

DigiCULT

Resource Discovery Technologies

for the Heritage Sector



Thematic Issue 6

June 2004





RESOURCE DISCOVERY TECHNOLOGIES FOR THE HERITAGE SECTOR

THEMATIC ISSUE 6



John Pereira

Introduction and Overview

5

Guntram Geser

Resource Discovery - Position Paper: Putting the Users First

7

Interview with Jussi Karlgren

Empowering the User through Context Aware Search Tools

13

Michael Steemson

DigiCULT Experts Seek Out Discovery Technologies for Cultural Heritage

14

Interview with Pia Borlund

Humanistic vs Machine Centred Information Retrieval Approach

21

Joemon M. Jose

Personalisation Techniques in Information Retrieval

22

Interview with Andreas Rauber

Building a Neural Network Based Digital Library

27

Douglas Tudhope and Ceri Binding

A Case Study of a Faceted Approach to Knowledge Organisation and Retrieval in the Cultural Heritage Sector

28

Selected Resources

34

The Rome Forum Participants and Contributors

35

DigiCULT: Project Information

39

Imprint

40





INTRODUCTION AND OVERVIEW

By John Pereira

FUNCTION AND FOCUS

DigiCULT, as a support measure within the Information Society Technologies (IST) Programme, provides a technology watch mechanism for the cultural and scientific heritage sector. Backed by a network of peer experts, the project monitors, discusses and analyses existing and emerging technologies likely to benefit the sector.

To promote the results and encourage early take-up of relevant technologies, DigiCULT has put in place a rigorous publication agenda of seven Thematic Issues, three in-depth Technology Watch Reports, as well as the DigiCULT.Info e-journal, pushed to a growing database of interested persons and organisations on a regular basis. All DigiCULT products can be downloaded from the project Website <http://www.digicult.info> as they become available. The opportunity to subscribe to DigiCULT.Info is also provided here.

While the DigiCULT Technology Watch Reports address primarily technological issues, the Thematic Issues focus more on the organisational, policy and economic aspects of the technologies under consideration. They are based on the expert round tables organised by the DigiCULT Forum secretariat. In addition to the Forum discussion, they provide opinions of other experts in the form of articles and interviews, case studies, short descriptions of related projects, together with a selection of relevant literature.

TOPIC AND CHALLENGE

This sixth Thematic Issue concentrates on the question of how resource discovery technologies can ensure that the high value, authoritative heritage information placed on the Internet is effectively found, retrieved, and presented to information seekers. In an effort to circumnavigate the expansive area of interesting technologies, this Thematic Issue will focus on *user-driven approaches* in heritage resource

discovery. But before leaving the open savannah for the user trails, it is important to note that, overall, the expansion of the Web has favoured the emergence of new search applications, usage patterns, and interaction paradigms. These developments will have, and in fact are already shown to have, a strong impact on the users' expectations of the technological applications, interfaces and additional tools offered in the heritage sector.

Tailoring heritage online services to users' immediate needs is essential. According to a *New York Times* article 'Old Search Engine, the Library, Tries to Fit into a Google World' (June 21, 2004), a survey of 1,233 students across the USA, 'concluded that electronic resources have become the main tool for information gathering, particularly among undergraduates'. This explains part of the recent development that major (American) library information centres co-operate with Google, Yahoo and other powerful search engines for indexing their catalogues.

A big issue in fact is how to deal with the massive information resources of the so called 'Deep Web' that today are missed by search engines, and include also content-rich, mostly publicly funded databases of libraries, archives and museums. If, as some experts suggest, Google or a consortium of major search engines would be allowed to index these resources and take on more of the role of a 'universal service' provider, is there scope for unbiased, non-commercial driven information presentation? And what is the heritage sector's response?

OVERVIEW

Setting the scene for this Thematic Issue, the position paper first concentrates on the issues raised by most information seekers' choice of powerful search engines over trying to find out and mine maybe more valuable 'Deep Web' databases. Then it provides a look at current and future resource discovery technologies from the perspective of the academic and educational user. While academic users already

seem relatively well served, in contrast the expectations of educational users expose heritage institutions to more fundamental challenges in providing online access to their resources. Taking the lead from the entertainment industry, educational users seek compelling and interactive resource discovery environments. The fact that users benchmark their educational offerings against the industry standards in the 'edutainment' sector needs to be recognised and designed into heritage online services.

Three interviews provide different perspectives in overcoming the still considerable barriers to meaningful search and retrieval of heritage resources. Jussi Karlgren offers a unique approach on how to recapture the added value provided by librarians and curators, which he believes has been lost in the digital revolution. He expresses the need for research into technologies that enable these information and subject experts to fill the role of information brokers online. This form of stewardship would ensure that users have authoritative information and would better assist them in the formulation of their search needs. Such support, if performed correctly, can eliminate the feeling of isolation and prevent search fatigue, which ultimately results in users accepting unsatisfying – or, worse, misleading – results.

Andreas Rauber begins from the standpoint that there may never be a perfect way to represent a piece of content, so now may be the time to redirect some research effort towards new approaches. His approach is to put objects in an information space rather than define a representation schema for each class of objects. He describes the use of an unsupervised neural system to classify and cluster objects according to content. The research has also expanded into music files with clustering techniques able to organise music collections into genres of similar music. The capacity to fine-tune the classification is built into the system.

Pia Borlund is a firm believer in user-driven technological development. Her research focus is to adapt retrieval systems to the user and not the other way round. To better understand user needs she has developed a new method for evaluating information retrieval systems. Initial trials indicate that the more cognitively different the representations are that point towards a certain document, the higher the probability that the document is relevant to a given set of criteria – an approach she believes will have application value for the cultural heritage sector.

Jose Joemon in his contribution begins with the dilemma that users are poor at formulating a query that fulfils the needs of a machine-based query system. He therefore sees the value in continued devel-

opment of retrieval tools based on personalisation methods. He presents unique approaches to refine the search based on implicit feedback techniques. The techniques include bundling of sources to overcome loss of context and a query-less interface based on object selection to define a user's information need. Whilst not as accurate as explicit feedback, it has been demonstrated that implicit feedback can be an effective substitute for explicit feedback in interactive information seeking environments.

Michael Steemson summarises the Forum's expert discussion, which asked, on behalf of cultural heritage clients everywhere: 'How can we find something if we don't know it exists?' The experts were also asked not to forget the backdrop of powerful commercial services that give the appearance of being comprehensive and offer instant gratification to their users. Consensus was reached on the need to think beyond existing general-purpose search engines to tools for different kinds of search behaviours capable of finding the rich cultural heritage resources online. Future research challenges, as a key focus of the discussion, revealed agreement on the need to channel effort towards a modular approach in building user-driven search tools and services, able to integrate the learning and research behaviours of users.

This Thematic Issue also contains a case study, which concentrates on how to make use of Knowledge Organisation Systems (KOS) such as thesauri and ontologies in Web-based resource discovery. Douglas Tudhope and Ceri Binding highlight the fact that there is a vast existing legacy of such intellectual systems, and collections indexed by using their concepts and controlled vocabulary, to be found within cultural heritage institutions. This rich legacy of KOS makes it possible to offer more intelligent search options than the current generation of Web search engines. However, today such KOS are often only made available on the Web for display and reference purposes. There is a need to integrate them more fully in indexing and search systems, in particular through the development of standardised application programming interfaces (APIs) and access protocols.

Finally, regular readers of our Thematic Issues will find that this time we did not use images from a heritage institution to illustrate the publication. Rather, Birgit Retsch from the DigiCULT Forum secretariat shot a series of 'resource discovery' photographs at the Archivio di Stato di Roma, where the sixth expert forum was held.



RESOURCE DISCOVERY - POSITION PAPER: PUTTING THE USERS FIRST

By Guntram Geser

In recent years considerable technological advances have been made in the field of Internet-based access to heritage collections and other information. Research has focused on the development of descriptive concepts such as metadata for retrieval purposes, interoperable digital repositories, and more sophisticated searching and browsing mechanisms.

Overall, the expansion of the Web has favoured the emergence of new search applications, usage patterns and interaction paradigms. At the leading edge there are, for example, techniques such as interactive information retrieval, recommender systems, information extraction and summarisation, retrieval of multimedia features and other advanced applications for mining of specialised collections.

There are still many open research questions as well as unsolved issues of uptake and implementation.¹ This position paper will not try to give an overview of these issues. Rather, it concentrates on user-focused questions such as: Which users may want to discover what kinds of cultural and scientific heritage resources, and in what ways? How do user needs and expectations match up with available and potential resource discovery concepts and applications in the heritage sector?

FACING THE GOOGLE IMPACT

The OCLC *Environmental Scan* report, issued in January 2004, reviews global issues affecting the future of libraries as well as providers of museum and archival information resources.² This extremely interesting report ties its strategic assessments around the central perspective of service to information consumers at the level of their current needs. It acknowledges that the Web has become the most significant engine driving changes with respect to information access, and that there is among librarians and other 'traditional' information professions 'a subdued sense of having lost control of what used to be a tidy, well-defined universe'.

The report contrasts the library, characterised as a world of order and rationality, with the anarchy of the free-associating, unrestricted and disorderly Web.

In the latter, it states, 'searching is secondary to finding and the process by which things are found is unimportant. "Collections" are temporary and subjective where a blog entry may be as valuable to the individual as an "unpublished" paper as are six pages of a book made available by Amazon. The individual searches alone without expert help and, not knowing what is undiscovered, is satisfied.'

As Google becomes synonymous with the word 'search' in the minds of Web surfers, many librarians worry that people are losing sight of other information sources, online as well as within brick and mortar libraries.³ The OCLC report prominently cites a content vendor's statement that 'Google is disintermediating the library', and makes it clear that libraries need to work hard - and may even need to re-invent themselves - in order to convince the 'Google generation' that they can offer a better information service. This may mean to strive towards what the report, paradoxically, suggests as a worthy goal: to achieve 'invisibility' - in the sense that the library communities' services become ubiquitous and fully integrated into the infosphere.⁴

INCREASING THE VISIBILITY OF HIDDEN HERITAGE RESOURCES

While libraries, archives, museums and other heritage institutions may need to become 'invisible' as service providers within the electronic information environment, at the same time they need to raise the public awareness of their information resources more strongly.



¹ A good overview of the state of play in information retrieval research is given in: James Allen et al. (2003): *Challenges in Information Retrieval and Language Modeling*, <http://www.sigir.org/forum/S2003/ir-challenges2.pdf>

² Online Computer Library Center: *The 2003 OCLC Environmental Scan: Pattern Recognition*. C. De Rosa, L. Dempsey and A. Wilson. January 2004, <http://www.oclc.org/membership/escan/toc.htm>

³ Note that another information community is also hit by Google's innovative services: News editors. Cf. Larry Dignan's commentary on the virtues of news.google.com: 'Who needs editors anyway?' (1 October 2002), <http://news.com.com/2010-1071-960207.html?tag=nl>

⁴ It may also be necessary to use every opportunity to get closer to potential users and showcase library services, as, for example, the reference librarians at the Simon Fraser University do with their mobile Ask Us Here! service. See Diane Luckow: Help for Google generation (13 November 2003), http://www.sfu.ca/mediapr/sfu_news/archives_2003/sfunews1130306.htm

This includes awareness of the sheer fact that these resources exist, how rich they are, and that mostly they can be accessed free of charge. If the institutions do not succeed in this, to most people outside the heritage communities these resources might as well not be there.

Many people are unaware that most of the rich and authoritative information accessible over the Internet is invisible to general-use search engines like Google. This information resides on the 'hidden' Web, which is largely comprised of content-rich databases from universities, associations, businesses, government agencies and memory institutions. As a matter of fact, the information in these often publicly funded databases is not being used to its full potential; in particular, it could be opened up more effectively to non-specialist information seekers.

Paul Miller, Director of the UK Common Information Environment initiative, notes that Google and other powerful Internet search engines manage to retrieve only very little from the richly structured information in heritage databases, and asks: 'How can the creators and curators of high value, authoritative, information ensure that their work is found and used? More than that, though: how can we work to build associations and relationships between these aggregations of world-class content such that the user with an interest in "The Titanic", say, can move in a structured fashion from photographs and documents in The National Archives to objects in museums, to learning resources from Curriculum Online, to multiple media context from the BBC? How do we achieve all of this in a manner that is adaptable to the varying needs of our users, whether they are school children, undergraduates, teachers, researchers,

or members of the general public with an ongoing thirst for knowledge? And how do we maximise the value of the current and ongoing investment in creation, curation and delivery of this information and the range of services built around it?'⁵ These are many questions and challeng-



How rich is the 'deep Web'?

Although search engines may boast about their ability to index the Web, their searching spiders (robot programs that crawl the Web) can harvest only a small amount of information from the 'deep Web', which is the much larger portion of the Web. The search company Bright Planet states that this is 'approximately 500 times bigger than the surface Web'. A study conducted by Bright Planet in 2000 also found that the 'deep Web' contained nearly 550 billion individual documents compared with the one billion documents of the surface Web; that there existed more than 200,000 deep Web sites; that ninety-five per cent of the deep Web is publicly accessible information (i.e. not subject to fees or subscriptions), and that it represents the largest growing category of new information on the Internet.⁶ The main reason why this information remains hidden from searching spiders is that accessing it requires typing something and/or scanning a page and selecting a combination of options. And, well, the searching spiders not only lack fingers for typing, but also a brain capable of judgement.⁷

es, and Miller makes it clear that the related problems 'are bigger than any one organisation or sector, and that the best solutions will be collaborative and cross-cutting; that they will be common and shared'. A major step in addressing the challenges is initiatives such as the Common Information Environment (CIE) that serve as an umbrella under which a growing number of organisations are working towards a shared understanding and shared solutions.⁸

DISCLOSURE OF HERITAGE RESOURCES

Today, heritage institutions invest much effort in providing online access to their descriptive metadata on collection objects and, increasingly, to digital surrogates of these objects. To support resource discovery, they need to expose metadata about their resources, so that it can be used by other applications and services, such as harvesters, distributed search mechanisms or alerting services. In order to facilitate exchange and interoperability between services, the institutions will need to provide item-level metadata conforming to a widely shared standard such as Dublin Core⁹ or, if the target is to serve learning and teaching communities, Learning Object Meta-

⁵ Paul Miller: Towards the Digital Aquifer: introducing the Common Information Environment In: *Ariadne*, Issue 39, April 2004, <http://www.ariadne.ac.uk/issue39/miller/intro.html>; one working prototype designed to demonstrate the CIE concept to users, content providers and policy makers is My Historic Environment described by Kate Fernie et al. at <http://www.mda.org.uk/conference2003/paper08.htm>

⁶ Michael K. Bergman: The Deep Web: Surfacing Hidden Value (August 2001). In: *The Journal of Electronic Publishing*, Vol 7, Issue 1, <http://www.press.umich.edu/jep/07-01/bergman.html>; also accessible at: <http://www.brightplanet.com/technology/deepweb.asp>

⁷ Cf. the University of California's tutorial 'Finding Information on the Internet', which gives a detailed explanation of why some resources remain invisible to spiders: <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html>

⁸ Organisations currently working to shape the CIE include Becta, The British Library, Culture Online, the Department for Education & Skills' (DfES) e-Learning Strategy Unit, the e-Science Core Programme, JISC, the Museums, Libraries and Archives Council (MLA, formerly Resource), The National Archives, the NHS' National Electronic Library for Health (NeLH), the Office of the e-Envoy and UKOLN.

⁹ Dublin Core Metadata Element Set, <http://dublincore.org/documents/dces/>



data (LOM).¹⁰ One major success in the enhancement of heritage resource discovery is the strong uptake of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in the sector.¹¹ This includes subject gateways¹² or resource discovery networks such as the UK RDN;¹³ large players such as the Library of Congress (e.g. for the American Memory Collection),¹⁴ as well as highly specialised communities such as institutions and scholars working on Renaissance emblem literature.¹⁵ Also Minerva Europe, the Europe-wide co-ordination initiative in the area of heritage digitisation, shows a clear interest in the OAI-PMH for digital resource discovery. It may well become a basis for a future European cultural heritage portal.¹⁶

A major reason for the success of OAI-PMH rests on its mandatory minimal requirement of unqualified Dublin Core (DC). However, this also leads to a downside. DC has been designed to support simple cross-domain resource discovery. When rich metadata is cross-walked to DC, the hierarchical structure of the original encoding is lost, i.e. its full richness cannot be expressed. But, overall, OAI-PMH has shown itself to be an efficient and effective way of metadata exposure and exchange.

