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

Future WWW Resource Discovery by Metadata Agents

Mark Madsen

Abstract

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. These approaches are inflexible and do not scale well.

The proposed 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. Searching will be carried out in a "just-in-time" fashion, search criteria will be encapsulated as metadata objects carried by mobile agents. These agents will conduct their searches based on user-specified criteria, expressed as scripts, to ensure the appropriateness of the resources returned.

This presentation summarises the primary issues for future WWW resource discovery schemes, describes the infrastructure that must be built to support those schemes, and gives a snapshot of ANSA's involvement in the both development of appropriate WWW standards and the prototyping of these ideas.

APM.1490.01 **Approved** 23rd May 1995

Project Management (confidential to ANSA consortium for 2 years)

Distribution:
Supersedes:
Superseded by:

Future WWW Resource Discovery Using Metadata and Agents

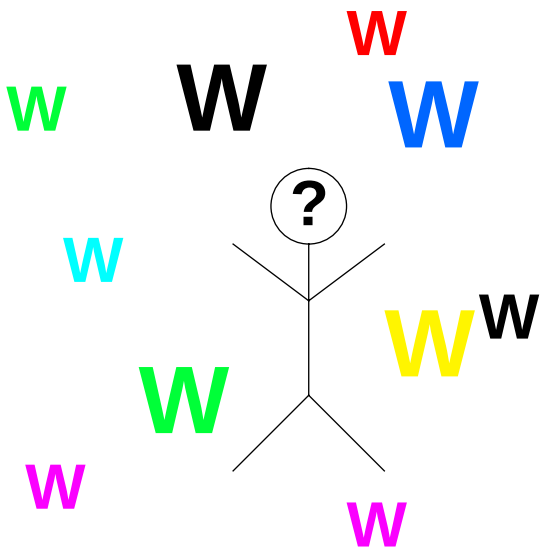
ANSA Technical Committee Presentation, May 1995

Mark Madsen

msm@ansa.co.uk

Presentation Overview

- The Resource Discovery Problem
- The Need for Metadata
- The Need for Agents
- The Required Infrastructure



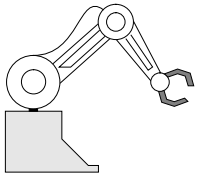
The WWW is ripe for exploration and exploitation

The Resource Discovery Problem

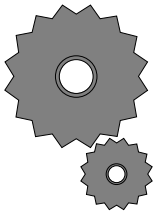
- Main problem is knowing what is out there
- Hierarchy of approaches
  - Robots using HTTP directly (eg [Lycos])
  - Custom indexes using brokers (eg [Harvest])
  - Autonomous agents using metadata
- Solution involves
  - resource location within the Web
  - navigation of the resource space

The WWW needs to support resource discovery from within

Web Robots



- **Robot characteristics are**
  - Web client running on a single host site
  - Contribute heavily to network load: one source states up to 40%
  - Many robots are criminally inefficient
  - Indexing is done in a just-in-case fashion
  - Categorisation is necessarily restricted to one-size-fits-all
  - Scaling is poor on all fronts

