Poseidon House
Castle Park
Cambridge CB3 0RD
United Kingdom

TELEPHONE: **Cambridge (01223) 515010**
INTERNATIONAL: **+44 1223 515010**
FAX: **+44 1223 359779**
E-MAIL: **apm@ansa.co.uk**

ANSA Phase III

Agents for Knowledge Resource Mapping in the World-Wide Web

Mark Madsen

Abstract

The World Wide Web is a largely passive environment containing enormous quantities of potentially valuable information in the form of hypertext structures. Much of the value and nearly all of the knowledge stored in the Web is contained in the way in which items are linked and cross-referenced, yet existing search engines, indexers, and robots tend to discard most of this useful navigational meta-information. This presentation will focus on how agents can be designed and utilised so as to generate maps of the Web that can be used by researchers and students interested in knowledge relating to specific subject areas. The problems of rapid validation of acceptance and rejection criteria will be given special prominence.

This paper is scheduled for presentation at CKBS-95 SIG Meeting: Intelligent Agents and the Next Information Revolution, Manchester Metropolitan University, May 9th, 1995.

APM.1473.01 **Approved** 2nd May 1995
External Paper

Distribution:
Supersedes:
Superseded by:

Agents for Knowledge Resource Mapping in the World-Wide Web

Agents for Knowledge Resource Mapping in the World-Wide Web

Mark Madsen

APM.1473.01

2nd May 1995

The material in this Report has been developed as part of the ANSA Architecture for Open Distributed Systems. ANSA is a collaborative initiative, managed by Architecture Projects Management Limited on behalf of the companies sponsoring the ANSA Workprogramme.

The ANSA initiative is open to all companies and organisations. Further information on the ANSA Workprogramme, the material in this report, and on other reports can be obtained from the address below.

The authors acknowledge the help and assistance of their colleagues, in sponsoring companies and the ANSA team in Cambridge in the preparation of this report.

Architecture Projects Management Limited

Poseidon House
Castle Park
CAMBRIDGE
CB3 0RD
United Kingdom

TELEPHONE UK	(01223) 515010
INTERNATIONAL	+44 1223 515010
FAX	+44 1223 359779
E-MAIL	apm@ansa.co.uk

Contents

1 Motivation for Mapping the Web

1.1 Overview of Resource Discovery: Problem and Solution

Resource discovery is one of the most important challenges facing any form of information system. The World-Wide Web is now the largest and most widely used information system ever constructed, and it is therefore necessary that it should have adequate methods for discovering relevant resources.

In general, the utility of any hypertextual system is strongly dependent on the availability of an effective categorisation of its components and a powerful means of navigation within the network. The purpose of this article is to show that an approach based on metadata provides for an effective categorisation mechanism, as well as resulting in a flexible class of cartographically based navigational strategies.

1.1.1 The problem

The World-Wide Web is a rapidly evolving information-rich environment. In its present form, it provides user access only to those resources which are explicitly known. Solving the problem of resource location is crucial to users who wish to exploit the Web's resource base, and therefore provides a commercial opportunity based on resource discovery systems and services.

Existing approaches to the resource discovery problem are based on Web robots and spiders which are used to build massive indexes of resources. This may be termed "just-in-case" resource tracking. The problems with these approaches are manifold: they scale badly and contribute heavily to network load (since they attempt to retrieve all all known resources for indexing), the indexing criteria are inflexible (since they are determined by the indexing system, not the end user), the resulting indexes are centralised and can be very large (since they are intended for multipurpose use), they are unable to provide any form of quality assurance (since resources may change in the long intervals between index updates), and have problems of authorisation (since they will be disallowed access to most sensitive or commercial sites).

1.1.2 The solution

The new technologies based on metadata objects and mobile agents will be able to solve all of these problems. Metadata is needed so that the appropriateness of resources with respect to specified criteria can be determined without requiring access to the resources themselves [MADSEN94]. Searching will be carried out in a "just-in-time" fashion, so that retrieved resources will be as up to date as possible. Search criteria will be encapsulated as metadata objects carried by mobile agents, which will be allowed access to sensitive sites depending on their levels of authorisation. These agents will conduct their searches based on user-specified criteria, expressed as scripts, to ensure the appropriateness of the resources returned.

All the components of this vision are in existence now. Most are widely available, while others will need to be added from programming models such as Java and Obliq [CARDELLI94].

The specific contribution of this article to the resource discovery problem is to show how mobile agents can build bespoke maps of the Web by using criteria relevant to a particular subject area or a specific user. These maps can then be used to deduce other interesting properties of the knowledge contained within a given region covered by the map. The cartographical knowledge so obtained can then be used as a basis for the decision to retrieve resources and resource sets from one particular area of the Web rather than another. An example of the utility of such an approach would be in maintaining the best possible mirror of a Web site.