RESOURCE DISCOVERY: WHAT MAKES IT A COMPLEX PROCESS?

Resource discovery involves the searching, locating and retrieving of information resources on computer-based networks, in response to queries of a human user or an automated mechanism. It also involves presenting the information to the user in the most appropriate form, and the capability to manage the found resources at either the find or the retrieve level. Therefore, the discovery process should support a query and browsing interface as well as other methods of user interaction (e.g. for personalisation).

The resource discovery requirement may be transient or may be based on a more permanent system that regularly notifies the user about relevant information. Advanced systems will utilise profile infor-

mation on the user in order to provide the most relevant information efficiently. However, ideally a system should also allow for serendipity, i.e. discovering valuable resources the users were not looking for intentionally.¹⁷

Interactive resource discovery is a complex process that includes multiple phases, is iterative and highly dynamic. In particular, it is an intellectual process, which involves (re-)formulating queries and examining candidate resources. It is also a learning process, in which users will successively develop and refine criteria of information granularity as well as their understanding of domain specific terminology, concepts and discourse. In this respect, the resource seeker can be compared to a traveller, who in communicating with local residents may use a phrase book, but if he does not learn their language he will not be capable of 'navigating' properly within their culture.¹⁸

IT'S THE USER, STUPID!

Most of the time we are all fast-moving electronic travellers who use simple phrase books, and want useful information as quickly as possible. This is why almost nobody – less than 0.5 per cent – uses advanced search options; why most people want to have a simple entry field into which they will type on average 1.3 words; and why most people will not look at more than one page of result list.

In consequence, general search providers will need to concentrate most of their energy on doing as good a job as possible with little input from the users, hiding compound queries behind point-and-click, and fitting more information on to search-result screens.¹⁹

In consequence also, nobody should expect heritage organisations to compete (in any sense of the word) with powerful search engines such as Google. Let the problem of large numbers of hits, information sources of doubtful relevance, and no guarantee of authenticity and reliability be the problem of the fast-moving e-traveller.

The appreciation of highly valuable structured heritage information resources as well as elaborated concepts and different perspectives of the Arts &

¹⁰ IEEE Learning Object Metadata (LOM), <http://ltsc.ieee.org/wg12/>

¹¹ OAI, <http://www.openarchives.org/>; for an overview on the use of the OAI-PMH in the heritage sector see Muriel Foulonneau (ed.): *Open Archives Initiative – Protocol For Metadata Harvesting. Practices of cultural heritage actors*, September 2003, http://www.oaforum.org/otherfiles/oaf_d48_cser3_foulonneau.pdf

¹² The DigiCULT Website's 'Resources' section also features Gateways to Cultural Heritage Resources, currently over 70 annotated links: <http://www.digicult.info/pages/resources.php>

¹³ The UK RDN is a collaboration of over 70 educational and research organisations, including the Natural History Museum and the British Library, <http://www.rdn.ac.uk>

¹⁴ The Library of Congress's experience as an early adopter of the OAI-PMH is described by C. R. Arms (2003): *Available and Useful: OAI at the Library of Congress*, <http://memory.loc.gov/8081/ammem/techdocs/libht2003.html>

¹⁵ See the DigiCULT special publication: *Digital Collections and the Management of Knowledge: Renaissance Emblem Literature as a Case Study for the Digitization of Rare Texts and Images*. February 2004, <http://www.digicult.info/pages/special.php>

¹⁶ Cf. Foulonneau 2003 (footnote 11), pp. 43–44. Footnotes 17–19 continued on page 10.

Humanities must stem from other sources, in particular, from inspiring teachers on all levels and the educational efforts of the heritage institutions themselves.

SERVING COMMITTED ELECTRONIC TRAVELLERS AND LEARNERS

Heritage institutions are working hard to create and expose metadata about the objects they hold in their often heterogeneous, multi-media collections (e.g. records, images, sound, video, etc.), and increasingly they create and provide access to repositories of digitised items. Heritage objects are also strongly related to different historical and cultural views and interpretations, whereas in other information domains terms and concepts may be taken as given and retrieved out of documents without much consideration of contexts. Consequently, the heritage sector is rich in knowledge organisation systems such as controlled vocabularies to assure semantic consistency and interoperability.

However, these huge efforts often seem to fail to result in an adequate return on investment – in terms of interest and appreciation, discovery and valuable uses of heritage resources and services. Is there a mismatch between these resources and services and the needs and expectations of broader user communities? If so, how large is the mismatch with respect to available and potential searching, browsing, navigation and other concepts and mechanisms?

In order to sort out what kinds of needs the heritage sector should be able to serve, in the overview below we concentrate on the two largest user groups, ‘academic’ and ‘educational’ users. Furthermore, the different user groups’ objectives and tasks, current modes of resource discovery, and potential advanced and future options are described.

Academic Users

Overall, academic user groups – scholars, university teachers and students – will seek highly structured and authoritative information resources and use resource discovery tools for domain experts. However, they will also welcome enhancements in tools that both inform and direct, including more intuitive browsing and personalisation mechanisms.

Main objectives and tasks

Scholars, university teachers and students research on, locate, identify and compare specific heritage objects with like examples. In order to validate a hypothesis in Arts & Humanities studies (e.g. history

of culture and science), they contextualise and interpret the objects, and prepare publications, lectures and seminars that consolidate, expand and mediate knowledge.

Current modes of resource discovery

There already exists a wealth of resource discovery services and tools for academic user groups such as online bibliographic information services, subject gateways or full-blown resource discovery networks such as the UK RDN.²⁰ With respect to online heritage databases, academic user groups will expect and use:

- | Finding aids for special collections and archives;
- | Multiple entries to collection information, e.g. search by subjects, object types, function, names, dates, etc.;
- | Advanced search, e.g. for devising combinations to mine collections, ideally across distributed, cross-domain databases;
- | Opportunities to use thesauri that display vocabulary terms in hierarchies and tree segments;
- | Mechanisms that suggest search terms, phonetic spellings, correction of typing errors;
- | Zoomable images, links to sources, a glossary function, and copyright information.

Advanced and future options

When browsing the topics of interest of, and contributions to, conferences in the area of information retrieval research (for example, ECIR or ACM SIGIR²¹), anyone will come to at least three conclusions: the area is extremely broad; there is much work in progress; and many topics will find a strong interest of academic user groups, such as topic detection and tracking, information extraction, text summarisation, collaborative filtering and recommender systems, natural language processing, cross-lingual and multilingual issues, image, audio and video retrieval.

However, the picture changes somewhat if we ask where scholars, university teachers and students will find convenience and productivity gains with respect to their main objectives and tasks (as described above). What we would expect to be of prime interest in resource discovery will then, for example, include:

- | A stronger integration of resource discovery within the ‘workbench’ or ‘toolbox’ of scholars and students, such as networked thesauri²² and reference management tools.
- | Personalisation and sharing of discovery strategies and results: For example, Historyguide.de offers to store favourite searches and automatically

¹⁷ For a more detailed description, see ‘Resource Discovery - A Definition’, Research Data Network CRC, Resource Discovery Unit, University of Queensland, <http://archive.dstc.edu.au/RDU/RD-Defn/>

¹⁸ Cf. Carl Lagoze: From Static to Dynamic Surrogates. Resource Discovery in the Digital Age. *D-Lib Magazine*, June 1997, <http://www.dlib.org/dlib/june97/06lagoze.html>

¹⁹ Cf. Tim Bray: On Search, the Users (17-06-2003), <http://www.tbray.org/ongoing/When/200x/2003/06/17/SearchUsers>

²⁰ UK RDN, <http://www.rdn.ac.uk>; among the eight subject gateways or ‘hubs’ of the RDN are a Humanities (HUMBUL) and an Arts & Creative Industries (artefact) hub.

²¹ ECIR04, <http://ecir04.sunderland.ac.uk>; ACM SIGIR 2004, <http://www.sigir.org/sigir2004/>

²² See the article by D. Tudhope and C. Binding in this issue.



push announcements via e-mail when fitting new metadata records appear on their gateway. In January 2004, it also introduced the opportunity to integrate regularly updated search results into one's own Website.²³

- | Systems that involve knowledge-based visualisation and contextualisation of resources, such as the one that is being built by the VICODI project, which concentrates on European history resources.²⁴
- | Generally, academic users, in particular university students, will also welcome resource discovery applications that provide state-of-the-art graphically driven interfaces for searching, browsing and navigation (see also below).

²³ History Guide/

InformationsWeiser

Geschichte: <http://www.historyguide.de>

²⁴ VICODI (Visual

Contextualisation of Digital

Content), <http://www.vicodi.org>; for a concise

overview see their

poster for the IADIS

e-Society 2004 conference,

http://www.vicodi.org/vicodi%20poster_143.ppt

²⁵ Cf. EP2010 Study:

Dossier on Digital Games

& Learning – Paradigms,

Markets and Technologies.

September 2003, http://ep2010.salzburgresearch.at/dossiers/ep2010_dossier_games-elearning.pdf

²⁶ VMC: <http://www.virtualmuseum.ca>

²⁷ Educnet: Sites disciplinaires et thématiques,

<http://www.educnet.education.fr/secondaire/disciplines.htm>

²⁸ SCRAN: <http://www.scran.ac.uk>

²⁹ Learning Curve, <http://learningcurve.pro.gov.uk>

³⁰ Cf. Fiona Cameron

(2003): 'The Next

Generation – 'Knowledge

Environments' and Digital

Collections, <http://www.archimuse.com/mw2003/papers/cameron/cameron.html>

Educational Users and Lifelong Learners

Educators in heritage institutions should play a strong role in gaining an understanding and monitoring the current changes of user expectations in the digital environment. Schoolchildren, students and a younger generation of teachers have grown up with rich, interactive and dynamic media. The level of what they expect from learning content, environments and tools in terms of interactivity, complexity, active exploration, as well as collaboration, is increasing rapidly. In fact, heritage institutions need to understand that their customers benchmark their online educational resources against the industry standards in 'edutainment', games, and other compelling interactive environments.²⁵

Main objectives and tasks

From the traditional viewpoint of instructional and textbook-based teaching, one will expect teachers to search for ready-to-use content and even worksheets for preparing lessons, student projects, and school visits to museums or heritage sites.

However, a newer generation of teachers – in line with state-of-the-art learning paradigms – will seek environments that engage learners in meaningful practices, fostering creative thinking and innovative problem solving. Schoolchildren and students themselves will be most interested in explorative e-learning opportunities.

Current modes of resource discovery

School teachers and students will explore targeted Websites such as the Virtual Museum Canada that offers a wealth of exhibits, a teacher centre, searching across all resources or all exhibits, and a MyMuseum function.²⁶ They will also use thematic Websites such as those on the French Educnet that include resources on art history, cinema, music and theatre.²⁷ Furthermore, there is a growing number of Websites that provide access to resources that are tied in with national curricula. Best practice examples here are, for example, SCRAN – the Scottish Cultural Resources Access Network²⁸ – or The National Archives' 'Learning Curve'.²⁹ However, heritage institutions today mainly provide opportunities to 'mine' digital collections such as image libraries. The main functionalities are searching and browsing of resources, a 'one size fits all' approach from school kids to 50+ lifelong learners.

Advanced and future options

Heritage institutions need to serve different search styles that are related to a range of information, learning and leisure activities. Many older users will tend to prefer simple, user-friendly searching and browsing tools, keyword searches or searches under known categories, and to follow linear browsing pathways (e.g. using hierarchical metaphors and thematic structures). Future search opportunities for these users may include natural language question and answer mechanisms and recommender systems (similar to features offered, for example, by Amazon).

In contrast, younger users prefer browsing options over specific searches, want to drive their own pathways through collections, and seek interactivity and action spaces (e.g. real time, 3D presentation, simulations, game-like discovery spaces).³⁰ In particular, youngsters will welcome resource discovery applications with state-of-the-art graphically driven interfaces for browsing and navigation, associative concepts such as mindmaps, visual prompts, and visualisation engines that dynamically connect related topics, ideas and objects. Such solutions will most likely be used on top of large online collections and within exhibition environments.

Some examples include:

- | The AquaBrowser technology developed by Medialab Solutions,³¹ which is being used by the Dutch Digital Heritage Association's Cultuurwijzer (Net Guide to Culture) Website for its 'freestyle surfing' Cultuurgrazer search function.³² The technology dynamically generates clouds of associations (e.g. Rembrandt – museum – paint) and result lists. A result list contains the documents that are closest to both the user's query and the cloud of suggested associations. The latter can be used to explore and find more relevant resources along different routes.
- | Automatically generated thematic 'trails' through collections such as those offered by Picture Australia³³, or theme generator tools, such as the HyperMuseum prototype.³⁴
- | Dynamic generation of timelines, chronologies, maps, and other contextualisation: e.g. the Metropolitan Museum of Art's 'Timeline of Art History'³⁵; or the Electronic Cultural Atlas Initiative (ECAI) projects, which use TimeMap software developed at the University of Sydney.³⁶ Ideally, such applications would allow for demonstrating relationships between events or objects, and accessing deeper layers of information.



SUMMARY

We started off with the OCLC *Environmental Scan* report's picture of a dreary future in which Google and other powerful search engines would 'disintermediate' heritage institutions in their role as high quality and authoritative information service providers. We found that the institutions should not fear and try to compete with such widely used search engines. Rather, they need to raise the public awareness and visibility of their resources more strongly, and work hard to support their academic and educational user groups, at the level of their expectations of information access and use. Otherwise the institutions' investment in creating digital collections, rich descriptive metadata, learning resources, and a broad range of information services will not lead to a high return on investment – in terms of interest and appreciation, discovery and valuable uses of heritage resources.

Overall, heritage institutions and networks have already achieved a lot in terms of Web-based disclosure and discovery of their information resources. However, as stated in the *DigiCULT Report*, 'discovery is the beginning, not the end'.³⁷ The institutions and networks will need to invest an extra effort in state-of-the-art environments, workbenches and tools for their core user groups, to allow for valuable uses of the discovered information resources by scholars, university and school teachers, schoolchildren and students, and interested lifelong learners.

³¹ Medialab Solutions: Overview of AquaBrowser technology, <http://www.medialab.nl/index.asp?page=about/technology>

³² See: <http://www.cultuurwijzer.nl;http://den.medialab.nl> and <http://erfgoed.medialab.nl>. Note that the latter Website allows for AquaBrowser, thesaurus-based as well as keyword-based searching.

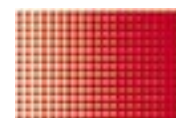
³³ Picture Australia: <http://www.pictureaustralia.org/trails.html>

³⁴ Stuer, P., Meersman, R. and De Bruyne, S.: The HyperMuseum Theme Generator System: Ontology-based Internet Support for the Active Use of Digital Museum Data for Teaching and Presentation. In: *Museums and the Web 2001*, <http://www.archimuse.com/mw2001/papers/stuer/stuer.html>

³⁵ <http://www.metmuseum.org/toah/splash.htm>

³⁶ ECAI: <http://www.ecai.org>; TimeMap project: http://acl.arts.usyd.edu.au/research/time_map/

³⁷ Cf. European Commission: *The DigiCULT Report. Technological landscapes for tomorrow's cultural economy*. Full Report. January 2002 (chapter IX.6 'New tools in the box'), available for download at <http://www.digicult.info/pages/report.php>





EMPOWERING THE USER THROUGH CONTEXT AWARE SEARCH TOOLS

AN INTERVIEW WITH JUSSI KARLGREN

By Joost van Kasteren

It seems simple: you type a few words and at the blink of an eye thousands of documents containing a wealth of information appear at your fingertips. 'It seems like progress', says Jussi Karlgren from the Swedish Institute of Computer Science*, 'but, in fact, a lot of embedded help has disappeared. In the offline world, a publisher packages the content along perceived user needs. And, a good librarian or bookseller will create easy-to-use search trails. In my view, special tools are needed to help people survive the digital information flood.'

