ambitious target is called The Semantic Web initiative, and has opened a broad spectrum of opportunities for improving the search and retrieve of information on the Internet. Of course this is not casual, but one of the main targets of this new scenario as pointed in [17] or [18]. However, the consolidation of a standardised way to interchange semantic information is just another step in the race for interoperability. Other battles are being fought to rationalise the way this information is

processed and search and retrieval are maybe the most important elements of the information feed chain.

The challenge is to find efficient and rational ways to exploit this new information that begins to be disseminated over the net, and that, despite of it is formalised in a standard way, it can be stored in different ways (embedded on HTML pages, in a database, in specific knowledge repositories, etc.) and it remains highly heterogeneous (an innumerable an unrestricted number of ontologies, potentially overlapped, can co-live in the Semantic Web).

DISTRIBUTED DIGITAL LIBRARIES. A GOOD EXAMPLE OF INTEROPERABILITY

Maybe the best example of a distributed search environment are digital libraries. Their solid tradition and their persistent interoperability effort have no equivalent in the Web context. A clear example of this is the ANSI/NISO Z39.50 Information Retrieval Application Service Definition and Protocol Specification [19], now an ISO international standard, whose first versions are from 1988, a year before the Web foundational document [20] was published. This gives an idea of how mature are the standards in this field, and how this enables the deployment of efficient distributed search networks and services. Recently, a Web Service based on the Z39.50 protocol has been released, SRW (Search/Retrieve Web Service [3]).

Another interesting and well-known example is the Open Archives Initiative (OAI [1]), a technical framework not intended to replace other approaches but to provide an easy-to-implement and easy-to-deploy alternative for different constituencies or different purposes than those addressed by existing interoperability solutions. The OAI harvesting protocol is meant to be agnostic to the nature of a data provider, since it supports those that have content with fixed metadata records, those that computationally derive metadata in various formats from some intermediate form or from the content itself, or those that are metadata stores or metadata intermediaries for external content providers [21]. It should be noted that the specific decision was to use unqualified Dublin Core [22] as the common metadata set. This decision was made based on the belief that the common metadata set in OAI is explicitly purposed for coarse granularity resource discovery. The OAI takes the approach of strictly separating simple discovery from community-specific description [23]. The harvesting protocol (OAI-PMH [24]) has been designed to be specially low-barrier, being compatible with any metadata format (but forcing repositories to offer Dublin Core). The difference between OAI and Z39.50-based approaches is described in [21]:

"*The OAI technical framework is intentionally simple with the intent of providing a low barrier for participants. Protocols such as Z39.50 have more complete functionality; for example, they deal with session management and results sets and allow the specification of predicates that filter the records returned. However, this functionality comes at an increase in difficulty of implementation and cost.*"

There are other interesting cases but it is not our intention to give here a complete state-of-the-art of distributed digital libraries. What is important is to understand that what has been achieved in digital libraries can be exported to other contexts. This implies that machine-friendly search interfaces, based on solid standards, can be also ambitioned for Web resources.

INTEROPERABILITY ADAPTORS FOR DISTRIBUTED INFORMATION SEARCH ON THE WEB

DESCRIPTION

An Interoperability Adaptor is a semantic-independent Web Service, with some simple search capabilities and the possibility to offer more precise semantic-based search mechanisms. This should allow specialised clients to find and use specific services without restrictions, and more generic clients to interact with all the services, independently of their specific features. This idea is strongly influenced by the OAI protocol (OAI-PMH [24]) that separates simple discovery from community-specific description [23]. However, and these are two important differences with OAI, on one hand the adaptors we are talking about are targeting any kind of information, including dynamic and static HTML pages, images, videos, audio, metadata, news or any data accessible through the Internet. On the other hand, these services are not, in principle, coarse grained (OAI-PMH does not offer precise mechanisms based on contents or metadata to filter the results of a source). Instead of this, they will offer the possibility to negotiate the query mechanism, that could range from conventional full-text queries to specific query languages like RQL. This negotiation will be based on WSDL [25] descriptions, that will specify the metadata formats and query languages supported by the service. For interoperability purposes, different layers of expressiveness and semantic-coupling should be offered when possible, for example a RSS-enabled [26] news provider could allow, apart of RSS-based queries, a more generic metadata harvesting mechanism (like TAP [27] for e.g.) or even a simple full-text search interface (applied considering the contents and also the metadata). This last feature, the definition of a simple access mechanism whose input is just a list of keywords (and maybe boolean operators) and whose output is well-formed XML but not constrained to any schema, could be the less restrictive layer of the service. This function could be used to expose for e.g. the work of conventional crawlers (generic or specialised), allowing the deployment of decentralised crawling networks where the metadata would flow from the sources to the consumers without the restrictions of semantic-dependent filters.