In order for the strategies described below to work correctly, it will be necessary for the Web to contain sufficient metadata describing each resource it contains. At present, little metadata is maintained, and there is an almost complete absence of standards for the storage and structure of such metadata. This issue, however, is currently a high priority of the IETF Working Group on Metadata, and it is expected that a standard for the storage of metadata for use by all future Web applications will be approved by the IETF inside the course of the next year.

2 Metadata Query Management

2.1 Overview of Metadata Querying

As was pointed out above, metadata querying is the key to searching the Web effectively, and is the central component of the mapping strategies that will be developed below. The discussion below is specifically geared towards illustrating the issues arising in the mapping problem. A more thorough, and general, treament of metadata query construction and processing is given in [MADSEN94].

2.1.1 The structure of metadata

For the purposes of this discussion, it is obviously best to avoid assumptions based on strongly restricting the form of the metadata held in the Web, since this is not yet finalised. However, for the purposes of illustration, it is simplest to think of metadata as being held in the form of attribute-value pairs, with all attributes being optional. The specific form that any value can take is also not important, but for the sake of conceptualisation, it will be simplest to think of the values as consisting of text strings.

It is important to note that, despite these simplifications, the techniques that are developed in this article are applicable, given appropriate supporting technologies, to a wide class of metadata structures and systems.

2.1.2 Constructing metadata queries

The searching strategies based on metadata are constructed from knowledge criteria supplied by the user on the basis of that user's understanding of the subject and their ability to describe the class of resource for which they are searching. This means that any search criterion supplied corresponds to a template that should match the metadata representing the appropriate resource.

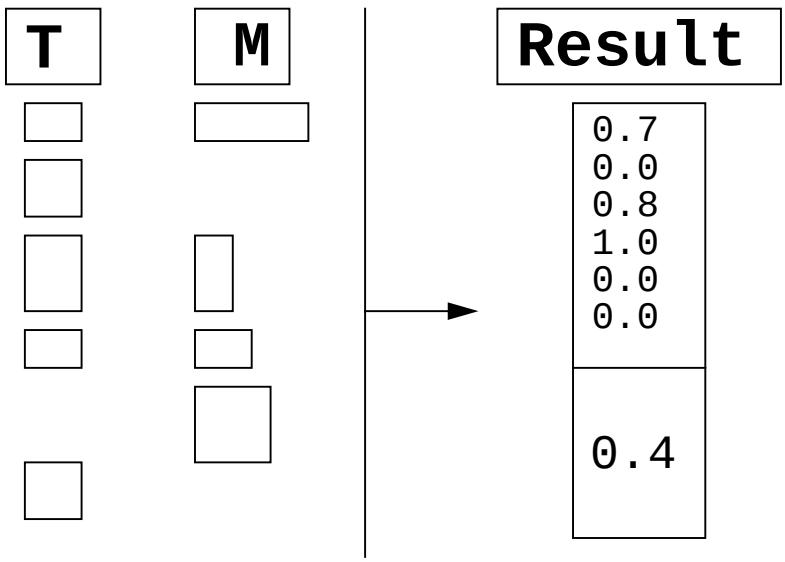
Typically, then, a metadata query consists of a query object: the data content of this object is the metadata template, while its methods are a specification of the manner in which the comparisons between metadata and query template are to be conducted. This is one of the respects in which "just-in-time" searching gives superior flexibility to the search methods supported by robot-constructed indexes: the latter must build their search criteria into the gathering process, while the former has the capability to specify what kinds of comparison are meaningful according to the kind of knowledge or resources which are being sought. At the same time, the proposed class of metadata-based schemes depends on the existence of high-quality metadata being maintained within the Web, or at least being accessible from the Web.

To make these ideas more concrete, let us illustrate the procedure by which a metadata query is constructed.

1. The user describes the elements required in the metadata template.

2. The system wraps these elements into the data of the metadata query object.
3. The user specifies the manner in which the query criteria are to be decided: this may be accomplished by choosing a predefined search format from a menu list, or it may be possible to define the algorithm or set of heuristics by which the criterion is to be decided. These can then be coded as scripts.
4. The system wraps the selected criteria into the methods of the metadata query object.
5. The metadata query object is transferred to one (or possibly a selection) of a set of query servers according to forward information held in the user's local system. The topic of forward information is not directly relevan to this article, so that little will be said about it here beyond the remark that forward information of the type required can be held effectively in a Whois++ centroid, a Harvest gatherer [BOWMAN94], an ANSA trader [THOMAS94], or some other WWW gateway, such as can be constructed within CORBA [EDWARDS95].

Figure 2.1: Illustration of the Metadata Matching Process



The action of the query object itself on another metadata object M can be set out as follows. Denoting the template metadata in the query by T, the query object Q computes the result via

Q: M -> R (T,M)

for each metadata object M to which it is allowed access. The result function will typically depend pairwise on the values of all the attributes possessed by both T and M, with both T and M being padded with any null-valued attributes needed to allow comparison with the other metadata object - this

process is shown in the abstract in Figure . Within the pairwise comparisons, the evaluation of the match could be computed, for example, as:

- The number of keywords contained in all of T, M, and a keyword list;
- The statistical similarity at a specified confidence level of the word-frequency tables for T and M;
- The number of common link anchors contained by T and M.