Karlgren is an expert in language technology and works with real users in an experimental setting. From his observations, the existing general tools for information retrieval generate a whole lot of bits and pieces that are very difficult to assess for the user, unless you are a specialist in the field. 'If I am a user out on a search for information on drums used by Shamans in Lapland, I do not only want a picture of the drums, but I also want to know what they are used for, what role they play in the Saami culture, and what the existing views of anthropologists are. With the tools for retrieval I have to figure it out all by myself. There is no go-between for digital information sources. The user has to do the work the publisher or the editor of a magazine used to do, i.e. judging the relevance of the information, arranging it in a logical order and interpreting it. It seems like progress but in the meantime the user has lost a lot of helping hands in trying to make sense of information.'

'The problem is that we are looking too hard for general tools to retrieve information that can be used by any user from now until eternity. A kind of Holy Grail, which is not very useful, because the use of information changes over time. Just look at newspapers from fifty or even twenty years ago. They are very different from today's papers, because they reflect changes in the publisher's anticipation of users' needs over time. We have to realise and to accept that the retrieval systems of today will be obsolete in five or ten years' time.'

Karlgren predicts that, as the amount of information is growing, there is a need for more specialised case

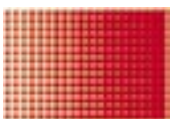
tools to support the user in finding the information he needs. A kind of information brokerage, either fully automatic or partly manual, that is comparable with the role publishers, editors and writers used to play and still play. There are already brokers on the Internet, for instance for booking hotels, but they are not fully geared towards users' needs.

Karlgren: 'Partly to protect their own interests they do not give you the phone number of the hotel. While you might need it because you have a baby with you and want to know if they can cater for him as well. What you actually want as a user is a broker who can point you in the right direction and gives you information you can use. If you are going to buy a vacuum cleaner you do not need the one-sided information of the seller, nor a list of all vacuum cleaners in the world, but you need someone to help you to express your needs and find the right information to fulfil that need.'

The need for tools tailored towards users' needs is very good news for cultural heritage institutions, according to Karlgren. 'Cultural heritage institutions are a bit wary about giving out information. Not so much because they want to make money out of it, but because of their reputation. They want the people to know that they provided the information, but they do not want other people to mess around with that information, thus – indirectly – giving them a bad name or without giving them the credit they deserve. If you have tailored tools for accessing the information, you can keep track of its usage.'

So, if during a search on the Internet someone stumbles on to information on Shaman drums from the north of Finland, it should be clear right away that the information is coming from, let's say, the Nordic Museum of Stockholm and that it is linked to information on the role of these drums in the Samen culture, the existing views on the subject and so forth. Or, to put it another way, you immediately present yourself as a trustworthy guide or broker to help the user fulfil his information needs.'

* <http://www.sics.se/~jussi/>



DIGICULT EXPERTS SEEK OUT DISCOVERY TECHNOLOGIES FOR CULTURAL HERITAGE

By Michael Steemson

For almost two years, DigiCULT Forums have been disentangling the World Wide Web, drilling through the reactionary rock strata of resistance to change and lighting up the shades of virtual communities, all in the cause of quickening cultural heritage sector interest in the new, exciting information society technologies.

Echelons of DigiCULT experts debating in places like Barcelona, Darmstadt, The Hague and Edinburgh have urged mobilisation of digital development forces inside museums, libraries, archives and galleries. It has been a hard campaign.

And so, for a change, the Sixth Forum, meeting in Rome, looked at the sector from the outside. It asked, on behalf of cultural heritage clients everywhere: 'How can we find something if we don't know it exists.' It was the perfect question to address in the 2,700-year-old capital of Italy where, if Vespa volleys and Fiat Uno flocks can be evaded, unsuspected heritage is to be found down every cellar or *strada con diritto di precedenza*.

What the DigiCULT Forum 6 academics, scientists and archivists sought were suitable resource discovery technologies ... digital processes to search cultural heritage (CH) institution cellars and *stradas* to offer professors and populace a whole-of-sector view of their subjects. It was a lot to ask of the Rome Eleven, ten men and one woman, as they gathered under the painted ceiling of the book-lined *Sala Alessandrina*, the virtual reading room of the *Imago II* digital imaging project at the Rome State Archives (*Archivio di Stato di Roma*) in the city's suitably named *Palazzo della Sapienza* (Palace of Wisdom).

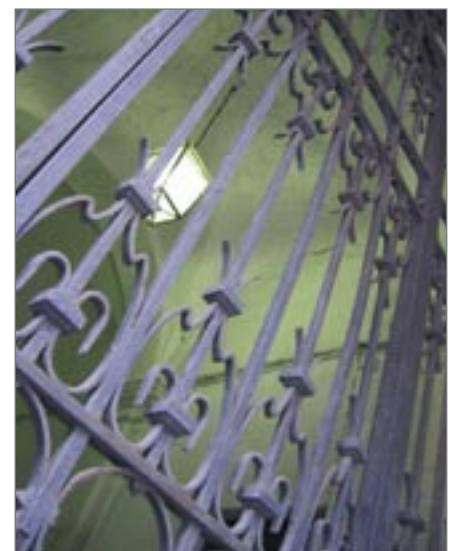
The search continues a series of round table debates for the European Commission's Information Society Technologies Directorate, acquainting cultural institutions with the wonders of information technology. Earlier forums had studied such subjects as the Semantic Web, digital learning objects and digital asset management.

Forum 6 launched into its work inspired by an erudite exposition on the enormity of the task from Swedish digital library scientist, **Traugott Koch**, the senior librarian for the Lund University libraries, earnest in patrician beard. In his dissertation, this *Journal of Digital Information* (JoDI) editor walked the experts through a list of concepts and approaches to resource discovery technologies. He had been doing a 'little research', he said, modestly (actually he has been at it for the best part of fourteen years), in 'automatic classification and resource discovery re-using traditional knowledge organisation systems, most recently with the EU's Renardus Consortium¹ service trying to provide cross-searching, cross-browsing in highly heterogeneous and distributed resources.'

This was certainly the stuff the Forum had assembled to discuss. The participants listened intently as the Swedish expert continued: the basic difficulty was that the objects to be sought were not self-describing and were isolated within large collections. Then, there were the sector's differing cultural and historical views of terms and concepts, its need for greater awareness of catalogue and retrieval metadata schemas, multi-media, multi-linguality, Web authoring tools, mark-up languages and so on.

Mr Koch expected the Forum would discover that 'there are not enough digitised catalogues of finding aids or whatever to speak about objects and documents' and would find a need for more sector co-operation and co-ordination in developing its digitisation standards. In the wider world, electronic access infrastructures – registries of vocabulary, semi-automatic metadata creation tools, knowledge management systems, authority databases and dynamic,

¹ Renardus: Internet search-browse access to participating European data provider records; <http://www.renardus.org>



adaptive interfaces – were developing fast but, he said, the heritage sector was not leading its own, unique researches, just following.

Forum Moderator, **Joemon M. Jose**, a multimedia information retrieval expert with the Computing Science Department at the University of Glasgow, Scotland, thought the experts' approach should be from three viewpoints: e-learning, 'virtual museums' and 'presenter', which he described as 'a virtual artist who wants to create a multimedia presentation or a documentary creator who wants to identify material for his project'. He went on: 'So the first question is: which available and potential resource discovery mechanisms are most relevant for heritage institutions?'

EVALUATING THE FOUND OBJECTS

But, Swedish Institute of Computer Science researcher, **Jussi Karlgren**, put his finger on a concern that needed solution before they did that: in other words, identifying and verifying the sources of information gathered by these mechanisms. All three of the Moderator's scenarios could be challenged to evaluate material assembled and presented by an automated system, he suggested.

'If, for example, I wanted an overview of Finnish history and found a Russian museum, a Finnish museum and a Swedish museum to give it, the three pictures would be completely different and so contradictory as to challenge the person receiving them, even cause distress. So how do you provide these intellectual challenges in ways that are obvious and yet get answers you can believe.'

Worse could happen, said Dr Karlgren. 'I can imagine scenarios where a highly prestigious institution would say "I will not be a part of it if that institution is part of it and if our information is mixed", which is completely likely.'

His countryman, Lund University librarian Koch, had an answer, though not an immediate one: 'This calls for provenance metadata to a much larger degree than we find today. There are lots of text and documents on the Web that you won't be able to identify, sometimes not even the organisation from which they emanated. You would not be able to identify who is the intellectual produc-

er or owner of the information, statement, opinion, or text passage.'

He went on in words aimed at institution curators: 'Because of these different views, there are much higher requirements for even small chunks of text passages and other stuff to be identified with the creator, the creator's affiliation and, maybe, point of view. That could not be neutral, of course, but it would later on allow text criticism and analysis of prevailing cultural and political views.'

Moderator Jose wondered if information conflicts might be detected by the search system, an idea that appealed to Dr Karlgren. He teased: 'Detecting conflicts automatically sounds like a really wonderful research project I would like to work on for the next ten years or so.'

However, he thought that users would be aware that conflicts were likely even if the information was authentic, authorised and reputable. He suggested: 'It is quite likely that Vikings are viewed differently in the north of France than they are in Scandinavia, for instance.'

He went on: 'Objective judgement is difficult on the exploration of northern Europe by seafarers or, alternatively, the rape and pillage of the northern seas by barbarians, unless you know that this information is from this source and this is from that; people trust this information source; this bit of information is used in many places and those in turn are reputable places.'

The way to deal with the problem was the use of provenance to identify sources and reveal how other users viewed their reputations. Then it would be a question of trust, in much the same way as one chose a newspaper. But, he wondered, how would such metadata evolve?

Dr **Andreas Rauber** had a solution. He is an associate professor in the Department of Software Technology and Interactive Systems at the Vienna University of Technology in Austria, a student of machine-learning techniques to make collections more accessible. Instead of assigning metadata to objects, the objects should be associated with information, he proposed.

He explained: 'Put the object into a kind of ... well, I wouldn't like to call it an ontology because that's way too structured, but put it into a kind of information space. Just as you have information about knowledge, so you have information about certain periods. Put the object into that and then describe the situation, the relationships. As soon as you add that description it becomes available to all the objects that are placed in that period.' 'You have different views of certain events then you place the object into them.'



That way the object is linked to the different interpretations rather than having the different interpretations assigned to the object.'

The procedure struck a chord with **Douglas Tudhope**, a researcher in hypermedia in the School of Computing at the University of Glamorgan, Wales, working with processes like interactive and automatic query expansion. What he called the 'raw object and its metadata' would come from the museum collection, he thought, 'But then, it is the act of creating this secondary information object which has got the context and some perspective. People could add to it and create layers. More attention could be given to the tool to create this and to what metadata is needed for this secondary information object.'

The conversation seemed to be getting a bit too lofty for the likes of the Forum's sole woman expert, Denmark's Dr **Pia Borlund**, an associate professor in the Information Studies Department at the Royal School of Library and Information Science in Aalborg. She wanted it brought down to earth again, suggesting: 'One thing we are also considering is whether we want to go for ideal or what is practical and possible.' Her field of expertise is evaluation of information retrieval and interactive systems including, as she put it, 'users and their use of information retrieval systems leading to information-seeking behaviour'.

But Moderator Jose had, apparently, rather liked the turn of the debate and suggested: 'Curators want what the current technology allows them but we also need to look beyond what the technology allows.'

Seamus Ross liked Dr Rauber's 'information space' process, too. Dr Ross is Director of the University of Glasgow's Humanities Advanced Technology and Information Institute (HATII) and the EU's Electronic Resource Preservation Network (ERPANET). He recalled undertaking a study in the British Museum for which nineteenth-century records had included drawings and a curator's notes but little provenance. There had been nothing with which to relate the object to material found in a similar context and it had been very difficult to come to an understanding of it. However, things improved, Dr Ross remembered.

'Over time, by relating that particular object to others we are able to build up a metadata picture of the kind of contexts that it had to come from. We attached the object to a whole set of metadata that we built up about other objects of similar categories and similar types. We need to recognise that metadata evolve and in fact our ability to trust in their relevance changes over time.'

KNOTTY METADATA STANDARDS

Then the discussion turned to questioning a requirement for standard metadata thesauri. Italian **Paolo Buonora** bemoaned the fact that 'in Italy, the only point that common metadata committees agree is administrative and management metadata. We can not arrive at any common point about a description of a cultural object'. Mr Buonora is director of the Rome State Archive's *Image II* digital imaging project in whose reading room the Rome Eleven cogitated.

He was not, however, sure that common metadata mattered. As he said: 'The reasons why historians and researchers in general love to come to the archives is that they never find exactly what they were looking for. All we need is a map so we do not get lost.'

Austrian Dr Rauber also wondered if the idea of a common metadata model was even attainable let alone necessary. Could existing technologies not adequately link differing representations of a single object, he wondered, adding: 'I am still pretty sure that a lot of the tasks that we want to solve can be done with state-of-the-art technology and without investing great sums. So a lot can be solved with very crude automatic approaches to a very high degree of user satisfaction.'

He had a formula for the thesis: 'My analysis leaves me to ask whether they would not take us faster towards quite reasonable results that may serve 80 per cent of our needs, and leave the remaining 20 per cent for discussions that we will probably never be able to resolve anyway.'

Dr Karlgren was at odds with his 80 per cent figure, saying glumly: 'I would say five per cent can be structured with the technology we have today.' But he readily acknowledged: 'This is a rapidly evolving field, finding contexts in texts, for instance, or recognising objects in pictures or events in video frames. This is something a lot of people are working on.'

'Of course, most are working with national security concerns or something, not cultural heritage. But this stuff will be percolating into our fields, so we can be certain that in 15 years from now there will be fairly good video parsing sequences that are able to extract information from video frames quite well. And, we can use them to structure the information. All we need to do now is to say that those things will be available and now we do not need to draw a complete metadata set.' Here, Traugott Koch preached caution, warning: 'We should be careful with our present discussion not to be drawn into conflicts about the view on metadata.'



He agreed that metadata need not be applied at the moment of creation, but it was important to be able to show that, to take a simple example, a street name or house number had been valid at a certain time. The International Council of Museums' common extensible semantic framework, CIDOC CRM,² showed there was much commonality in the concepts museums used as basic categories. These were constant things that would not be re-invented in 100 years.

'I would argue that behind the metadata there must be a controllable, stable ontology of categories, not what kind of data you slot in or what connections you make at a given moment.'

Dr Ross had a cheerful example of the process being used by London's Tate Gallery when digitising 50,000 art works. He explained: 'They had the metadata as you describe, they knew what it was, where it came from, when it was made. They actually took that and linked their catalogue to the digitisation of the visual object. But then they added subject metadata to every one of those visual objects – between five and 40 pieces of information.'

It had transformed how users accessed the collection, he told the Forum. 'Before they used to go and search for artist and period, now they search by topic, by subject. Unfortunately, I have to say that the most common topic they searched for was "nude" but, still, you would have noted that it still changed how they did it.'

Glamorgan University's hypermedia researcher Douglas Tudhope had a last word on the subject before Moderator Jose dragged the Rome Eleven back onto enabling technologies themes. 'There is no dichotomy between using metadata and language-based approaches or content-based analysis. We can do it both ways. Museum collection management systems have the building blocks to create metadata and many of them conform to controlled terminologies. Their associated text descriptions can be analysed by

information retrieval techniques and from these we can link to the wider worlds. It is a matter of using a combination of techniques.'

SEEKING ENABLING TECHNOLOGIES

Traugott Koch offered Google as an enabling 'candidate technology' saying: 'A full text search in Google will solve most of our technological problems, and just in time.' But how would users find their answers in a digital archive of TV documentaries, for instance? 'What technologies are needed to make this possible, that a teacher in Stockholm in five years' time can really search through all the digital collections of TV features?'

Douglas Tudhope wanted 'tools for different kinds of search behaviours'. The need was for search metadata to facilitate browsing and what he described as 'serendipitous discovery' for users.