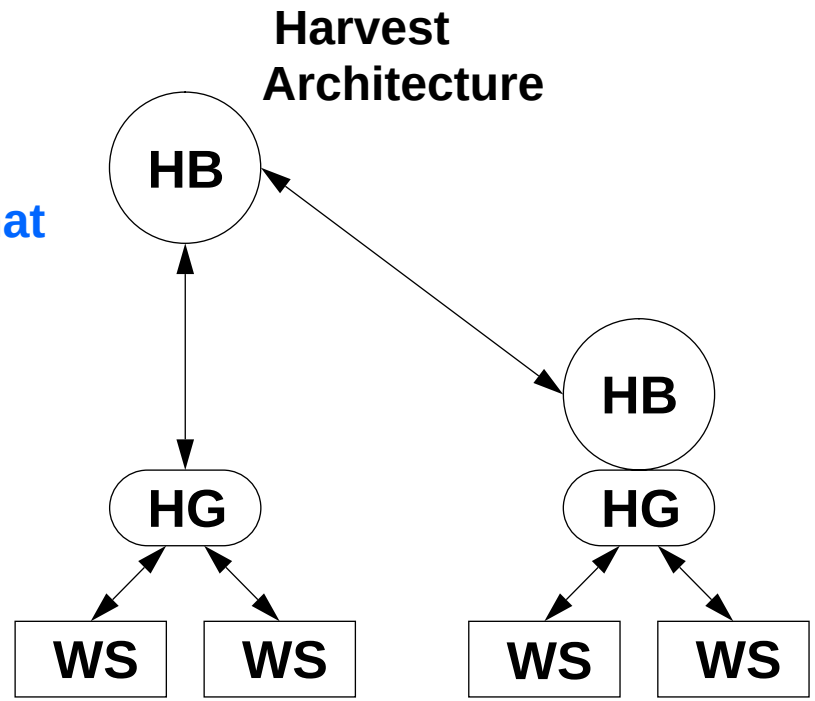


- **Robot-generated problems have led to**
  - Realisation of the need for robot/agent ethics [Eichmann]
  - Standards for exclusion of robots from servers [Koster]

Robots are not the long term solution

Custom Indexing Systems

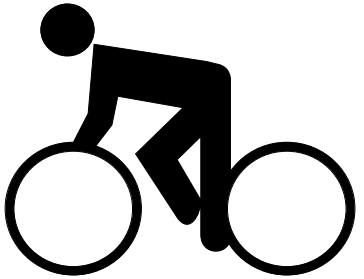
- Harvest is the archetypal custom indexing system
- Components are
  - Broker
  - Gatherer
  - Summary Object Interchange Format
- Advantages of Harvest
  - minimal wastage (SOIF)
  - good scaling (brokering)
  - shallow search (gathering)
  - narrow focus (start points)
  - customisable (individuality)



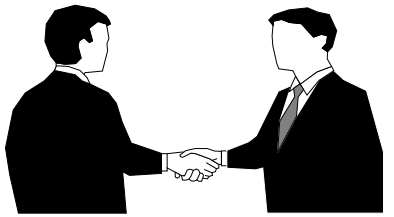
Brokering is essential to resource discovery strategies

Autonomous Agents

- **Agent characteristics [Madsen]**
  - lightweight: processing load controlled by server
  - directed: agents are sent only to good prospects
  - highly specific: agents can restrict searches on the fly
  - personalisable: agents can reflect and exploit their owner's knowledge



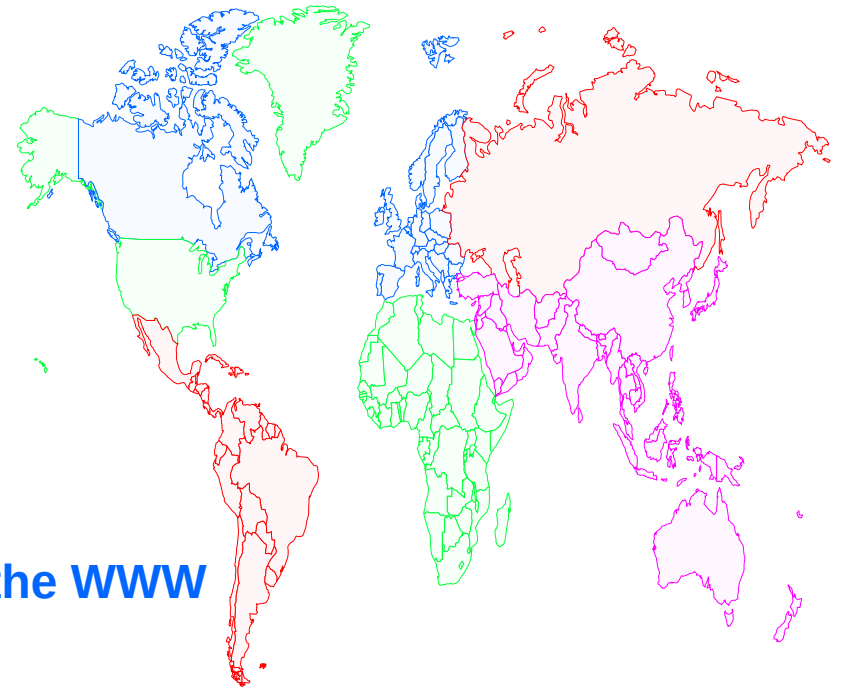
- **Agents need metadata to perform their functions**
- **Agents collaborate well with brokers supplying forward metadata: the brokers act as routers for the agents**