AN EXAMPLE

Fig. 1 shows an example of use of the adaptors. The scenario is the market of spanish online newspapers and their search services. We have some experience in this domain because we have designed and developed an advanced metasearch engine. It allows to express non-trivial queries, that the system resolves against the metadata extracted from the html results pages of about ten different spanish newspapers [14]. The generation of the http calls and the extraction of the metadata proved to be a hard task, overall considering that the formats of the interface and the results pages changed very often. In other parts it is usual that online news sources offer the possibility to harvest the information in RSS format, but that does not happen in Spain. What we thought then is that it would be fine that our work with each one of the newspapers could be reused by other information aggregators, and we immediately thought in deploying web services for these purpose. But, what metadata format the services should return? And what query languages or protocols they should know? We thought that it would be interesting to face this questions with a open approach, allowing the coexistence of different web services, offering different search features but having all of them a common interface and a common way to be described in WSDL. Published in UDDI [28] registers, the sources could be located and featured dynamically by the aggregators, taking profit from the existent UDDI infrastructure. In the example the grey shapes purpose is to highlight the fact that the adaptors have not been created nor maintained by the newspapers, we have developed them to feed our system but we have decided to make them open. The

"Basic Search" of the figure is a simple adaptor whose interface accepts only a list of keywords and returns results in XML. The deployment of services with different levels of semantic coupling aims to empower interoperability.