The Eleven talked of time-line links, Dublin Core values, workflow management, provenance and flexible access systems. Seamus Ross liked this last point. It could help in many museums where heritage archivists shied away from grouping objects outside the 'anachronistic form' of their physical collections. 'You would keep objects that were found together or in the context of one another but now you can use new technology to re-group objects of a similar type.'

The Eleven thought there was need for a study of online usage and ways of letting owners know when their objects were being used. They talked of official directories, the University of California's Alexandria Digital Library³ collections of geographically referenced materials and access services, and the drawings of the 'father of modern archaeology', the eighteenth-century German art historian, Johann Winckelmann.⁴

Dr Jose sought debate on information retrieval technologies but Swedish computer scientist Dr Karlgren challenged: 'We need to make a distinction between the technologies that will develop irrespec-

² International Committee for Documentation of the International Council of Museums (ICOM-CIDOC): CIDOC Conceptual Reference Model, <http://cidoc.ics.forth.gr>

³ Alexandria Digital Library Project, <http://www.alexandria.ucsb.edu>

⁴ Johann Joachim Winckelmann, see Minnesota State University e-museum reference: http://www.mnsu.edu/emuseum/information/biography/uvwxyz/winckelmann_johann.html

tive of what we say and those that we in this room need to request from the developers.'

Talk of better text indexing or image understanding technology by the Rome Eleven would make little difference because 'so many people are already clamouring for them', he said. Needs specific to digital cultural heritage had to do with veracity and 'all those things that make museums, archives and libraries more interesting than bookshops'.

He stressed: 'We urgently need a technology for better understanding of similarities between documents in terms of usage, for example. Maybe if we say this loudly enough here it will actually make a difference because that is not currently a big issue in the research field.'

The Google search process – 'typing 1.87 words into a window, looking at 60,000 documents in a Google window, then at 12 of them and then you read one of them' – was one way, he said, but the process did not look into layers of information to cover what Pia Borlund had earlier called 'the perceived difference between what I know and what I need to know to do some task'.

The Forum set about building its own model of the required process, a system that, as Traugott Koch described, 'invites you to start at some point in a hierarchy, trickling down, exploring up and down and wandering through the hierarchies or networks'.

The experts knew that users were often unsure what they wanted, but knew what it was when they saw it. Search engines had to be able to notice how users acted during searches and try to detect what type of search was being undertaken: specific information, a home page, entertainment, or just ideas. The engines, therefore, had to be aware that objects were similar in many different ways in different information spaces.

Sometimes information was just stumbled upon. Sometimes it stood out because of the way it had been used previously. Connecting these options could create a powerful way of modelling information paths without taking them beforehand.

Then put 'attitude' into the mix, a connection between objects of critical importance, and add 'trust'. Dr Karlgren commented wistfully: 'Allowing this sort of thing to emerge dynamically is important ... to make a system that can extract this sort of information and make it usable later. I think that is the technological support I would be most happy to build.'

In fact, some systems were already making analyses on some of these different layers, he thought, but nonetheless 'focusing on the particular needs with-

in the cultural heritage sector might prove beneficial'. The Moderator brought them back to practicalities again, asking: 'So what kind of technology allows us to build this? Are there any technologies that allow us these layers of information, and enrich as we interact with them?'

Glamorgan University's Douglas Tudhope had good news: 'In the NKOS [Networked Knowledge Organization Systems] network,⁵ we are currently looking at standards for protocols for using knowledge organisation systems, classification thesauri for different kinds of use which could be search or browsing. I believe in the next few years there will be some sort of standard protocols that are going to be the building blocks to enable these different approaches.'

Seamus Ross could see value in this. 'Participation in the use of an information resource increases its value and usefulness over time,' he contended.

Lund digital librarian Traugott Koch thought museums could learn something from the library Web services described by **Lorcan Dempsey**,⁶ Vice-President and Chief Strategist of the Online Computer Library Center⁷ (OCLC) in Dublin, Ohio, US. He had devised a library service made up of thousands of Web services providing 'all the smart atomic bits and pieces and they can be built into learning object repositories or whatever scenario you might want to use'.

Traugott Koch said: 'I would ask museums to be active in the creation of such services themselves with things they have to do anyhow, to decompose some of the things they do in order to show how they could be reused by others in the first place.'

PEER-TO-PEER OR NOT

Dr Jose asked about peer-to-peer retrieval. It sent queries to relevant peers. It was more democratic. Usage information could be added to it to create the kind of retrievals the experts were talking about.

The others were dubious. 'A fascinating technology but it requires all peers to share an understanding of the goal and representation of the data', said one, adding: 'You need to make sure that you are not just equal but are basically identical.' Another thought: 'These ideas have been around for a long while and they have not been solved.'

Jussi Karlgren wondered what he would tell a small cultural heritage institution if it telephoned to ask about digitising its collection. He mused: 'Should I start discussing the benefits of client-server versus peer-to-peer solutions? I think not. I should probably

⁵ Networked Knowledge Organisation Systems, <http://nkos.slis.kent.edu>

⁶ Lorcan Dempsey, <http://www.oclc.org/research/staff/dempsey.htm>

⁷ Online Computer Library Centre (OCLC), <http://www.oclc.org>



tell them to get a really good scanner and make sure to put the provenance information in the right place and scan index cards before the objects themselves – that sort of thing.

‘Or should I warn them off digitising, tell them to wait a decade or so because then we will have sorted out everything?’

That brought wry chuckles from around the table, too, and Seamus Ross murmured disquietingly: ‘Small and medium-sized institutions have absolutely no understanding of how to represent resources.’ To be fair, he had prefaced the remark with a strong call for search tools that could help by drilling down to find ‘the Medici letter behind the splash front page’. He declared: ‘What we really need is search tools that drill into catalogues to bring back and aggregate the stuff so that it has some value. As well as addressing the representation issues, that might improve search engines.’

Douglas Tudhope recommended study of the UK **mda**’s work SPECTRUM (Standard Procedures for Collections Recording Used in Museums) XML⁸ that standardised some representation issues and the ANSI/NISO standard Z39.50⁹ protocol for Web retrievals. Traugott Koch thought the Renardus project had produced surprising results in its survey of Web usage. ‘Only 20 per cent of users come through the Website front doors. The other 80 per cent jumps from somewhere into the middle of the service via search engines.’

It was, he said, a reality no-one had designed for and gave a poor result to users. ‘They end up on such a page and it is totally out of context. They do not know what to do, they do not know what it means, and they do not know where it comes from or what is the logical feature before and afterwards.’ If the heritage sector was to build on being discovered this way, using Google or other big search engines, it needed to redesign most of its Web pages so that each carried its own context and help information, he said.

The Eleven spoke of passage-level retrieval, enriched interfaces, the Web-based virtual museum ‘Smithsonian without Walls’,¹⁰ the virtual learning environment ADEPT (Alexandria Digital Earth Prototype),¹¹ Amazon.com’s ‘recommender technology’, and implicit and explicit feedback systems.

However, none of these applications, Jussi Karlgren suggested, should be used by cultural heritage institutions. He thought: ‘They should provide their information in such a form that third party systems can access the information. And we should give them the information they need to provide for those systems as well as they can.’

There was consensus among the experts that institutions could encourage users to provide their personal data in exchange for better information and solutions. Traugott Koch agreed enthusiastically. ‘That is exactly what commercial companies do. They almost force academic users to enter IDs and passwords which not only enables individualised search histories and mappings of interest but allows connection to an individual research field,’ he said.

A QUESTION OF MONEY

The search for answers was reaching its climax. Moderator Jose wanted to pin down a combination of usage information, content based on information retrieval techniques and metadata terminology. He guided the Forum: ‘It allows us to personalise and adapt to a particular situation. It provides a wide variety of visualisation and interaction techniques.’

Douglas Tudhope thought: ‘The combination of these is interesting and, in itself, novel.’

The Forum thought that, fundamentally, the cultural heritage sector was not vastly different from other information domains. It could, therefore, benefit from most of the technology developed for the wider sphere. Simple tools for creating metadata would be useful. Existing tools were ‘somewhat cumbersome’

⁸ mda (Museum Documentation Association): SPECTRUM XML, http://www.mda.org.uk/index_rs.htm
⁹ ANSI/NISO standard Z39.50. Explanatory detail, Library of Congress Web, <http://lcweb.loc.gov/z3950/agency/>
¹⁰ Smithsonian without Walls Website, <http://www.si.edu/revealingthings/>
¹¹ ADEPT, Alexandria Digital Library Project, University of California, Santa Barbara, <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Smith/>



but it should not be difficult to make easier ones, they thought.

One big difference, said Dr Karlgren, was that most heritage information providers were not in it just for the money. More important was the institution branding and systems had to cater for that. He argued that institutions need not provide the meta-data tools but should have clear standard interfaces to their information and 'then rely on people to want to find their information properly'. His words opened a large can of worms.

Dr Ross did not agree. He wanted to know: 'Why should they give up ownership and management of their content? What is the motivation to provide the rights and access to their content; for you to provide an additional interface?'

'The question is whether or not the holdings of museums, libraries and archives is a public good that should be publicly accessible for free or whether they should be a commercial actor and generate income,' he insisted. Traugott Koch was placating. 'It is a very fundamental political discussion. We should not dare to make any recommendations to these institutions about what they should give away to other parties independent of business models and what business they are in.'

The argument about costs went on. Metadata and index input could never be entirely automatic, despite the beliefs of what Traugott Koch called 'our bosses and techie freaks ... and a lot of directors who might still believe that they are only a few years away from total automatic systems'.

Dr Ross feared that 'without significant automated processes the metadata cost necessary to catalogue these things is going to mean either outsourcing it to China or finding ways to automate the processes.'

Traugott Koch left it, saying: 'We know that the decisions about business models and tasks are dependent on the people who pay for it.'

And so, after this frosty moment, the Sixth Forum closed amicably enough. The conversation ended, as it had begun, with words from Traugott Koch. The experts had offered all the advice and guidance they could, but nothing altered the fact that cultural heritage would eventually have to come to terms with the new facts of digital life. They had the Sixth Sense to see that.





HUMANISTIC VS MACHINE CENTRED INFORMATION RETRIEVAL APPROACH

AN INTERVIEW WITH PIA BORLUND

By **Joost van Kasteren**

‘Most information retrieval systems are based on poorly conceived notions of what users need to know and how they should retrieve it. They do not reflect real user needs and search behaviour.’ According to Pia Borlund we should adapt retrieval systems to the user and not the other way around. This goes for scientific libraries, which is the field she is working in, and also for the cultural heritage sector with its collections of objects and documents.

Borlund is an associate professor at the Royal School of Library and Information Science in Denmark*. She developed an alternative approach to the evaluation of interactive information retrieval (IIR) systems. Alternative to the still dominating evaluation approach, which was developed in the sixties by Cyril W. Cleverdon at the Cranfield Institute of Technology, which focused on recall (= the fraction of relevant documents retrieved) and precision (= the fraction of retrieved documents that are relevant).

Borlund: ‘The Cranfield model treats information need as static, i.e. entirely reflected by user request and search statement, while in real life information needs can be dynamic. If you start with a topic that is new for you, you need to know what is going on, and you need to develop the right jargon. While searching, your information need “matures”, so to speak; it is dynamic in nature. It might become static. Many researchers check every week what articles have been pre-printed on the Web.’

Secondly, the Cranfield model promotes a rather limited concept of relevance in the sense that it uses only topical relevance. Borlund: ‘In my view relevance is a multidimensional concept which is also dynamic in nature. What are relevant changes over time? So relevance is – again – about understanding user needs and user behaviour. As a librarian you can never know what is relevant to a user. You might know the topic but you would not know why he is looking for information on that topic nor his angle.’

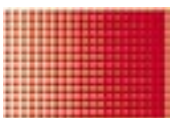
Borlund developed a new method for evaluating information retrieval systems taking into account the dynamic nature of user needs and the multidimensional and dynamic nature of relevance. It combines

a system-oriented approach like the Cranfield model with the user approach, which measures user’s satisfaction regardless of the actual outcome of the search. The method involves 25 to 30 test persons who are told a short ‘cover story’ to trigger a simulated information need. The cover story allows free user interpretation of the situation and hence heterogeneous information needs, to mirror a real world situation. The cover story also functions as a platform for situation-dependent relevance. With the story in mind people are asked to use the retrieval system to fulfil their simulated information need. Performance of the system is then measured by assessing how well the information needs of the test persons have been fulfilled, both in objective terms (did they find all the relevant information?) and in subjective terms (were they satisfied with the result?). Relative Relevance and Ranked Half Life are used as alternative performance measures (alternative with respect to the aforementioned recall and precision measures) that can handle non-binary degrees of relevance.

The evaluation method is used in the TAPIR (Text Access Potentials for Interactive Information Retrieval) research project. The project investigated the use of several cognitive representations of a document. One representation might be the author’s own perception of the work expressed in title and full text. Another might be derived from the indexing by a descriptor or from citations given to the work by other authors. The assumption is that the more cognitively different the representations are that point towards a certain document, the higher the probability that the document is relevant to a given set of criteria. ‘These poly-representation algorithms can be very useful in many domains, including the cultural heritage sector.’

Borlund believes that an important guideline in designing, developing and testing information retrieval systems is to start with the information need of the user. Not only to unlock scientific literature, but across other domains where collections are to be made accessible. ‘You need to change perspective’, she says. ‘Retrieval is not about cataloguing objects or information, but about information needs.’

* <http://www.db.dk/pb/>



PERSONALISATION TECHNIQUES IN INFORMATION RETRIEVAL

By Joemon M. Jose

INTRODUCTION AND OVERVIEW

Searching for information from digital archives is part of our day-to-day life. Formulating a good query is the first step in this process. It is widely acknowledged that query formulation, the transformation of a user's information need into a list of query terms, is a difficult task. Users don't find this process of serialising their thoughts to be intuitive, often leading to poor queries and a widening of the gap between the actual need and stated request. The problem can be circumvented once users learn how documents are indexed and also have a basic idea of how the retrieval engine judges relevance. However, this knowledge is not something we can take for granted.

This problem is magnified when the information need is extremely vague ('I don't know what I'm looking for, but I'll know it when I find it'). This common scenario typically results in the formulation of short queries, consisting of 1-3 non-discriminatory terms. This leads to the associated problem of information overload where millions of documents are returned for a broad query, e.g. in the context of Web search. Certain users may trawl through pages and pages of results but in reality most will view the first few result pages before accepting these to be the best possible results, albeit mistakenly in most cases. This typically leads to extended or unproductive search sessions where any initial time constraints are waived and users become dissatisfied. In addition, they unknowingly miss relevant documents.

During the search process, a user's information need is constantly developing and so therefore is the query or queries associated with the search session. As users think, digest, interpret and problem solve, this information need can be influenced by any number of factors, for example, the quality of search results or the content of a particular document. Current search techniques do not address these fundamental problems adequately.

As part of a recently concluded project¹ funded by the Engineering and Physical Sciences Research

Council (EPSRC), the Information Retrieval Group of the University of Glasgow's Department of Computing Science² worked on the development of prototypes to model users' current and long-term information needs.

Three systems based on the implicit feedback techniques are described below and evaluated based on a task-oriented and user-centred approach for the evaluation of personalised IR systems. Based on the commonly applied relevance feedback systems, the implicit technique uses relevance indicators gathered unobtrusively from searcher interaction, to modify the initial query. Whilst not as accurate as explicit feedback, it has been demonstrated that implicit feedback can be an effective substitute for explicit feedback in interactive information seeking environments.

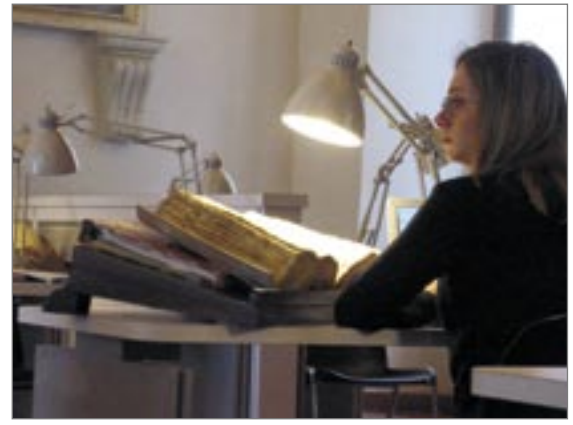
In the 'fetch' approach the system models users' long-term information needs and proactively fetches relevant documents for the user. The system uses a bundling technique allowing users to develop strategies for coping with the loss of context occurring when a variety of independent sources are viewed together. Over a period of time, through the observation of this bundling, the agent can build an accurate profile of the multifaceted information need and eventually recommend relevant Web pages without the need for users to mark documents explicitly.