Agents are essential for extracting info from the WWW

The Need for Metadata

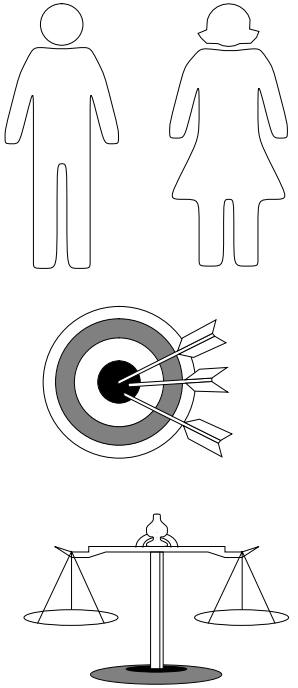
- **Resource metadata used for**
  - resource location
  - classification and ranking
- **Cartographical metadata used for**
  - restricting robot accesses
  - guiding agents
- **Legacy metadata used for**
  - windows onto legacy applications
  - gatewaying into the older parts of the WWW



Metadata is the lever needed to use the WWW

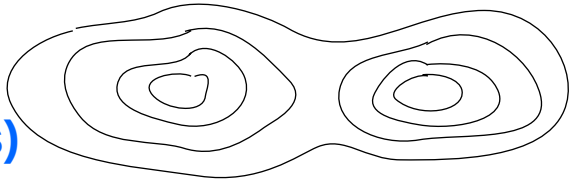
The Need for Agents

- **Need to use their characteristics of**
  - **personality**
  - **specificity**
  - **targetability**
  - **scalability of searches**
- **Need to use their capabilities for**
  - **mapping**
  - **navigating**

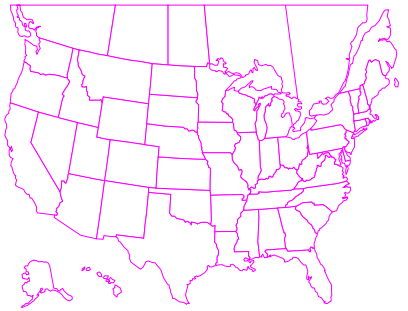


Agents can explore every nook and cranny of WWW

Cartographic Agents on the WWW



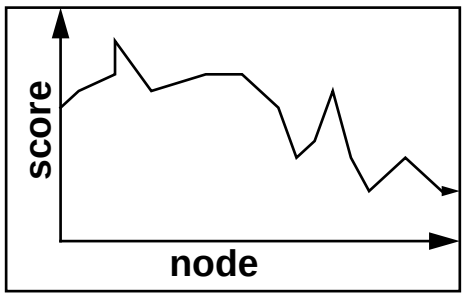
- WWW map is a set of nodes and *contours*
  - nodes are resources (HTML pages, CGI scripts)
  - contours are *isometric levels* of knowledge valuation for a subject
- Maps preserve structural information (connection and direction)
- Maps depend on the search criteria
  - different agents build different maps
  - agents make better cartographers than robots
- Maps complement indexes
- Maps provide a guide to locations of concentrations of resources

