Proceedings of the Workshop on the Reuse of Web-based Information

Edited by
Anne-Marie Vercoustre,
Maria Milosavljevic and
Ross Wilkinson

Report Number CMIS 98-111
June 1998

CSIRO Mathematical
and Information Sciences
723 Swanston St. Carlton VIC3023
Telephone: (03) 9282 2610
Facsimile: (03) 9282 2600
www.cmis.csiro.au

Proceedings of the Workshop on Reuse of Web-based Information

Held in Conjunction with
The Seventh International World Wide Web Conference

Brisbane, Australia --- April 14, 1998

WELCOME TO THE GATHERING PLACE

Edited by:

Anne-Marie Vercoustre, INRIA, France, Anne-Marie.Vercoustre@inria.fr
Maria Milosavljevic, CSIRO-CMIS, Australia, Maria.Milosavljevic@cmis.csiro.au,
Ross Wilkinson, CSIRO-CMIS, Australia, Ross.Wilkinson@cmis.csiro.au

CSIRO Report Number CMIS 98-111
June 1998

URLs:
Reuse Workshop: www.mel.dit.csiro.au/~vercous/REUSE/WWW7-reuse.html
WWW7 Conference: www7.conf.au
CSIRO-CMIS: www.cmis.csiro.au
INRIA: www.inria.fr

Workshop Theme:

The workshop is focusing on the *reuse* of information that is currently delivered over the Web. By reuse, we mean the assembling of information from Web-based sources, and the restructuring of that information to either build a new Web application or to import it into a non-Web-based application, such as a database. The target research areas include (but are not limited to):

- mining, exploring and visualizing the Web,
- the structural extraction of relevant information, e.g., the Web as a large database,
- the semantic extraction of relevant information, e.g., the role of natural language processing and intelligent agents,
- the a posteriori restructuring or remodelling of Web pages and sites,
- the extraction of objects from the Web and their integration with legacy applications, or with other sources of information.
- the seamless integration of Web-based information, e.g., the design of information 'brokers'
- the development of descriptive models for Web information, and
- the role of XML in Web information reuse applications.

As the Web has grown the reuse of Web-based information has become increasingly important. Much of the existing research has focused on related issues such as 1) improving information search and retrieval capabilities, 2) enhancing the display of information in browsers, and 3) designing Web interfaces for legacy applications.

Advances in these areas have enabled a massive amount of information to be made available on the Web. The point is that people do not search for information just for the sake of it; they search for information because they need it and want to make use of it. The real challenge is to make the best use of that information since desired information is often distributed piecemeal among a number of sites, servers, and pages, each with its own disparate structure. The goal of reusing Web-based information is to extract and assemble existing information that a user is interested in and to deliver that information in a form that is directly usable by users or applications. Full automated access to Web servers makes for some fairly compelling applications, e.g. integration of corporate purchasing systems with e-commerce Web sites.

This workshop aims to draw together a number of research groups taking different approaches to the reuse of Web-based information, in order to promote the cross-fertilisation of ideas and highlight the prospects for future collaboration. The workshop will be a inter-disciplinary exploration into the reuse of Web-based information in controlled environment such as Intranets, as well as any part of the whole Web. Reuse of Web information in various contexts contributes to reducing the duplication of the information, while keeping it adapted to the real needs of the users.

Acknowledgment

We thank INRIA, France and CSIRO-Mathematics and Information Science, Australia, for supporting the chair and co-chairs in the organisation of this workshop, and the WWW7 Conference Committee for managing the registration and all the facilities.

Program Committee:

- Thomas Baker, Asian Institute of Technology, Thailand
- Stephen Green, Microsoft Research Institute, Language Technology Group, Sydney, Australia.
- Curtis Dyreson, James Cook University, Townsville, Australia
- Craig Lindley, CSIRO-Mathematics and Information Science, Sydney, Australia.
- Udi Manber, University of Arizona, USA
- Giansalvatore Mecca, Universita' della Basilicata and Universita' di Roma TRE, Italia
- Philipp Merrick, webMethods, Fairfax, VA, USA
- Pr. Alberto O. Mendelzon, University of Toronto, Computer System Research Institute, Canada
- Maria Milosavljevic (co-chair), CSIRO-Mathematics and Information Science, Sydney, Australia.
- David Levy, Xerox PARC, Pao Alto, USA.
- Pr. Leon Sterling, University of Melbourne, Intelligent Agent Laboratory, Australia
- Ross Wilkinson (co-chair), RMIT and CSIRO-Mathematics and Information Science, Melbourne, Australia.
- Anne-Marie Vercoustre (chair), INRIA-Rocquencourt, France, (and CSIRO-Mathematics and Information Science, Australia)