RELEVANCE FEEDBACK SYSTEMS

Relevance feedback (RF) is the main post-query method for automatically improving a system's representation of a searcher's information need. However, the technique relies on explicit relevance assessments provided by the searcher (i.e. indications of which documents contain relevant information) and creates a revised query attuned to those marked.

RF systems suffer from a number of problems that hinder its practical usage. RF depends on a series of relevance assessments made explicitly by the searcher. The nature of the process is such that searchers must visit a number of documents and explicitly mark each as either relevant or non-relevant. This is

¹ EPSRC project: POW! Personalisation of Web Searches through Ostension & Summarisation, <http://www.dcs.gla.ac.uk/~jj/projects/pow/pow.html>
² <http://ir.dcs.gla.ac.uk>



a demanding and time-consuming task that places an increased cognitive burden on those involved. Documents may be lengthy or complex, users may have time restrictions or the initial query may have generated a poor result set. Therefore, searchers may be unwilling to provide relevance feedback.

RF systems typically adopt a binary notion of relevance: either a document is relevant, or it is not. If a document is only partially relevant the approach requires the searcher to choose between these two extremes, which means that there is no middle ground. In such circumstances it can be difficult for searchers to make decisions on what documents to assess as relevant.

Implicit RF, in which an IR system unobtrusively monitors search behaviour, removes the need for the searcher to explicitly indicate which documents are relevant. The technique uses implicit relevance indications, gathered unobtrusively from searcher interaction, to modify the initial query. Although not as accurate as explicit feedback, it has been demonstrated that implicit feedback can be an effective substitute for explicit feedback in interactive information seeking environments.

IMPLICIT FEEDBACK SYSTEMS

Implicit feedback systems remove the responsibility of providing explicit relevance feedback from the searcher. These systems infer which pages are relevant by analysing user actions such as the time taken to read pages. A variety of 'surrogate' measures can be used (hyperlinks clicked, mouseovers, scrollbar activity, etc.) to unobtrusively monitor user behaviour and estimate searcher interests. Through such means, it is possible to estimate document relevance implicitly. If a user 'examines' a document for a long time, or if a document suffers a lot of 'read wear', it is assumed to be relevant. The motivation for this idea is similar to that used to promote adaptive search systems, which develop and enhance their knowledge of searcher needs incrementally from inferences made about their interaction. Such systems aim to help struggling

searchers, who may have problems finding what they are looking for. In a similar way, implicit feedback based systems in concert with the searcher use implicit monitoring of interaction to generate an expanded query that estimates information needs. We have explored these ideas in two different applications: Web Search Interfaces and Image retrieval.

Personalised Interfaces for Web Search Systems

We present an interface (Figure 1), which we use to engender interaction and hence generate more evidence for the techniques we employ. Our interface works as an adjunct to search systems. The system takes user queries and despatches them to the Google Web search engine.

The returned results are processed and presented to the user. We use the top 30 documents from the Google results and summarise each of them with respect to the query. By this, we generate a maximum of four sentences from each document, which are also known as top ranked sentences. These sentences provide a much more granular representation of the document, with respect to the query.

In general, top ranked sentences, document titles, the document summary, and each sentence in the summary within its context of occurrence in the document are the various representations a user can view on the interface.

In our approach searchers can interact with different representations of each document. These representations are of varying length, are focused on the query and are logically connected at the interface to form an interactive search path. The representations present a higher level of granularity than the full text of documents, allowing the implicit approach to concentrate on the most relevant parts of individual documents.

Our objective is to help the users in information seeking activities. During the search process there are a number of possibilities for personalising. One is the crystallisation of the user query as the user is exposed to more and more relevant information. In another situation, the underlying information need itself changes. Information needs are dynamic and may

change in the light of new information. Over time, we apply statistical methods to successive lists of query expansion terms and use the resultant evidence to predict the degree of change (or development) in a searcher's information need. Different degrees of per-

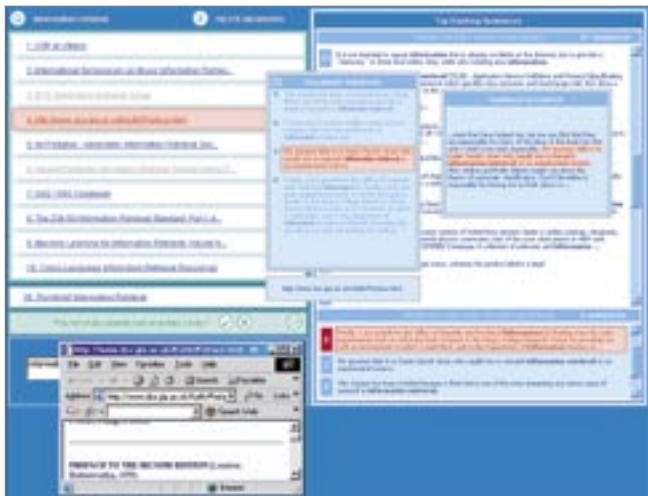


Figure 1: Interface of the Implicit Feedback based Web search system

ceived change result in different interface responses.

The interface therefore offers two forms of support: the implicit selection of terms to expand the query and an estimation of information need change. Through implicitly monitoring interaction at the results interface, searchers are no longer required to assess the relevance of a number of documents, or indeed consider entire documents for relevance.

Our approach makes inferences based on interaction and selects terms that approximate searcher needs. It uses the evidence it gathers to track potential changes in information need and tailor the results presentation to suit the degree of change in need. Large changes in perceived information needs result in new searches, but smaller changes result in less radical operations, such as re-ranking the list of retrieved documents or re-ordering representations of the documents presented.

The main aim of our approach is to develop a means of better representing searcher needs while minimising the burden of explicitly reformulating queries or directly providing relevance information. Devising systems that adapt to the information needs of those who use them is an important step in developing systems to help struggling searchers find the information they seek.

Image Retrieval

We have explored a similar approach for image retrieval. We have developed an interface (Figure 2) in which a user starts browsing with one example image. Subsequently a new set of similar images are presented to the user.

As a next step, the user – through selecting one of the returned documents – updates the query, which now consists of the original image and the selected image from the set of returned candidates. After a couple of iterations the query is based on the path of images.

In this approach, the emphasis is placed on the user's activity and the context, rather than predefined internal representation of the data. A path represents a user's motion through information, and taken as a whole is used to build up a representation of the instantaneous information need.

In a nutshell, it supports both browse-based and query-based approaches. It supports a query-less interface, in which the user's indication of the rele-

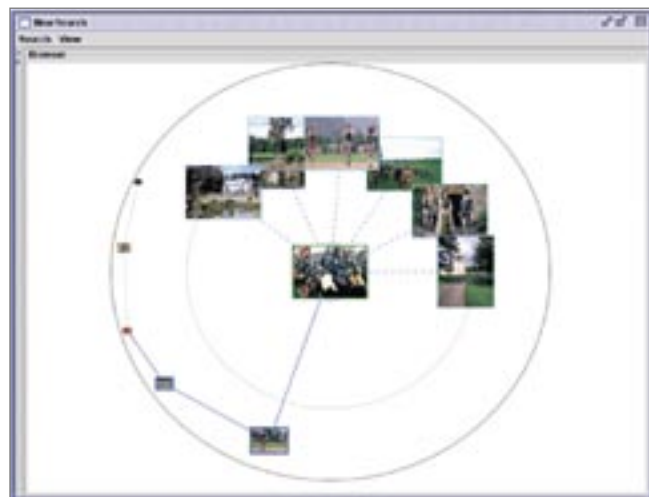


Figure 2: Image search interface

vance of an image – by selecting an image – is interpreted as evidence of its being relevant to his current information need. Therefore, it allows direct searching without the need for formally describing the information need.

FETCH: A PERSONALISED INFORMATION RETRIEVAL TOOL

In the above applications, we were modelling users' current information needs using interaction between the system and the user. In the approach below, we model users' long-term information needs

and proactively fetch relevant documents for the user. For this, we developed a system called Fetch (Figure 3), which is explained below.

A query is executed via the search area (1) and results are returned together with a query-biased summary (2) for each link (3) in the result set. Links



Figure 3: Fetch interface

can then be dragged on to the workspace (4) and grouped together with similar documents to form bundles (5) analogous to the way in which related documents are placed in the same folder on a desktop. Bundles on the workspace are also represented in the overview panel (6) in order to complement the flexibility of the workspace with a more structured view. The agent will at some future time analyse the bundles belonging to each user and formulate a new query for each. The system notifies the user of this new information by changing the colour of the bundle on the workspace from green to red (7). By double-clicking the updated bundle, instead of opening the bundle, a new search will be initiated using the associated query with results returned as before. Relevant links can then be dragged into new or existing bundles in the same fashion as before. The list of query terms can also be edited in the query editor (8) based on the quality of the first result set. Iterations of this form continue as long as the contents of the bundle are updated and thus the user's changing information need can be captured. The agent also checks for updated links on the workspace, alerting the user by changing the colour of the link icon from green to red.

Thus, Fetch adopts the flexible environment while incorporating a bundling technique that allows users to develop strategies for coping with the loss of context occurring when a variety of independent sources

are viewed together. Over a period of time, through the observation of this bundling, the agent can build an accurate profile of the multifaceted information need and eventually recommend relevant Web pages without the need for users to mark documents explicitly.

The Fetch interface gives a good visualisation of the information space. The bundles are visible on the workspace and it can be moved around. In this way, users are able to associate spatial clues with the bundles. For example, one could say all bundles on the top-right corner deal with museums.

EVALUATION OF PERSONALISED SEARCH SYSTEMS

The activity of evaluation has long been recognised as a crucially significant process through which information retrieval systems reach implementation in a real-world operational setting. Evaluative studies are concerned with assessment of the quality of a system's performance of its function, with respect to the needs of its users within a particular context or situation. The direction of such studies is commonly determined, and thus implicitly validated by the adoption of some kind of structured methodology or evaluative framework.

The traditional evaluative study of IR systems derives from the Cranfield Institute of Technology's projects in the early 1960s and survives in the large-scale experiments undertaken annually under the auspices of the Text Retrieval Conference (TREC). However, on many occasions the suitability of such a framework for the evaluation of interactive systems is questioned.

We have been following a task-oriented and user-centred approach for the evaluation of personalised IR systems. Our approach was based on the adoption of simulative work task environments to place the user in a realistic information seeking scenario. The basic idea is to develop simulated search tasks. Such tasks will allow the user to simulate an actual working environment and thereby better judge the relevance of documents from an actual information need perspective. In addition, such a situation facilitates a realistic interaction in a laboratory setting. The system, tasks and users are distributed with a proper statistical design in order to avoid learning bias.

RELEVANCE TO CULTURAL HERITAGE ARCHIVES

The systems introduced above demonstrated the use of implicit feedback. User evaluations have shown the usefulness of these approaches in resolving the information needs of the searchers. We believe that these techniques are directly transferable to the stakeholders of digital cultural heritage archives.

The users of cultural heritage archives are not always trained in the use of advanced computing facilities. Personalised search techniques enhance the quality of interaction for such users. They allow them to explore the collection effectively. Such techniques will also be the basis for search agents. If the system can identify and crystallise the search need from the interaction then we can create agents to search other collections and fetch more relevant documents.

References:

Martin, I. and Jose, J. M.: Fetch: A Personalized Information Retrieval Tool, Proceedings of the RIAO 2004 conference, April 2004, Avignon, France.

Martin, I. and Jose, J. M.: A Personalised Information Retrieval Tool, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Pages: 423-424, Toronto, Canada, July-August 2003, ACM Press.

White, R. W., Jose, J. M. and Ruthven, I. G.: A task-oriented study on the influencing effects of query-biased summarization in Web searching, *Information Processing & Management*, Volume 39, Number 5. Pages: 707-734, 2003, ISSN: 0306-4573.

White, R. W., Jose, J. M. and Ruthven, I. G.: A Granular Approach to Web Search Result Presentation. Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003). Pages: 220-227, Zurich, Switzerland, September 2003.

White, R. W., Ruthven, I. G. and Jose, J. M.: Finding relevant documents using top ranking sentences: An Evaluation of Two Alternative Schemes, Proceedings of the 25th SIGIR Conference on Research and Development in Information Retrieval. Pages: 57-64, Tampere, Finland, August 2002, ACM Press.





BUILDING A NEURAL NETWORK BASED DIGITAL LIBRARY

AN INTERVIEW WITH ANDREAS RAUBER

By Joost van Kasteren

Building digital collections of documents, images or music demands a lot of effort in analysing these 'objects', assigning metadata and designing ways of accessing them. As the amount of electronic information is growing exponentially, the build-up of digital collections has not kept pace. 'Why don't we do it the other way around,' says Andreas Rauber, 'and put objects in an information space, like a librarian puts books on a certain topic on the same shelf. But then automatically.'

Rauber is associate professor at the Vienna University of Technology, Department of Software Technology and Interactive Systems.* One of the projects he is involved in is the SOMLib Digital Library, a library that – automatically – clusters documents by content. Rauber: 'Retrieving useful information from digital collections is rather difficult. You have to formulate a query, specifying large numbers of keywords, and the result is often a list of thousands of documents, both relevant and irrelevant. Compare this with a conventional library or a large bookshop and the way we approach the stored information. Books are organised by topic into sections, so we can locate them easily. In doing so, we also see books on related topics, which might be useful for us.'

The conventional library system inspired Rauber and his co-workers to develop a digital library along the same lines, based on the self-organising map, the SOM in SOMLib. The SOM is an unsupervised neural network that can automatically structure a document collection. It does so by vectoring full text documents and then classifying them according to content. Rauber: 'Our SOM does not focus on specific words, so documents about river banks are separated from documents on banks as financial institutions and from documents about banks to sit on.'

Being a neural network, it has to be trained by presenting input data in random order. Gradually the neural network is fine-tuned until it clusters most of the documents correctly. Rauber: 'Sometimes you get

a strange result. One time we presented the text of a stage play to our SOM and it ended up in the news section. When we took a closer look, it came out that it was a very realistic play.'

To handle large document collections they have developed a hierarchical extension, the GHSOM (Growing Hierarchical SOM), which results in something like an atlas, says Rauber. 'An atlas of the world contains maps of the continents, of the countries on these continents and of regions and sometimes even cities. In a GHSOM you go from an overview of the different sections to an overview of the different compartments in a section to the topics within that compartment.'

The search has been made easier by adding a user interface, LibViewer, which combines the spatial organisation of documents by SOMLib with a graphical interpretation of the metadata, based on Dublin Core. Documents are assigned a 'physical' template like hard cover or paper and further metadata such as language, last time referenced and size of document. The metadata have a graphical metaphor so it seems as if you are standing in front of a bookcase, with all kinds of different books. Some thick, some thin, some green, some yellow and some look as if they have been extensively used, and others look brand new.

The SOM can also be used for 'objects' other than documents. Rauber has developed the SOM enhanced Jukebox (SOMeJB), which is built upon the same principles. It is used to organise pieces of music, for instance mp3 files, according to sound characteristics. Rauber: 'The result is a music collection organised in genres of similar music. So you get one cluster of heavy metal, one of classical music and one of Latin music. It differs from standard repositories of music in that you do not have to access through text metadata, like name of the artist or title of the song. You might use it even to search for songs of a certain genre on the Internet.'