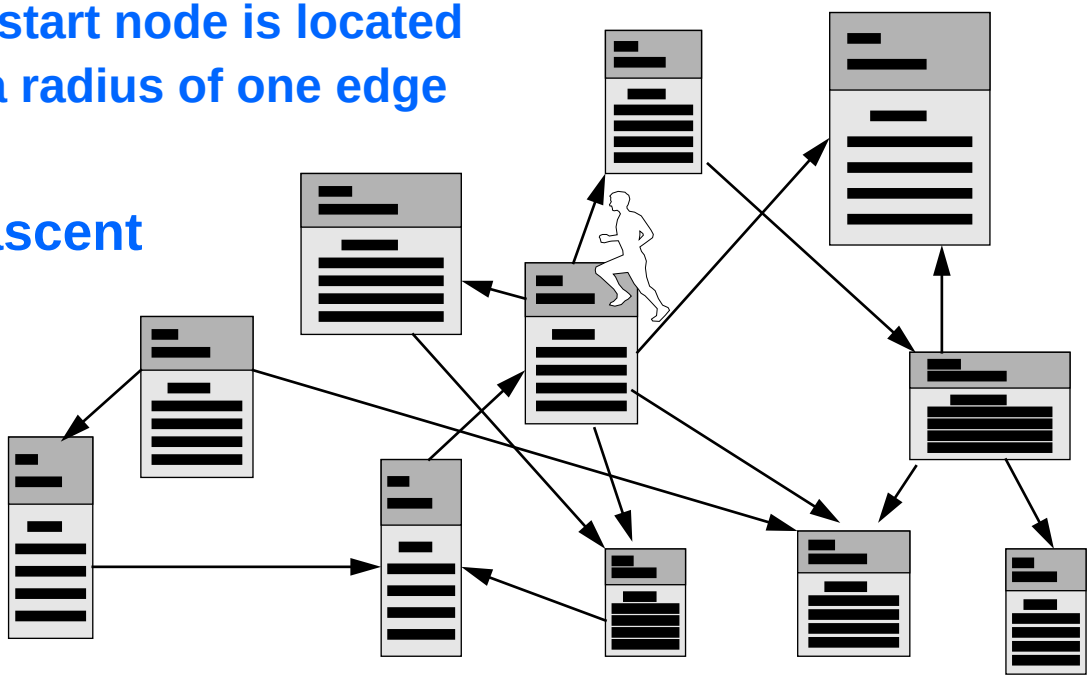


WWW maps are a valuable resource

Building A WWW Map

- Agent process for WWW map construction
  - migrate to server where start node is located
  - evaluate objects within a radius of one edge
  - repeat
- Can restrict to steepest ascent





Map metadata facilitates discovery services

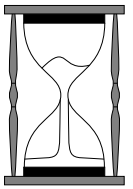
Infrastructure for Resource Discovery

- **Infrastructure for Metadata**
  - standards: URC proposal soon to go before IETF [Daniel]
  - storage and management facilities widely supported
  - sophisticated query construction methodologies (Vex, [Microcosm])

- **Infrastructure for Agents**
  - ethics for both agent behaviour and treatment
  - safe environments for agents and hosts
  - host-based navigation facilities

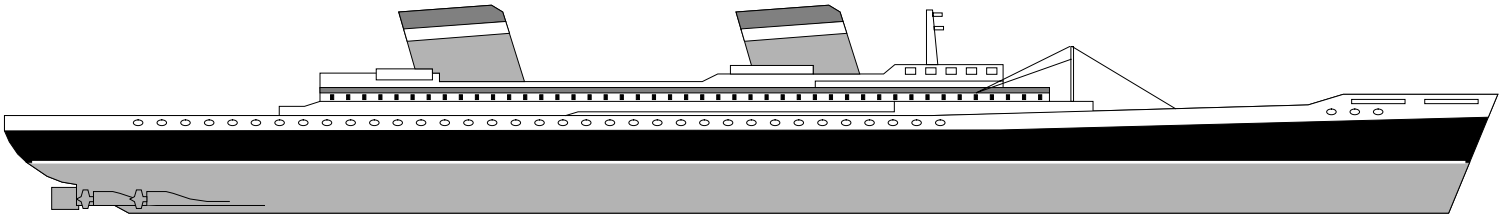
- **The ANSA Contribution**
  - Metadata: standards work for IETF: URC specification
  - Agents: prototyping of ideas using existing ANSA technology