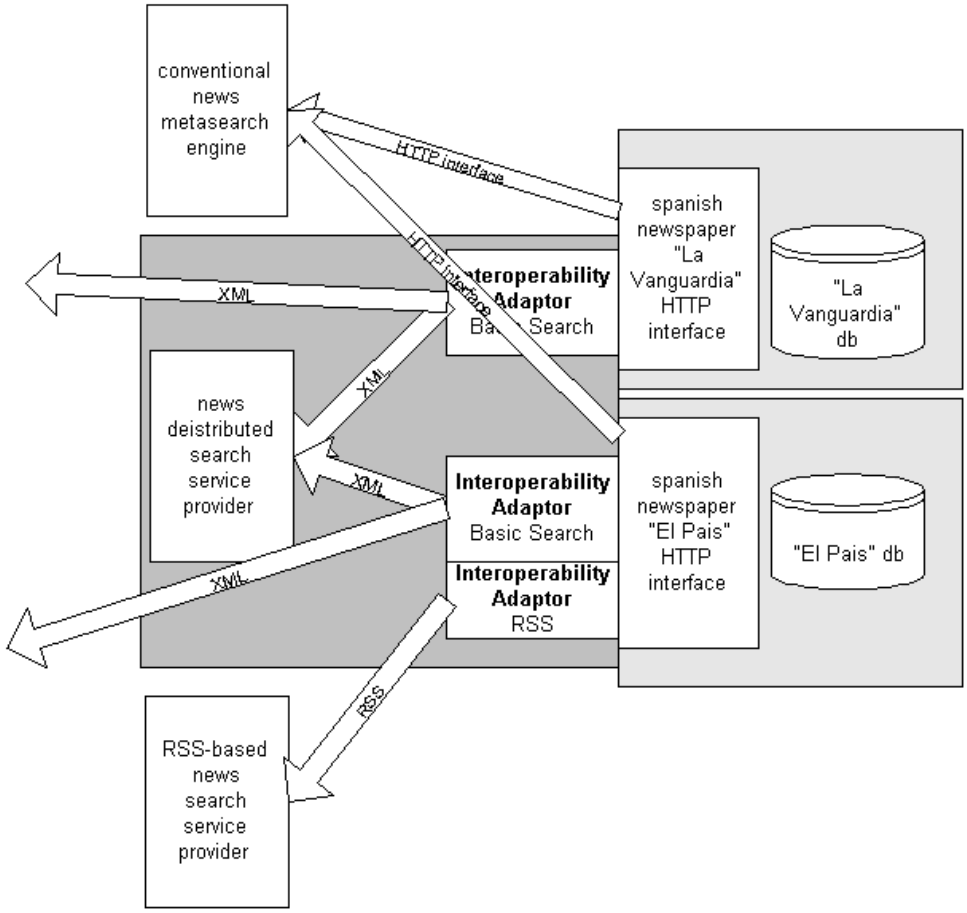

FIGURE 2 – Interoperability Adaptors Example

EXTERNAL IMPLEMENTATION

Ideally, the adaptors should be developed and maintained by the repositories administrators. However, we have developed a mechanism, inherited from the experiences of [14], to prove that in some cases and with some limitations this can also be done externally. This means that the back-end of the adaptor interfaces directly with the http interface of the source. Unstructured HTML content can be transformed in well-formed XML with tools like XML Tidy [29], that discards malformed fragments to respect the XML specifications. We have broadly tested it in the development of a semantic-enabled metasearch engine with success. Once serialized in XML, the metadata of the content can be extracted by applying a template, based on XSLT [30], XML Query [31] or simply XPath [32]. We have tested the second alternative by using hand-coded XML Queries without problems [14]. However, we are studying a process to automate the metadata extraction but we still have not definite results. It is interesting, for decoupling reasons, not to mix the metadata extraction mechanism with the specific metadata model required for the adaptor. Instead of this, it is better to define a new template that translates the raw metadata extracted to the adaptor specific format. The natural way to do it is by using XSL Transformations.

NEXT STEPS. SEARCH SERVICES NETWOKS

Till now, we have talked about a mechanism to standardize the way Web contents and Web search interfaces are accessed. However, we could imagine the advantages of generalizing this standard in other environments, like databases, knowledge repositories or digital libraries for example. In this situation we could keep talking about Interoperability Adaptors, or maybe better to call them simply Search Services. These kind of Web Services, providing standardised search interfaces and different levels of semantic capabilities, could constitute the required building blocks to build scalable distributed search networks that allow to access heterogeneous content providers and the information they are offering through the Internet. We talk about 'networks' in plural because specialisation is the future of Web search, and the only way to face the semantic heterogeneity problem. So we defend a model where multiple kinds of specialised search services and search aggregators co-live, constituting different networks that can be partially interconnected.

RELATED WORK

Several works from different contexts have similarities with what we have exposed here. Maybe SRW, from the digital libraries environment, seems apparently the more similar. SRW is a Web Service interface for Z39.50, and has the advantages and disadvantages of this protocol. SRW comes with XCQL, a very expressive query language specifically designed for the service. Z39.50 is a very powerful protocol for searching and retrieving bibliographic information, however it is also rather complex and implies some assumptions about the metadata attached to the contents. There are other initiatives, like GILS [2], that provide a smaller subset of Z39.50 aiming to improve the deployability of the service. The difference between OAI and Z39.50-based approaches is described in [1]:

"*The OAI technical framework is intentionally simple with the intent of providing a low barrier for participants. Protocols such as Z39.50 have more complete functionality; for example, they deal with session management and results sets and allow the specification of predicates that filter the records returned. However, this functionality comes at an increase in difficulty of implementation and cost.*"

CONCLUSIONS AND FUTURE WORK

Distributed search in the Web has to face the problem of interacting with interfaces designed for human consumption. This has constrained the success of the most part of Web distributed search approaches, like metasearch, incapable to face the challenge to adapt to heterogeneous and changing human-oriented interfaces. We have suggested an approach that can improve this situation by defining standardised and low-barrier access-points to information repositories, inspired in digital libraries protocols like OAI-PMH and implemented over Web Services. The target is to enable the deployment of distributed search environments capable to face the inherent heterogeneity of the Web context. To promote the migration from conventional http-based search interfaces to this new scenario, we have shown a possible way to implement the adaptors externally (without impacting the sources), but ideally they should be deployed and maintained by the target system administrators. In the medium-term, the results returned by these services can be used also by the emerging Semantic Web applications, that will benefit from initiatives that promote the dissemination of machine-understandable information. Now, we are working to refine a first specification of the adaptors. This implies defining the operations exposed by the Web Services and the way the services are described in WSDL. We are considering the possibility to create or reuse an ontology of Web objects that allows to explicit what kind of elements the adaptor is returning, independently of the schema or schemas it is using for these objects. We are also discussing

the necessity to define a compulsive minimum subset of functionality for the services, that could be a basic keywords-based interface returning just well-formed XML.