* <http://www.ifstuwien.ac.at/~andi>



A CASE STUDY OF A FACETED APPROACH TO KNOWLEDGE ORGANISATION AND RETRIEVAL IN THE CULTURAL HERITAGE SECTOR

By Douglas Tudhope and Ceri Binding

INTRODUCTION: KNOWLEDGE ORGANISATION SYSTEMS IN DIGITAL HERITAGE

The trend within museums and digital heritage institutions to unlock the information in their collections involves opening up databases, previously the domain of the IT department, to a new range of users. These might, for instance, be members of the public searching a museum Website for information relating to an object which has been in the family for generations or they might be curators looking to create a virtual exhibit¹ from the objects in the collections database. There is a need for tools to help formulate and refine searches and navigate through the information space of concepts that have been used to index the collection. When technical terms are involved, a ‘controlled vocabulary’ is generally used to index the collection – if both searchers and indexers draw on the same standard set of words then the synonym mismatch problems common with Web search engines can be avoided. Controlled vocabularies provide a means to standardise the terms used to describe objects, by limiting the indexing vocabulary to a subset of natural language.

These controlled vocabularies have long been part of standard cataloguing practice in libraries and museums and are now being applied to electronic repositories via thematic keywords in resource descriptors. Metadata sets for the Web, such as Dublin Core, typically include the more complex notion of the Subject of a resource in addition to elements for Title, Creator, Date, etc. However, controlled vocabularies can do more than simply supply a list of authorised terms. They play a significant role, particularly when used to provide a mediating interface between indexed collections and users who may be unfamiliar with native

terminology. This applies to both existing collection databases and new collections of records, which may be ‘born digital’ but can be categorised and indexed using the same structures and techniques.

Knowledge is structured and organised so that a user can explore a network of related concepts to find the most appropriate one for a given situation. The different types of Knowledge Organisation System (KOS) include classifications, gazetteers, lexical databases, ontologies, taxonomies and thesauri.² There is a vast existing legacy of these intellectual knowledge structures (and indexed collections) to be found within cultural heritage institutions. A library might use the Dewey Decimal Classification (DDC), for example, while a museum might use the Art and Architecture Thesaurus (AAT). Other large, widely used KOS include AGROVOC,³ CABI,⁴ Library of Congress Subject Headings, MeSH,⁵ and many others.⁶ On the other hand, a large number of smaller KOS have also been designed to meet the needs of specialist applications or subject areas. In the UK, the **mda** (Museum Documentation Association)⁷ has facilitated the development of several specialised thesauri, such as the Archaeological Objects Thesaurus, the Railways Object Names Thesaurus and the Waterways Object Names Thesaurus.

NETWORKING KOS SERVICES

The rich legacy of KOS makes it possible to offer search options that go beyond the current generation of Web search engines’ minimal assumptions on user behaviour. However, this will require new thinking on the services that KOS can offer to the digital environment. Traditionally, attention has focused on methods for constructing KOS, with a view to their being used as reference material in print form. New possibilities have emerged

¹ e.g. the Science Museum’s Exhiblets: <http://www.sciencemuseum.org.uk/collections/exhiblets/index.asp>

² Gail Hodge gives a useful summary: http://nkos.slis.kent.edu/KOS_taxonomy.htm

³ Food and Agriculture Organization of the United Nations: AGROVOC Multilingual Thesaurus (Arabic, Chinese, English, Français, Español, Português), <http://www.fao.org/agrovoc/>

⁴ CAB International: CAB Thesaurus, a controlled vocabulary resource [over 48,000 descriptive terms] for the applied life sciences, <http://www.cabi-publishing.org/DatabaseSearchTools.asp?PID=277>

⁵ National Library of Medicine: Medical Subject Headings (MeSH), <http://www.nlm.nih.gov/mesh/>

⁶ For indexes of KOS on the Web, see <http://www.lub.lu.se/metadata/subject-help.html> and [http://www.w3.org/2001/sw/ Europe/reports/thes_links.html](http://www.w3.org/2001/sw/Europe/reports/thes_links.html)

⁷ **mda** (Museum Documentation Association), http://www.mda.org.uk/index_rs.htm



with online catalogues and Web search systems. Many existing KOS have been published and made available for Web-based access. However, they tend not to be fully integrated into indexing and search systems. Their interfaces are typically primarily designed for display purposes and are not appropriate for direct programmatic access. The lack of any standardised application programming interface (API) hinders attempts at interoperability.

In this case study, we first briefly introduce one particular KOS, the thesaurus, together with the faceted approach to KOS design. We then describe the FACET Project which investigated the potential of faceted thesauri in retrieval and reflect on some of our experiences during the project. We finish by discussing some key concerns for realising the potential of thesauri in Web-based systems, particularly the issue of access protocols.

THESAURI

The thesaurus is one of the most commonly used controlled vocabulary indexing tools in museums.⁸ It is a type of KOS structured by a core set of semantic relationships which are specified by international standards (BS5723, ISO2788, ANSI/NISO Z39.19).

Hierarchical relationships as illustrated in Figure 1 structure broader (more general) and narrower (more specific) concepts in relation to a given concept, and allow a thesaurus to be visualised as a series of concept hierarchies. Associative relationships describe somewhat looser connections between concepts. Equivalence relationships specify terms that can be considered as effective synonyms for a concept, according to the scope and objectives of a particular thesaurus. Thus a major thesaurus will typically include a large entry vocabulary – a network of linguistic equivalents, colloquial terms and synonyms which can be used to channel searchers towards the formal controlled vocabulary indexing terms ('preferred' terms, determined by established literary warrant). Used in this way the entry vocabulary acts as an effective mediation device, educating users about the nature and terminology of the search domain and providing direct links to terms actually used in index-

ing. Scope notes, used in conjunction with any evidence provided by the relationship structure, assist in the disambiguation of thesaurus terms. They also communicate the limits of the context within which a given term may be applied.

FACETED APPROACH

Many people recommend a faceted approach to thesaurus (and related KOS) design. For example, the AAT is a large thesaurus (about 125,000 terms), organised into seven facets (and 33 hierarchies as subdivisions) according to semantic role: Associated concepts, Physical attributes, Styles and periods, Agents, Activities, Materials, Objects and optional facets for time and place. According to the AAT online Web pages,⁹

Hierarchical relationships (BT/NT)	Associative relationships (RT)
<p>burnishing (polishing)</p> <p>BT <i>polishing</i></p> <p>NT <i>ball burnishing</i></p> <p><i>Processes and Techniques</i></p> <p>. <i>finishing (process)</i></p> <p>.. <i>polishing</i></p> <p>... burnishing (polishing)</p> <p>.... <i>ball burnishing</i></p> <p>... <i>electropolishing</i></p>	<p>burnishing (polishing)</p> <p>RT <i>burnishers</i> <i>(metalworkers)</i></p> <p>RT <i>flat burnishers</i></p> <p>RT <i>polishing irons</i></p> <p>RT <i>tooth burnishers</i></p> <p><i>burnishers (metalworkers)</i></p> <p>RT burnishing (polishing)</p>
Equivalence relationships (UF/USE)	Scope notes (SN)
<p>burnishing (polishing)</p> <p>UF <i>burnished</i> <i>(polished)</i></p> <p><i>burnished (polished)</i></p> <p>USE burnishing (polishing)</p>	<p>burnishing (polishing)</p> <p>SN Making shiny or lustrous by rubbing with a tool that compacts or smooths</p> <p><i>burnishing (photography)</i></p> <p>SN Method of obtaining a glossy surface on collodion prints by pressing them between rollers</p>

Figure 1: Examples of thesaurus structure from the Getty Art & Architecture Thesaurus

⁸ Willpower Information: Thesaurus principles and practice, <http://www.willpower.demon.co.uk/thesprin.htm>

⁹ Getty: Art & Architecture Thesaurus On Line, http://www.getty.edu/research/conducting_research/vocabularies/aat/about.html

'Facets constitute the major subdivisions of the AAT hierarchical structure. A facet contains a homogeneous class of concepts, the members of which share characteristics that distinguish them from members of other classes. For example, the term *marble* refers to a substance used in the creation of art and architecture, and it is found as a preferred term (*descriptor*) in the Materials facet. The term *Impressionist* denotes a visually distinctive style of art, and it is listed as a preferred term in the Styles and Periods facet.

Homogeneous groupings of terminology, or *hierarchies*, are arranged within the seven facets of the AAT. A broader term provides an immediate class or genus to a concept, and serves to clarify its meaning. The narrower term is always a type of, kind of, example of, or manifestation of its broader context. For example, *orthographic drawings* is the broader context for plans (*drawings*) because all plans are orthographic.'

Thus facets (almost always) constitute mutually exclusive groupings of concepts. Single concepts from different facets are combined together when indexing an object – or forming a query. It is a much simpler and more logical organisation than attempting to form one single hierarchy that encompasses all the different possible combinations of objects and materials and agents. Faceted thesauri or classification systems include the AAT, BLISS¹⁰ and MeSH, and faceted approaches are being increasingly employed in Web design.¹¹

Faceted browsing interfaces to Web databases have become popular recently. For example, the Flamenco system¹² dynamically generates previews of query results as the user browses different facets. This is an elegant browsing implementation. However, in some networked situations, there may be a separation between terminology services and collection level services, and query preview may be impractical where several databases might be involved. In such circumstances, faceted approaches to searching may be helpful and this was one of the main aims of the FACET project.

THE FACET PROJECT

FACET (Faceted Access to Cultural hEritage Terminology) is a recently completed research project,¹³ funded by the UK Engineering and Physical Sciences Research Council (EPSRC), which investigated the potential of the thesaurus in retrieval. The project was carried out in collaboration with the J. Paul Getty Trust, who provided the AAT – the

primary thesaurus used in the project, and the UK National Museum of Science and Industry (NMSI). An extract of the NMSI Collections Database acted as a testbed for the project. CHIN (Canadian Heritage Information Network) and **mda** (Museum Documentation Association) acted as advisors to the project.

FACET builds on work that started in 1991 when the University of Glamorgan was commissioned to develop a hypermedia museum exhibit on local history from the photographic archives of the Pontypridd Historical and Cultural Centre. This inspired an earlier research project to investigate a query-based approach to navigation and retrieval, rather than relying on *a priori* fixed links. Access routes were time, space and, as subject index, the Social History and Industrial Classification.¹⁴

Our aim in FACET was to make use of facet structure in retrieval. The interface allows users to select terms from appropriate facets and combine them in a query. It is possible to conduct highly specific searches by combining concepts from different facets. This has the potential for very precise results if an item in the collection is indexed by these same terms.

However, in many cases it is unlikely that exactly the same combination of terms will have been used in indexing. Perhaps a term has been omitted or the searcher may have chosen a more specific concept than the indexer thought appropriate. Alternatively, there may not be any exactly matching objects in the collection, although there might be similar objects potentially of interest.

Toni Petersen, then Director of the Getty Art and Architecture Thesaurus Project, outlined key unsolved issues for system designers seeking to take advantage of the AAT in retrieval (in a discussion of the National Art Library database at the Victoria and Albert Museum, London), which was inspirational for some of our key research aims:¹⁵

'The major problem lies in developing a system whereby individual parts of subject headings containing multiple AAT terms are broken apart, individually exploded hierarchically, and then reintegrated to answer a query with relevance.'

Our solution involved a technique known as *semantic expansion*, whereby search terms are supplemented by additional terms representing similar concepts, based on their relative positions and relationships within the thesaurus structure. For example, a searcher interested in items made of *ebony* may also be interested in items made of a specific type

¹⁰ Bliss Classification Association: Bibliographic Classification (BC2 or Bliss), <http://www.sid.cam.ac.uk/bca/bchist.htm>
¹¹ Rosenfeld, R. and Morville, P. 2002: Information Architecture for the World Wide Web (2nd ed.). O'Reilly.
¹² Flamenco Search Interface Project, <http://bailando.sims.berkeley.edu/flamenco.html>
¹³ FACET project, <http://www.comp.glam.ac.uk/~FACET/>
¹⁴ Tudhope, D.: Geographical Access to Museum Hypermedia Exhibits, *mda Information*, Vol 2, No 3, <http://www.mda.org.uk/info23t.htm#Hypermedia>
¹⁵ Petersen, T. 1994: The National Art Library and the AAT. *Art and Architecture Thesaurus Bulletin*, 22, 6-8.

of *ebony*, such as *black ebony*, *marblewood*, or *kaki* (Japanese *ebony*). The thesaurus structure specifies the precise nature of the relationship between each of these terms, allowing a search process to expand the initial query automatically to include closely related terms. Having determined a degree of closeness between thesaurus terms, we then determine a suitable result set by comparing the expanded query with indexing terms. The results are displayed as a ranked list in order of decreasing relevance to the initial query. Figure 2 illustrates the Results window where a particular result (with overall match of 56 per cent) has been double-clicked to show the degree of match of individual terms in the query (*armchairs, brocading, mahogany, Edwardian*). Note that no query term matched exactly but all had partial matches to semantically close index terms. Relevance to the searcher will depend on context. The point is to provide a semantic expansion option for the user when exact matches are not available.¹⁶

Various standalone prototype systems were developed in an iterative design and evaluation cycle. Evaluation sessions were conducted on two major standalone prototypes. Data gathered included transcripts of think-aloud sessions, screen capture videos, user action logs and observation notes. The sessions involved 23 users (the vast majority being museum-related, including cataloguers, collections management and curators). Participants were given an introduction which walked them through a training task. They were then asked to carry out a number of additional search tasks, corresponding to typical museum enquiries. Our focus was on the process of user interaction and searching behaviour with the system. Evaluation of the first prototype illuminated important design issues, including how best to support the controlled vocabulary search process. Major enhancements to the interface followed.

The standalone system focused on the collections from the National Railway Museum,¹⁷ which is part of NMSI. The NRM Furnishings collection, which includes objects from the ‘Palaces on Wheels’ Royal Train collection (such as Queen Victoria’s saloon), offered rich detail for indexing with the AAT. For example, the general description of one such object reads: ‘*Carver chair, Oak with oval brocade seat. Prince of Wales crest on back from Royal Saloon of 1876*.’ Analysis of evaluation data from a second prototype is still ongoing but has fed into the design of the Web demonstrator.¹⁸ This was one of the final outcomes of the project and explored how the techniques from the standalone systems could be employed within a Web environment as dynamically generated

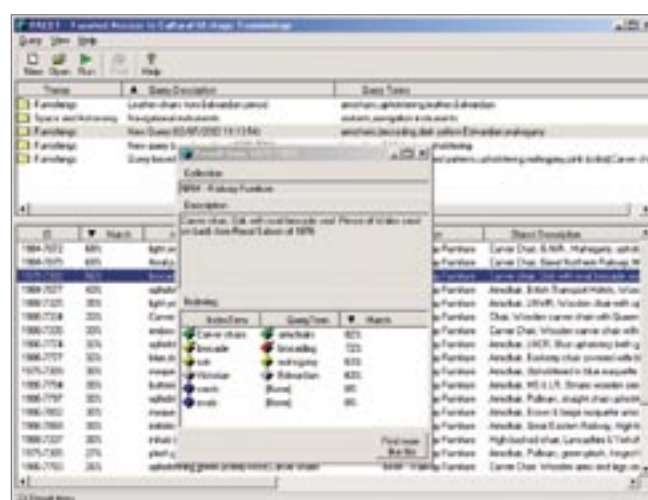


Figure 2: Example from FACET standalone system showing a result from semantic expansion

Web components. The interface does not rely on pre-built static HTML pages; thesaurus content is generated dynamically.

We also experimented with a number of smaller specialist thesauri, including draft versions of the **mda**'s Railway Object Names Thesaurus and Waterways Object Names Thesaurus, together with the Alexandria Digital Library's Feature Type Thesaurus. Through participating in the work of the Railways Terminology Working Group during the development of the thesaurus and in its peer review process, we came to appreciate some of the immense work involved in developing even a smaller thesaurus. This includes reaching consensus¹⁹ both on choice of terminology and on how it should be organised in hierarchies.

In addition to providing the AAT, the Getty Vocabulary Program's²⁰ Web interface and User's Guide to the AAT Data Releases (P. Harpring ed.) were useful sources of ideas for developing FACET's own interface, as was CHIN's experience in developing Artefacts Canada,²¹ one of the first Web-based 'virtual museum' applications with an interface incorporating a major thesaurus, such as the AAT.