The essential points made in this section are that metadata queries

1. themselves contain the comparison criterion constructed as metadata;
2. contain the methods necessary to compare the metadata elements of the query and the target resource;
3. can express their criterion metadata and the associated methods in a richly customisable fashion.

These points show that metadata queries can straightforwardly be wrapped as objects, and that they should be created, delivered and managed by agents. In fact, the metadata query objects are already close to being agents themselves. However, for conceptual reasons it is simpler to think of them as being delivered by agents, which are responsible for the more complicated details of interactions with the remote systems, handling issues of authorisation, and maintenance of responses to queries. This way is also more efficient, because a single agent can handle multiple queries and multiple response objects as well.

3 Agents for Mapping the Web

3.1 Overview of Web Mapping

As was pointed out above, metadata querying is best handled by agents which can act autonomously. Together wit the query construction and implementation methods already described, these agents can build Web maps using the methods developed in this section.

3.1.1 Metadata topography

At this stage, it is useful to introduce the idea of the metadata subweb of radius N centred on M, which is the collection of all metadata objects that can be reached by starting at M and following no more than N links (some may be reachable in less if there are cycles in the graph).

The topography of the Web is now an easy concept to develop: a contour map of a subset of the Web can be constructed for a given metadata query object by having an agent that begins at a given resource, and propagates outwards by following links to other resources that are relevant (within the criterion set of the metadata query) to the query. Irrelevant resources will be shown as being at the base level of the contour space, while resources with a high relevance value R will lie on the tops of hills, mountains, and mesas. (Despite the language being used here, it may not be possible to represent the map on a 2-dimensional plane, since the Web can be arbitrarily nonplanar.)

The maps so derived will be valuable resources in their own right, since they can be used to deduce where the best starting points for searches will be, and where the richest concentrations of knowledge lie: for example, it will generally be a better idea to start mining the Web knowledge in the centre of a mesa than on top of a single high peak. Vector maps that show the prevailing trends and directions of change of the topography can be derived from repetitive explorations, and may be able to show how a particular area of knowledge has suddenly begun to grow and receive rapid input.

3.1.2 The role of the agent

As was pointed out above, the task of carrying the queries that map out the Web is ideally suited to the capabilities of mobile agents, since they can carry the necessary authorisations with them, and they have the possibility of using the map that they build up as an ongoing guide to the search for further relevant reosurces. They can also checkpoint their current locations via a trading or name resolution service as described above, so that they can be contacted by their owner, users, or sibling agents for the purpose of providing updated information even as they continue their task.

In this regard, agents have a number of obvious advantages over traditional robot technologies. One is that, because agents operate by exclusion of possible resources (rather than selection from amongst all possibilities, as do robots), they can be expected to consume fewer Internet resources per search. Another

is that, because the metadata search agents described here make use of distributed processing support [CARDELLI94] they can exist within a world of far more finely grained levels of security and authorisation than robots. The technology for secure remote processing is not new [REES93], but with the widespread availability of new implementations (in supporting languages such as Safe-Tcl, Obliq, and Java), the bottleneck lies mostly in the current lack of Web infrastructure support. When it is widely available on the Web, this technology will result in the metadata search agents being able to validate (or reject) the relevance of searched metadata sets rapidly.

In conclusion, then, the present article has demonstrated a novel and potentially useful role for agents in mapping the Web's resources for future use for purposes of research and commerce. It is urgent that the IETF standards be implemented that will allow the development of these and other experimental initiatives to begin enhancing the World-Wide value of the World-Wide Web.

References

[BOWMAN94]

Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U., Schwartz, M.F., *Harvest: A Scalable, Customizable Discovery and Access System*; Technical Report CU-CS-732-94, Dept of Computer Science, University of Colorado - Boulder, 1994.

[CARDELLI94]

Cardelli, L., *A Language with Distributed Scope*; DEC SRC Report, November 1994.

[EDWARDS95]

Edwards, N.J., *Object Wrapping for WWW: The Key to Integrated Services?*; APM.1464, Architecture Projects Management Ltd, April 1995.

[MADSEN94]

Madsen, M.S., Fogg, I.S. & Ruggles, C.L.N., *Metadata Systems: Integrative Information Technologies*; Libri 44(3):237-257, 1994.

[REES93]

Rees, R.T.O., *The ANSA Computational Model*; APM.001, Architecture Projects Management Ltd, February 1993.

[THOMAS93]

Thomas, G., Beasley, M., Hoffner, Y., *Data Management for an Enhanced Trader*; APM.1162, Architecture Projects Management Ltd, November 1994.