We believe that best ways to search and retrieve information in the Internet are possible. Initiatives like the Semantic Web pursue also this target, but till now their effort has been mainly directed to define mechanisms that improve the expressiveness of metadata. Now it is clear that it is also important to define how an why these metadata are going to be used, and how to face their potential heterogeneity. Time is demonstrating that if it is relatively easy to achieve agreements and standards about protocols and machine-understandable languages, it is not so easy to achieve the same about the semantic relationships of objects (digital or real), that each person views from a different perspective. Trying to overcome this difficulty, a lot of research work is being done to find good ways to map or combine ontologies, but this has proven to be a hard task, and its inherent complexity points to explore other ways to exploit the emerging semantic information. We have take this way, and our idea of Web Services providing different levels of searching functionalities and semantic coupling could serve to promote the use and dissemination of these metadata .

NOTES AND REFERENCES

1 The Open Archives Initiative. See http://www.openarchives.org/.
2 Global Information Locator Service (GILS) . See http://www.gils.net/index.html.
3 SRW. See http://lcweb.loc.gov/z3950/agency/zing/srw/specifications.html
4 Arms, W.Y.: Digital libraries. Digital libraries and electronic publishing. Cambridge, Ma.: MIT Press. (2000)
5 Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press New York Addison-Wesley (1999).
6 Kobayashi, M., Takeda, K.: Information Retrieval on the Web. ACM Computing Survey (2000).
7 Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems (1998).
8 Lawrence, S., Giles, L.: Searching the World Wide Web. Science (1998).
9 Bergman M.: The deep web: Surfacing hidden value. White paper, BrightPlanet.com (2001). See http://www.completeplanet.com/tutorials/deepweb/index.asp.
10 Raghavan, Sri., Garcia-Molina, H.: Crawling the Hidden Web. Proceedings of the Twenty-seventh International Conference on Very Large Databases (2001).
11 Chen, H., Lally, A., Zhu, B., Chau, M.: HelpfulMed: Intelligent Searching for Medical Information over the Internet. See http://citeseer.nj.nec.com/chen03helpfulmed.html.
12 Dreilinger, D.: Integrating Heterogeneous WWW Search Engines. Masters Thesis, Colorado State University. May, 1995.
13 Selberg, E., Etzioni, O.: The MetaCrawler Architecture for Resource Aggregation on the Web. IEEE Expert (1997).
14 Tous, R., Delgado, J.: Advanced Metasearch of News in the Web. Proceedings of the International Conference on Electronic Publishing 2002. (ElPub 2002)
15 Peig, E., Delgado, J. and Perez, I.: Metadata interoperability and meta-search on the web. Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2001).
16 Brickley, D. and Guha, R.: Resource Description Framework (RDF) Schema Specification (Proposed Recommendation), W3C (World Wide Web Consortium) (1999). See http://www.w3.org/TR/1999/PR-rdf-schema-19990303)

17 Berners-Lee, T., et al.: World-Wide Web: The Information Universe Electronic Networking: Research, Applications and Policy, Vol 1 No 2, Meckler, Westport CT, Spring (1992)
18 Berners-Lee, T.: Work in progress Sep 1998. Semantic Web Road map. See http://www.w3.org/DesignIssues/Semantic.html.
19 Z39.50. See ftp://ftp.loc.gov/pub/z3950/official/part1.pdf
20 Berners-Lee, T.: Information Management: A Proposal. CERN DD/OC, March 1989.
21 Lagoze, C., Sompel, H.: The Open Archives Initiative: Building a low-barrier interoperability framework (2001). ACM/IEEE Joint Conference on Digital Libraries.
22 Dublin Core. See http://dublincore.org/.
23 Sompel, H., Lagoze, C.: The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2), February 15, 2000.
24 The Open Archives Initiative Protocol for Metadata Harvesting. See http://www.openarchives.org/OAI/openarchivesprotocol.html.
25 Web Services Description Language (WSDL) 1.1. See http://www.w3.org/TR/wsdl.
26 RDF Site Summary (RSS) 1.0. See http://web.resource.org/rss/1.0/spec.
27 R. Guha, Rob McCool.: TAP: A Semantic Web Platform. See http://tap.stanford.edu/tap.pdf
28 UDDI Specification. See http://www.uddi.org/specification.html.
29 HTML Tidy. See http://www.w3.org/People/Raggett/tidy/.
30 XSL Transformations (XSLT). See http://www.w3.org/TR/xslt 12.
31 XML Query. See http://www.w3.org/TR/xquery/ 13.
32 XPath. See http://www.w3.org/TR/xpath.