Collaboration with various NMSI staff and their experience with the collection management database and public inquiries proved extremely useful throughout the project. One motivational example for the research was a (1997) public request to the Science Museum for information on *eighteenth century European celestial navigation instruments*. At the time, this request highlighted some difficulties with existing techniques since initial query terms did not generate easy matches. Multiple queries to several fields in the

¹⁶ Details of the algorithm are given in <http://www.glam.ac.uk/soc/research/hypermedia/publications/jcdl02.pdf>

¹⁷ National Railway Museum, <http://www.nrm.org.uk>

¹⁸ FACET Web interface – Demonstrations, <http://www.comp.glam.ac.uk/~FACET/webdemo/>

¹⁹ From the launch press release: 'The mda Railway Terminology Working Group has wide representation from organisations with an interest in railway collections and information. The group formed in

1996 following the mda Terminology Conference and includes staff from the National Railway Museum, York; London's Transport Museum; Beamish, North of England Open Air Museum; the Historical Model Railway Society; Heritage Railways Association; English Heritage, National Monuments Record Centre; and mda.'

²⁰ Getty Vocabulary Program, http://www.getty.edu/research/conducting_research/vocabularies/aat/

²¹ CHIN, Artefacts Canada, http://www.chin.gc.ca/English/Artefacts_Canada/index.html

collections database were required, using terms such as *Astrolabe* and *Octant*, from different AAT hierarchies. Revisiting the issue, FACET's semantic expansion on *navigation instruments* was able to short-cut this process. Items from the collection indexed by *sextants, astrolabes, etc.* now resulted from a single query.

Our initial research aims focused on the potential of thesauri as semantic retrieval tools. However, we came to realise that many of the techniques were equally applicable as tools to assist the indexing or cataloguing process. The indexer also needs to map from the terminology employed in an object description to the controlled terminology used by the thesaurus. The indexer may wish to browse in order to explore the precise local context of a term in the thesaurus. The indexer may even wish for automatic suggestions by a thesaurus-based indexing system of likely terms to use. We reported on initial investigations that suggested some potential for semi-automatic approaches at the 2002 **mda** conference,²² in the same session where Helen Ashby reported on the development of the Railway Object Names Thesaurus.²³

THESAURUS REPRESENTATION AND ACCESS PROTOCOLS

With the advent of a viable distributed information architecture in the form of the Web comes the prospect of a global online museum. The use of common standard representations potentially enables searching across multiple collections, effectively blurring the physical boundaries between institutions.²⁴ A prerequisite for this is the use of standardised protocols for query and access to content. The adoption of common standards has the benefit of enabling a logical division of effort. KOS resources, search interfaces, cataloguing/indexing/mapping tools and indexed collections using common thesaurus protocols may all be developed by separate institutions, and may be physically hosted in separate locations. This would have the effect of allowing the institutions to concentrate on their core area of specialisation, tapping into the domain knowledge of others as and when appropriate. It also reduces the potential for duplication of effort and the duplication of data.

Linda Hill and colleagues have argued for 'a general KOS service protocol from which protocols for specific types of KOS can be derived'.²⁵ The idea is to provide programmatic access to KOS content

through the various types of (Web) services mentioned above, as opposed to thinking only of interactive human interfaces. Thus, in future a combination of thesaurus and query protocols might permit a thesaurus to be used with a choice of search tools on various kinds of database. This includes not only controlled vocabulary search applications but also collections without controlled metadata. For example, semantic query expansion services could be used with both free text and controlled vocabulary indexed collections.

A variety of interchange format specifications for the representation and dissemination of thesaurus data have been developed and are in use today. These are tagged text formats such as the MARC21 Authority format as used by the J. Paul Getty Trust, XML-based formats such as the ZThes DTD,²⁶ and RDF representations such as SWAD-Europe's SKOS-Core schema.²⁷ In order to facilitate distributed thesaurus access, a platform-neutral access protocol should be used to manipulate thesaurus data. Protocols for retrieving thesaurus data are closely linked to thesaurus representation formats. The CERES,²⁸ Zthes and ADL²⁹ protocols are reviewed in a recent paper, which also reports on the FACET Web demonstrator.³⁰ The Simple Knowledge Organisation System (SKOS) API is a more recent development, which defines a core set of methods for programmatically accessing and querying a thesaurus based on the SWAD-Europe project's SKOS-Core schema.³¹ The NKOS Website³² and discussion list are good sources of information on new developments. NKOS Workshops at ECDL conferences allow discussion on KOS data exchange and standards issues. The workshop at ECDL 2003³³ included reports from the US NISO and UK BSI Thesaurus Standards Groups, currently considering revisions to the existing standards.

FUTURE ISSUES FOR NETWORKING KOS IN DIGITAL HERITAGE

The Web demonstrator was our first step in exploring issues underlying networked access to KOS, something we intend to pursue in future work. Results from FACET show that best-match (ranked result) approaches can be applied to KOS-based queries via semantic expansion of query terms. The Web interface also showed that semantic expansion may be employed as a browsing tool when wishing to hide some of the complexity of hierarchical structures.

Existing KOS already have rich resources to offer digital heritage Web applications, notwithstanding

²² Tudhope, D., Binding, C., Blocks, D. and Cunliffe, D. 2002: Thesaurus based indexing and retrieval, <http://www.mda.org.uk/conference2002/paper15.htm>

²³ Ashby, H. 2002: Cocks and Blowers - decoding Railspeak, <http://www.mda.org.uk/conference2002/paper13.htm>

²⁴ For a brief review of pioneering interoperability projects, such as Aquarelle and CIMI: Dawson, D. 1998. Aquarelle to Z39.50: the A to Z of access to cultural heritage information. *New Review of Hypermedia and Multimedia*, 4, 245-254.

²⁵ Hill, L., Buchel, O. and Janée, G. 2002: Integration of Knowledge Organization Systems into Digital Library

Architectures (Position Paper for 13th ASIS&T SIG/CR Workshop, 17 November 2002), http://alexandria.sdc.ucsb.edu/~lhill/paper_drafts/KOSpaper7-2-final.doc

²⁶ Zthes: A Z39.50 Profile for Thesaurus Navigation, <http://zthes.z3950.org/profile/current.html>

²⁷ SWAD-Europe Thesaurus Activity, <http://www.w3.org/2001/sw/Europe/reports/thes/>



future Semantic Web developments which will tend to be more resource intensive. The critical issue facing KOS in Web environments is that existing standards are based in the print world and are not concerned with data interchange formats. KOS intellectual resources can be exploited in searching. However, the lack of standardised access and interchange formats currently impedes the wider use of such resources in the distributed Web environment. Programmatic access requires commonly agreed protocols building on lower-level standards, such as Web services. The development of common KOS representation formats and service protocols is closely linked. Progress needs to be made on both dimensions if standards are to be achieved. A service protocol should be expressed in terms of a well defined but extensible set of KOS data elements and relationships, with the relationship type a parameter to the protocol commands. This would allow the specialisation of the current thesaurus relationships.

Users tend to be unaware of the relative effectiveness of different search techniques on any particular collection and need assistance in search strategies. Support for translating from user-specified free text terms to appropriate controlled vocabulary terms is particularly important, both in disambiguating homographs (a choice of KOS concepts) and in suggestions where the user is having difficulties in locating suitable controlled terminology. More work needs to be done on simple 'search box' interfaces, where KOS are used behind the scenes to support free text search. However, service protocols also need to be able to support the development of innovative and responsive Web interfaces that encourage different types of users to take full advantage of the resources offered by KOS for searching digital heritage collections.

A substantial amount of intellectual effort is expended in the initial compilation of a complex knowledge structure such as a thesaurus. Additionally, the evolving nature of language and culture dic-

tates that, once compiled, frequent maintenance of such resources is required in order for them to remain relevant to contemporary audiences. Ease of dissemination, maintenance and use is essential for the fruits of this effort to be fully realised. A balance needs to be struck between maintaining a standard version of a commonly used KOS for interoperability, while also allowing local institutions to tailor and augment it. This might, for instance, involve adding more specific local concepts as leaf hierarchies. In the long run, there may also need to be agreement on intellectual property right issues, and in some cases possibly licensing models, if we wish to make available for general use thesauri developed by a wide spectrum of organisations.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (Grant GR/M66233/01). We also wish to thank Helen Ashby, Ann Borda, Alice Grant, Sarah Norville, Charlotte Stone and other staff from the National Museum of Science and Industry for their assistance, together with the J. Paul Getty Trust for provision of the AAT, and **mda** and CHIN staff for helpful advice.

²⁸ CERES Thesaurus Protocol and browser, <http://ceres.ca.gov/thesaurus/>

²⁹ ADL Thesaurus Protocol, <http://www.alexandria.ucsb.edu/thesaurus/protocol>

³⁰ Binding, C. and Tudhope, D. 2004: KOS at your Service: Programmatic Access to Knowledge Organisation Systems. In: *Journal of Digital Information*, Vol 4, Issue 4, <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Binding/>

³¹ SWAD-Europe Thesaurus Activity: SKOS API, <http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html>

³² NKOS Website, <http://nkos.slis.kent.edu/>

³³ Tudhope, D.: ECDL 2003 Workshop Report: Networked Knowledge Organization Systems/ Services (NKOS): Evolving Standards, In: *D-Lib Magazine*, September 2003, <http://dlib.ejournal.ascc.net/dlib/september03/09inbrief.html#TUDHOPE>



SELECTED RESOURCES

Some basic reading:

'Resource Discovery - A Definition', Research Data Network CRC, Resource Discovery Unit, University of Queensland, <http://archive.dstc.edu.au/RDU/RD-Defin/>

Heting Chu: Information Representation and Retrieval in the Digital Age. Medford: Information Today Inc, 2003.

Tim Bray: On Search, the Series (2003), <http://www.tbray.org/ongoing/When/200x/2003/07/30/OnSearchTOC>

Journal of Digital Information (JoDI), Theme: Information Discovery (Editor: Traugott Koch, Lund University), <http://jodi.ecs.soton.ac.uk/?theme=d>

Research challenges:

Challenges in Information Retrieval and Language Modeling. James Allen et al. (2002). September 2002, <http://www.sigir.org/forum/S2003/ir-challenges2.pdf>

Conferences & workshops:

ACM SIGIR Conference on Research and Development in Information Retrieval, <http://www.sigir2004.org>

European Conference on Information Retrieval (ECIR), <http://ecir04.sunderland.ac.uk>

Search Engine Meeting 2004: White Papers and Presentations, <http://www.infonortics.com/searchengines/sh04/04pro.html>

Text REtrieval Conference (TREC) workshop series, <http://trec.nist.gov>

Search engines:

Searchenginewatch: <http://searchenginewatch.com>

Duncan Parry: Searching for success: an update on search engine developments (15 April 2004), <http://www.freepint.com/issues/150404.htm#tips>

The hidden Web:

Michael K. Bergman: The Deep Web: Surfacing Hidden Value (08/2001), <http://www.brightplanet.com/technology/deepweb.asp>

The University of California's tutorial 'Finding Information on the Internet' gives a detailed explanation of why some resources remain invisible to spiders: <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html>

Discovery and retrieval of specific resources:

Manuscripts

Jan Roegiers: Integrated Resource Discovery and Access of Manuscript Materials: the User Perspective. In: *LIBER Quarterly*, Volume 13 (2003), No. 3/4, <http://liber.library.uu.nl/publish/articles/000039/>

Images

Technical Advisory Service for Images (TASI): Searching the Internet for Images, <http://www.tasi.ac.uk/resources/searchingresources.html>

TASI review of image search engines (February 2003), <http://www.tasi.ac.uk/resources/searchengines.html>

Music

J. Stephen Downie (2003): Music information retrieval. In: *Annual Review of Information Science and Technology* 37: 295-340. Available from http://music-ir.org/downie_mir_arist37.pdf

Audiovisual in general

Alan Smeaton: Access to Archives of Digital Video Information. A comprehensive presentation at the Search Engine Meeting 2004, <http://www.infonortics.com/searchengines/sh04/slides/smeaton.pdf>

Documentaries

Tobun Dorbin Ng, Howard D. Wactlar (2002): Enriching Perspectives in Exploring Cultural Heritage Documentaries Using Informedia Technologies. In: Proceedings of the 4th International Workshop on Multimedia Information Retrieval, http://www-2.cs.cmu.edu/~wactlar/ACMMM02_Culture.pdf

Weblogs

Phil Bradley: Search Engines: Weblog search engines (30 July 2003). In: *Ariadne*, Issue 36, <http://www.ariadne.ac.uk/issue36/search-engines/>



THE ROME FORUM PARTICIPANTS AND CONTRIBUTORS

Ceri Binding, Hypermedia Research Unit, School of Computing, University of Glamorgan, UK

Ceri Binding is a Research Fellow in the School of Computing, University of Glamorgan. He graduated with a BSc in Computer Studies from the University of Glamorgan in 1997 and worked as an Analyst Designer/Programmer for Hyder IT before joining Glamorgan in 2000. He had responsibility for development work on the FACET project and implemented various standalone and Web systems for the project. The Web demonstrator can be seen online at <http://www.comp.glam.ac.uk/~FACET/webdemo/>. Related research interests include Knowledge Organisation Systems, intelligent Web-based retrieval and interface design. His publications can be viewed at <http://www.comp.glam.ac.uk/~FACET/publications.asp>

Jacques Bogaarts, Nationaal Archief, The Netherlands

After working as a mathematics teacher, Jacques Bogaarts was trained as an archivist at the Dutch Archiefschool. From 1992 to 1997 he was involved in the PIVOT project of the Nationaal Archief. He helped to develop policies and methods that led to a new way of appraising archives on a functional base. From 1997 to 2001 he was senior consultant on records management at the Dutch Ministry of Water Management Transport and Public Works. Since 2001 he has been a senior consultant at the Nationaal Archief in The Hague. His work concerns digital preservation and access. He is a member of the International Council on Archives (ICA) working group on Current Records in Electronic Environments (CER), and participates in DPL-light, a consortium consisting of major cultural heritage institutions of The Netherlands, four universities and some partners from industry. DPL-light aims to develop state-of-the-art tools and methods for the digitisation of and access to cultural heritage.

Pia Borlund, Royal School of Library and Information Science, Denmark

Pia Borlund, PhD, is Associate Professor at the Aal-

borg branch of the Royal School of Library and Information Science where she lectures, researches and publishes on methods for evaluation of (interactive) information retrieval systems (e.g. simulated work task based approach), experimental design, information seeking behaviour, and concepts of relevance. Dr Borlund is a member of the Editorial Board of the *International Journal of Information, Processing & Management*, and sits on numerous conference committees, e.g. SIGIR, ECIR, CoLIS, AMR and NordI&D. She is involved in the Nordic NORS-LIS research school, and is a visiting professor at Lund University, Sweden. Dr Borlund is, together with research professor Peter Ingwersen, co-founder of the TAPIR research project, which forms a stimulating environment for research on information retrieval, information seeking behaviour, HCI-related issues, as well as bibliometric/webometric activities. See: <http://www.db.dk/pb/default.htm>

Paolo Buonora, Archivio di Stato di Roma, Italy

Paolo Buonora has a degree in Philosophy from the University of Rome 'La Sapienza' (1976). He worked in the Italian State Archive Administration from 1978, where he was first involved in editing the Guida Generale degli Archivi di Stato italiani. From 1986 he worked in the Soprintendenza archivistica per il Lazio, surveying audiovisual archives, municipal archives; from 1989 to 1991 he worked at Perugia University, engaged in doctoral research on 'Urban and rural history'; and from 1991 to 1994 he was employed again in the Soprintendenza archivistica per il Lazio. After 1994 he worked in the Archivio di Stato di Roma, where he was responsible for the photograph service and several working groups on informatics applications in archival documentation. From 1997 until the present time he has planned and directed the Imago II project in the Archivio di Stato di Roma. See: <http://www.asrm.archivi.beniculturali.it>

Guntram Geser, Salzburg Research, Austria

Dr Geser is Head of Information Society Research at Salzburg Research, Austria. In the eCulture domain his recent work includes the strategic study Digi-

CULT: Technological Landscapes for Tomorrow's Cultural Economy (2001) and DigiCULT Forum (2002–present); see: <http://www.digicult.info/pages/report.php>. He also co-authored the strategic study EP2010: The Future of Electronic Publishing Towards 2010 (September 2003) for the European Commission's Directorate-General for the Information Society; see: <http://ep2010.salzburgresearch.at>. Dr Geser studied Communication and Political Science at the University of Salzburg and Telematics Management at the Donau-University Krems. Before joining Salzburg Research, he worked on research projects in the fields of media history and cultural studies in Berlin (Technische Universität; Deutsche Film- und Fernsehakademie) and Amsterdam (Instituut for Film- en Televisiewetenschap; Nederlands Filmmuseum). He lectured at the University of Vienna on science journalism and served as media consultant for the Austrian Cultural Service.