WWW needs new standards in place urgently



The ANSA Contribution: Metadata Standards

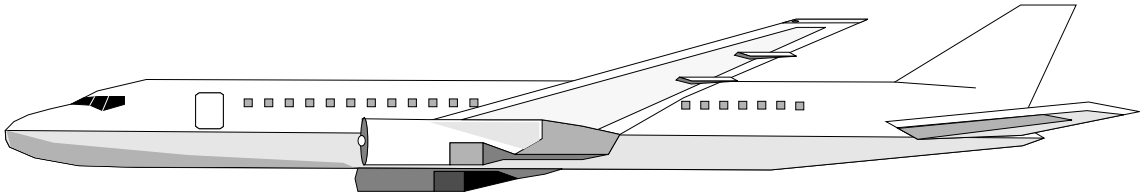
- Metadata standards work under the auspices of IETF
- ANSA collaborating with Ron Daniel (Los Alamos National Labs)
- Metadata will be encoded as Universal Resource Characteristics
- URC Requirements and Scenarios draft nearly complete
- Further work is concentrating on the draft URC Specification
- Floating ANSA ideas into the WWW via IETF standards



ANSA is working to define new WWW infrastructure

The ANSA Contribution: Prototyping Work

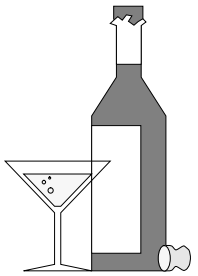
- Aim is to prototype ideas using ANSA technology
  - Use existing technology where possible (Matchmaker)
  - Build new technology where necessary (Changeling)
- Metadata searching prototype under development uses
  - Changeling webserver to incorporate custom search engine
  - URC repository built from the Matchmaker with a Tcl gateway
- Need to release code to make ANSA fly in the WWW



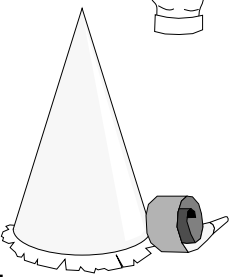
ANSA is working to develop new WWW capabilities

Summary

- Resource discovery is the most important problem for the reconstruction of the WWW
- Metadata is the lever necessary to query, locate and evaluate the resources hidden in the WWW
- Agents provide the most flexible, controllable, and cooperative solution to resource discovery through mapping activities
- The infrastructure needs a major overhaul to support the requirements of resource discovery
- The existing WWW does not suffice: it is time for WWWng



The fun is only beginning!



References

- [Eichmann] David Eichmann “Ethical Web Agents” <URL:http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/eichmann.ethical/eichmann.html>
- [Harvest] Bowman, Danzig, Hardy, Manber, Schwartz, “The Harvest Project” <URL:http://harvest.cs.colorado.edu/>
- [Koster] Martijn Koster “World Wide Web Wanderers, Robots and Spiders” <URL:http://web.nexor.co.uk/mak/doc/robots/robots.html>
- [Lycos] <URL:http://lycos.cs.cmu.edu/>
- [Madsen] Mark Madsen, “Agents for Knowledge Resource Mapping in the World-Wide Web” APM.1473
- [Microcosm] Wendy Hall, “The Microcosm Hypermedia Research Project” <URL:http://bedrock.ecs.soton.ac.uk/Microcosm/>
- [Daniel] Ron Daniel, “Universal Resource Characteristics” <URL:http://union.ncsa.uiuc.edu/HyperNews/get/www/URCs.html>