Joemon M. Jose, Department of Computing Science, University of Glasgow, UK

Dr Joemon Jose is a Lecturer in the Department of Computing Science of the University of Glasgow. He completed his PhD in Multimedia Information Retrieval from the Robert Gordon University, Aberdeen. He is an active information retrieval researcher and supervises a number of PhD students. He is the principal investigator of the EPSRC funded project on personalisation of Web searches through ostension and summarisation. Dr Jose is on the programme committees of many Information Retrieval and related conferences. His current research interests include adaptive and context sensitive IR, video summarisation and browsing, peer-to-peer information retrieval and the task-oriented and user-centred evaluation of interactive Retrieval Systems.

See: <http://www.dcs.gla.ac.uk/~jj/>

Jussi Karlgren, Swedish Institute of Computer Science (SICS), Sweden

Jussi Karlgren received his Candidate of Philosophy in Computational Linguistics and Mathematics in 1988, a Licentiate of Philosophy in Computer and Systems Sciences 1992, and a PhD in Computational Linguistics in 2000 – all at Stockholm University. He has since 1990 been employed at SICS, the Swedish Institute of Computer Science, in various instantiations of the Language and Interaction Laboratory. Previously he has worked at SISU, the Swedish Institute for System Development, been a visiting student at Columbia University, a programming assistant at Xerox PARC, and an assistant research scientist in the

PROTEUS project at New York University. During the academic years 1997–99 he substituted for Kimmo Koskenniemi as professor of computational linguistics in Helsinki; he has also taught and supervised students at Stockholm University and at the Royal Institute of Technology in Stockholm, and co-ordinated an EU-funded research project. In addition, he is member of the board of the Joint Group for Swedish Computer Terminology. Jussi Karlgren's primary research interests are dialogue and communication, taken broadly enough to include human-computer dialogue design, natural language interaction, information retrieval, and language typology. Currently Jussi Karlgren is responsible for the research theme of information access and information refinement at SICS.

See: <http://www.sics.se/~jussi/>

Traugott Koch, Knowledge Technologies Group (NetLab), Lund University Libraries, Sweden

Traugott Koch is Senior Librarian and Digital Library Scientist at the Knowledge Technologies Group (NetLab), Lund University Libraries in Sweden. In parallel, he is a member of the Knowledge Discovery and Digital Library Research Group (KnowLib) at the Department of Information Technology at Lund University. Traugott Koch has during the last 14 years worked mainly on European Union, Nordic and national digital library projects, e.g. the EU projects DESIRE I and II, EULER, a TEMPUS project in Lithuania, Renardus and E-Culture Net, the Nordic Metadata Project, EELS (the Engineering Electronic Library Sweden) and the project 'Intelligent components in a distributed Digital Library'. At present, DELOS: Network of Excellence on Digital Libraries, is his main project on an international level. Koch's main areas of interest are knowledge organisation, classification and indexing; automatic classification; Semantic Web; metadata; quality controlled subject gateways on the Internet; information discovery and retrieval; and digital library development. He is Theme Editor Information Discovery of the electronic peer reviewed journal *Journal of Digital Information* (JoDI), <http://jodi.ecs.soton.ac.uk/?theme=d>, and a member of several other editorial and project boards. He has been involved in the development and maintenance of Dublin Core Metadata Initiative DCMI standards since 1996 and is a member of the DCMI Usage Board and of the DCMI Advisory Board. See: <http://www.lub.lu.se/netlab/staff/koch.html>

John Pereira, Salzburg Research, Austria

Mr Pereira graduated from the University of Wollongong, New South Wales, Australia in 1996 with a

double major BA in Industrial Relations and Legal Studies. Recently he completed a certified E-Business Manager programme from the Donau University in Krems, Austria. Prior to joining Salzburg Research he was Customer Relations Manager within the Northern Division at Sony DADC. At Salzburg Research, Mr Pereira has been involved in the successful marketing and organisation of a European-wide multimedia contest, which included organising relevant conferences, exhibitions and seminars that promote best practice in digital content publishing. In 2001-2002 he managed a pilot project with SES Astra to provide satellite based interactive television to SMEs' for education and training. He is currently project manager of DigiCULT Forum, an EU funded initiative to provide mission-critical information in the selection and use of digital technologies for Europe's heritage organisations.

Andreas Rauber, Department of Software Technology and Interactive Systems, Vienna University of Technology, Austria

Dr Rauber is Associate Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology. Having been a member of the Faculty of the Vienna University of Technology since 1997, he joined the National Research Council (CNR) of Italy in Pisa in 2001, followed by a stay at the French National Institute for Research in Computer Science and Control (INRIA) in Paris, in 2002, both as an ERCIM Research Fellow. He received the Cor Baayen Award of the European Research Consortium for Informatics and Mathematics in 2002. Dr Rauber is a member of the Association for Computing Machinery (ACM), The Institute of Electrical and Electronics Engineers (IEEE), the Austrian Society for Artificial Intelligence (ÖGAI), and serves on the board of the IEEE Technical Committee on Digital Libraries (TCDL). He is actively involved in several research projects in the field of Digital Libraries, including the DELOS Network of Excellence on Digital Libraries. His research is focused on the organisation and exploration of large information spaces, as well as Web archiving and digital preservation. See: <http://www.ifs.tuwien.ac.at/~andi>

Seamus Ross, HATII & ERPANET, University of Glasgow, Scotland

Seamus Ross, Director of Humanities Computing and Information Management at the University of Glasgow, runs HATII (Humanities Advanced Tech-

nology and Information Institute) of which he is the founding director (<http://www.hatii.arts.gla.ac.uk>). He is also Principal Director of ERPANET (Electronic Resource Preservation and Network) (IST-2001-32706) a European Commission activity to enhance the preservation of cultural heritage and scientific digital objects (<http://www.erpanet.org>). He is a lead partner in The Digital Culture Forum (DigiCULT Forum, IST-2001-34898), which works to improve the take-up of cutting edge research and technology by the cultural heritage sector. He is leader of the Preservation Cluster of the Network of Excellence, DELOS: a Network of Excellence on Digital Libraries. Before joining Glasgow he was Head of ICT at the British Academy and before that a technologist at a company specialising in knowledge engineering. He holds a doctorate from the University of Oxford. His recent publications include: *Digital Library Development Review*, National Library of New Zealand, (Wellington, 2003), http://www.natlib.govt.nz/files/ross_report.pdf; 'Cyberculture, cultural asset management and ethnohistory: Preserving the process and understanding the past', *A/chivi/ & C/omputer*, XII.1/ (2002), and he was co-author of *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation* (2003), <http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf>

Pasquale Savino, Information Science and Technology Institute, CNR, Italy

Pasquale Savino is senior researcher at the Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of the Consiglio Nazionale delle Ricerche (CNR).





He graduated in Physics at the University of Pisa, Italy, in 1980. From 1983 to 1995 he has worked at the Olivetti Research Labs in Pisa; since 1996 he has been a member of the research staff at CNR's Istituto di Elaborazione della Informazione in Pisa, working in the area of multimedia information systems. He has participated in and co-ordinated several EU-funded research projects in the multimedia area: MULTOS (Multimedia Office Systems), OSMOSE (Open Standard for Multimedia Optical Storage Environments), HYTEA (HYperText Authoring), M-CUBE (Multiple Media Multiple Communication Workstation), MIMICS (Multipart Interactive Multimedia Conferencing Services), and HERMES (Foundations of High Performance Multimedia Information Management Systems). Recently, he co-ordinated the EU-IST project ECHO (European Chronicles on Line). His current research interests are multimedia information retrieval, multimedia content addressability and indexing. See: <http://faure.isti.cnr.it/~savino/>

Douglas Tudhope, Hypermedia Research Unit, School of Computing, University of Glamorgan, UK

Douglas Tudhope is Reader in the School of Computing at the University of Glamorgan and leads the Hypermedia Research Unit. Before joining the University of Glamorgan, he worked as a programmer/analyst at the University of California, San Diego, for the Hubble Space Telescope Project. His main current research interests are in the intersecting areas of digital libraries, hypermedia and networked knowledge organisation systems (NKOS). He co-organised NKOS workshops at ECDL2000 and ECDL2003. He directed the EPSRC-funded FACET project, which investigated the integration of thesauri into the search interface, together with semantic query expansion techniques for searching and browsing. This project was carried out in collaboration with the UK

National Museum of Science and Industry and the J. Paul Getty Trust. Other research interests include the application of interactionist social science perspectives to HCI and Participatory Design. Since 1977, he has been Editor of the journal *New Review of Hypermedia and Multimedia* (formerly *Hypermedia*; <http://www.tandf.co.uk/journals/titles/13614568.asp>). He serves as a reviewer for various journals and international programme committees. See: <http://www.comp.glam.ac.uk/pages/staff/dstudhope/>

Pavel Zezula, Faculty of Informatics, Masaryk University, Czech Republic

Pavel Zezula is a professor of informatics at the Faculty of Informatics, Masaryk University, Brno, Czech Republic. Currently, he is also the vice-dean for research in the same faculty. His professional interests are centred on storage structures and algorithms for content-based retrieval in non-traditional digital data types and formats, such as similarity search and the exploration of XML structured data collections. He has a long record of co-operation with the CNR in Pisa, Italy, and has participated in several EU-funded projects: MULTOS (ESPRIT No. 28: Multimedia Office Server), FIDE (ESPRIT BRA No. 3070: Formally Integrated Data Environment), HERMES (ESPRIT LTR No. 9141: Foundations of High Performance Multimedia Information Management Systems), SCHOLNET (IST-1999-20664), DELOS Working Group (ESPRIT LTR No. 21057) and DELOS NoE (IST-1999-12262). He has been Programme Committee Member (Chairman) of several prestigious international conferences (EDBT, VLDB, ACM SIGMOD). He has publications in top-level journals (ACM TODS, ACM TOIS, *VLDB Journal*), books, and conference proceedings (VLDB, ACM PODS, EDBT).



DIGICULT: PROJECT INFORMATION

DigiCULT is an IST Support Measure (IST-2001-34898) to establish a regular technology watch that monitors and analyses technological developments relevant to and in the cultural and scientific heritage sector over the period of 30 months (03/2002-08/2004).

In order to encourage early take-up, DigiCULT produces seven Thematic Issues, three Technology Watch Reports, along with the e-newsletter DigiCULT.Info.

DigiCULT draws on the results of the strategic study 'Technological Landscapes for Tomorrow's Cultural Economy (DigiCULT)', which was initiated by the European Commission, DG Information Society (Unit D2: Cultural Heritage Applications) in 2000 and completed in 2001.

Copies of the DigiCULT Full Report and Executive Summary can be downloaded from or ordered at: <http://www.digicult.info>.

For further information on DigiCULT please contact the team of the project co-ordinator:

Mr Guntram Geser,
guntram.geser@salzburgresearch.at
Tel: +43-(0)662-2288-303

Mr John Pereira,
john.pereira@salzburgresearch.at
Tel: +43-(0)662-2288-247

Salzburg Research Forschungsgesellschaft
Jakob-Haringer-Str. 5/III
A - 5020 Salzburg Austria
Tel: +43-(0)662-2288-200
Fax: +43-(0)662-2288-222
<http://www.salzburgresearch.at>

Project Partner:

HATII - Humanities Advanced Technology and Information Institute
University of Glasgow
<http://www.hatii.arts.gla.ac.uk/>
Mr Seamus Ross, s.ross@hatii.arts.gla.ac.uk

The members of the Steering Committee of DigiCULT are:

Philippe Avenier, Ministère de la culture et de la communication, France
Paolo Buonora, Archivio di Stato di Roma, Italy
Costis Dallas, Critical Publics SA, Greece
Bert Degenhart-Drenth, ADLIB Information Systems BV, The Netherlands
Paul Fiander, BBC Information & Archives, United Kingdom
Peter Holm Lindgaard, Library Manager, Denmark
Erich J. Neuhold, Fraunhofer IPSI, Germany
Bruce Royan, Concurrent Computing, United Kingdom



DigiCULT Thematic Issue 1 – Integrity and Authenticity of Digital Cultural Heritage

Objects builds on the first DigiCULT Forum held in Barcelona, Spain, on 6 May 2002, in the context of the DLM Conference 2002.

DigiCULT Thematic Issue 2 – Digital Asset Management Systems for the Cultural and Scientific Heritage Sector builds on the second DigiCULT Forum held in Essen, Germany, on 3 September 2002, in the context of the AIIM Conference @ DMS EXPO.

DigiCULT Thematic Issue 3 – Towards a Semantic Web for Heritage Resources builds on the third DigiCULT Forum held on 21 January 2003, at Fraunhofer IPSI, Darmstadt, Germany.

DigiCULT Thematic Issue 4 – Learning Objects from Cultural and Scientific Heritage Resources builds on the fourth DigiCULT Forum held on 2 July 2003, at the Koninklijke Bibliotheek – National Library of the Netherlands, The Hague.

DigiCULT Thematic Issue 5 – Virtual Communities and Collaboration in the Heritage Sector builds on the fifth DigiCULT Forum held on 20 October 2003, at Napier University, Edinburgh, Scotland.

DigiCULT Thematic Issue 6 – Resource Discovery Technologies for the Heritage Sector builds on the sixth DigiCULT Forum held on 9 March 2004, at the Archivio di Stato di Roma, Rome, Italy.

IMPRINT

This Thematic Issue is a product of the DigiCULT Project (IST-2001-34898).

Editors:

Guntram Geser and John Pereira, Salzburg Research

Authors:

Guntram Geser, Salzburg Research
Joemon M. Jose, University of Glasgow
Joost van Kasteren, Journalist
John Pereira, Salzburg Research
Michael Steemson, Caldeson Consultancy
Douglas Tudhope and Ceri Binding,
University of Glamorgan

Images:

Images on pages 24 and 25: *Courtesy of University of Glasgow, Department of Computing Science, Information Retrieval Group. Used with permission.*

Image on page 31: *Courtesy of University of Glamorgan, School of Computing, Hypermedia Research Unit. Used with permission.*

All other images: *Photographs taken by Birgit Retsch at the Archivio di Stato di Roma. Courtesy of Salzburg Research. Used with permission.*

Graphics & Layout:

Jan Steindl, Salzburg Research

ISBN 3-902448-03-2

Printed in Austria.

© 2004

DigiCULT Publications offer a valuable resource of mission-critical information in the selection and use of digital technologies for Europe's heritage organisations

Resource Discovery Technologies for the Heritage Sector:

This sixth Thematic Issue concentrates on how resource discovery technologies can ensure that the high value, authoritative information of heritage institutions is effectively found, retrieved, and presented to Internet users. With a key focus on the user, the Issue looks into user-driven approaches in interactive resource discovery. Expert opinion suggests that offering easy to use services and tools able to integrate the research and learning needs and behaviours of their users may form one of the heritage institutions' answers to the dominance of general-purpose global search engines.

However, along with ensuring state-of-the-art interactive access and presentation, the heritage sector will also need to raise the public's awareness to, and visibility of, its online resources in a more profound manner. Otherwise it faces the risk that the large investment required in creating digital collections, rich descriptive metadata, study and learning material, will fail to realise a high return – in terms of interest and appreciation, discovery and valuable uses of heritage resources.

DigiCULT Consortium:

salzburg | research



UNIVERSITY
of
GLASGOW

www.digicult.info

ISBN 3-902448-03-2