List of Participants

- **Lisa Casey**, City of Sydney, Town Hall House, Sydney, Australia, lcasey@cityofsydney.nsw.gov.au,
- **Kerry Cody**, Department of Primary Industries, Brisbane, Australia, codyk@dpi.qld.gov.au,
- **Joseph Davis**, University of Wollongong, Australia, joseph_davis@uow.edu.au, (accepted paper)
- **Petros A. Demetriades**, King's College London, UK, petros@dcs.kcl.ac.uk (accepted paper),
- **Curtis Dyreson**, James Cook University, Australia, curtis@cs.jcu.edu.au, **PC member**, (accepted paper),
- **Carlos F. Enguix**, University of Wollongong, Australia, cfe01@wumpus.uow.edu.au (paper accepted),
- **Brendan Hills**, CSIRO-CMIS, Sydney, Brendan.Hills@cmis.csiro.au,
- **Tanja Joerding**, Institute for Softwaretechnology, Dresdes, Germany, tj4@inf.tu-dresden.de,
- **Ralph Lawson-Smith**, Radiology Westmead Hospital, Westmead NSW, Australia, irls@imag.wsahs.nsw.gov.au,
- **Richard Lin**, Monash University, Australia, rlin@indy10.cs.monash.edu.au, (accepted paper),
- **Craig Lindley**, CSIRO-CMIS, Sydney, Australia, Craig.Lindley@cmis.csiro.au, **PC member**, (accepted paper),
- **Kim Marriott**, Monash University, Victoria, Australia, marriott@cs.monash.edu.au (accepted paper),
- **Stef Miladinova**, Rochedale South, Qld, Australia, smiladin@writeme.com,
- **Maria Milosavljevic**, CSIRO, Sydney, Australia, Maria.Milosavljevic@cmis.CSIRO.AU, **(Co-Chair)**,
- **Matthew Montebello**, University of Wales, Cardiff, U.K., m.montebello@cs.cf.ac.uk, (accepted paper),
- **Ms. Hiromi Morikawa**, JSRD, Shinjuku-ku, Tokyo, hmorikawa@jsrd.or.jp,
- **Max Mühlhäuser**, University of Linz, Austria, max@tk.uni-linz.ac.at (accepted paper),
- **Judith Nicholls**, Fermilab, Batavia, IL, U.S.A, nicholls@fnal.gov
- **Kee Ong**, NetFranchise, Burlingame, California, USA, keeong@pacbell.net
- **Meg O'Reilly**, Southern Cross University, Lismore, Australia, moreilly@scu.edu.au,
- **Nadine Ozkan**, CSIRO-CMIS, Sydney, nadine.ozkan@cmis.csiro.au,
- **Anand Rajaramon**, Junglee Corp., Sunnyvale, CA, USA, Anand@junglee.com (accepted paper)
- **Joanna Richardson**, Bond University, QLD, Australia, richardj@bond.edu.au,
- **Francois Rouaix**, INRIA, France, Francois.Rouaix@inria.fr,
- **Philipp Ruetsche**, Web Office ETHZ, Zurich, Switzerland, ruetsche@sl.ethz.ch
- **Simon Ryan**, Telstra Research Laboratories, Clayton, VIC, Australia, s.ryan@trl.oz.au,
- **Bill Simpson-Young**, CSIRO, Sydney, Australia, Bill.Simpson-Young@cmis.csiro.au
- **Jim Stinger**, Hewlett-Packard, Palo Alto, USA, stinger@hpl.hp.com (accepted paper),
- **Martin Ström**, LinnéData AB, Sweden, martin.strom@lig.linnedata.se,
- **Anne-Marie Vercoustre**, INRIA, France, Anne-Marie.Vercoustre@inria.fr, **(Chair)**,
- **Ross Wilkinson**, CSIRO, Melbourne, Australia, Ross.Wilkinson@cmis.csiro.au, **(Co-chair)**

Workshop Program:

Tuesday, 14th April

9:00-9:15 **Welcome and Introduction**
Anne-Marie Vercoustre

9:15-10:40 **Session 1: Searching for Information**
Moderator: Ross Wilkinson

A personal Evolvable Advisor for WWW Knowledge base Systems
M. Montebello, W.A. Gray, S.Hurley

A Jumping Spider: Restructuring the WWW Graph to Index Concepts that Span Pages
Curtis E. Dyreson

Transforming the Internet into a Database
Anand Rajaraman

10:40-11:10 *Break*

11:10-12:10 **Session 2: Extracting Information**
Moderator: Anne-Marie Vercoustre

Component Advisor: A Tool for Automatically Extracting Electronic Component Data from Web Datasheets
Malu Castellanos, Qiming Chen, Umesh Dayal, Meichun Hsu, Mike Lemon, Polly Siegel, Jim Stinger

Database Querying on the World Wide Web: UniGuide- An Object-Relational Search Engine for Australian Universities
Carlos F. Enguix, Joseph G. Davis, and Aditya K. Ghose

12:10-13:30 *Lunch*

13:30-14:30 **Session 3: Invited Talk, by Philipp Merrick, webMethods, U.S.A** .
Replaced at the last minute by another talk:

XML: The Key for Reuse?, Anne-Marie Vercoustre

14:30-15:30 **Session 4: Modelling Information**
Moderator: Curtis Dyreson

The FRAMES Project: Reuse of Video Information using the World Wide Web
C. A. Lindley, B. Simpson-Young, and U. Srinivasan

Typing Concepts for the Web as a Basis for Re-use
Max Mühlhäuser, Ralf Hauber, Theodorich Kopetzky

15:30-16:00 *Break*

16:00-17:20 **Session 5: Presenting Information**
Moderator: Maria Milosavljevic

An Architecture for Web Visualisation Systems
Petros A. Demetriades and Alexandra Poulovassilis

The Role of Reactive Typography in the Design of Flexible Hypertext Documents
Rameshsharma Ramloll

Using Constraints for Flexible Document Layout
Alan Borning, Richard Lin, Kim Marriott and Peter Stuckey

17:20-17:30 **Closing**
Moderator: Craig Lindley

Report on the workshop

The workshop attracted about 25 people, mostly from Australia but also from the USA, UK, Austria, Germany and France. It was combined with the **workshop on Flexible Hypertext**.

The workshop included four main sessions, an invited talk and a concluding session. The first three sessions had a theme corresponding to the main steps in reusing information, namely:

- Searching and finding information
- Extracting information
- Combining/Presenting information

while a forth theme, *modeling Information*, was regarded as central to the whole issue of reuse.

The steps involved in the process of information reuse were highlighted in Anne-Marie's introduction. She also pointed out that extracting information may amount to following links among pages to find the relevant fragments or further links. Therefore searching and extracting are tightly related and the allocation of papers to one of the other session was somehow arbitrary.

Modeling information appeared to be a condition for extracting and reusing. The two papers in this session considered a priori modeling for supporting future reuse, rather than using a model for discovering and extracting information, like in some other papers (eg. A Jumping Spider and papers in session 2).

The last session focused on the flexible presentation of information, which is another important aspect in the reuse of existing information.

Unfortunately, for personal and professional reasons, Philipp Merrick from webMethods was unable to attend the workshop and the WWW7 Conference. At the last minute his presentation on using XML for e-business was replaced by an introduction to XML and its potential benefit for reuse, by Anne-Marie Vercoustre. Her presentation focused on the research perspective rather than the business perspective on XML.

After drawing some conclusions we had an unplanned wrap-up session with the Workshop on Hypertext Functionality and the WWW, co-chaired by Carolyn Watters and Fabio Vitali. This also resulted on a BOF session on Virtual Documents, and the formation of a special interest group on this topic.

Finding/Searching Information

In order to reuse information, we first require access to that information, either by: finding out what is relevant (by searching); accessing well-known places where information is regularly updated; being notified that new information has come; or a combination of the above.

Matthieu Montebello underlined the inefficiency of search engines and browsing regarding precision and recall, and arose the issue of such evaluation criteria in a world-wide environment. He focused on meeting user's need and interest and proposed a system that prompts the user with new URLs selected by peers, according to a dynamically evolving user's profile.

Curtis Dyreson made the distinction between *static* and *dynamic* approaches for search tools. A static search is done at periodic intervals, off-line, for example by a robot, and results are stored locally for further querying and reuse. A dynamic search occurs when a tool crawls part of the Web on-line for extracting information each time the information is requested. Curtis described a static system to index concepts that span more than one page.

Anand Rajaman presented the architecture of a commercial system developed by Junglee and used for integrating book offers from various electronic bookshops, job opportunities, etc. It was really "reuse at work". The system includes a tool box for building wrappers on sites, linguistic feature extraction, as well as a data verifier that checks the conformance of data type to expectations, in case the server's structure had changed without notification.

General discussion

A first question was: is searching for reuse different than usual searching?
One partial answer is that searching for reuse is in the context of a specific reuse task, rather than the user's task.
We had a long discussion on evaluation of search and reuse in the context of the Web, compared to TREC evaluation.
The problems include:
- the validation of data: TREC evaluation works on a well-established and fixed corpus. For web reuse there is a need for testing the validity (stability) of data for each wrapper (agent), regarding both the structure and the content.
- the nature of the web, where there may be a tension between "relevance" for a customer and "attractiveness" for the vendor. So the question "what is relevance" or "meaning" (eg. for images) is yet more complex than for classical IR.

Evaluation should:
- be user-based
- set the task involved in reuse (relevance for reuse)
- evaluate the efficiency (how much reuse is possible?), regarding the instability of data.

Extracting Information

Jim Stinger presented a tool for automatically extracting electronic component data from web datasheets. His system does not rely on the data format (tags) for extracting information. Instead it uses domain knowledge about products, and vendor specific catalogues which model the knowledge embedded in particular vendors' datasheets. The navigator visits vendors' pages, traverses along promising links or fills forms as required, and returns a list of pages to the extractor. The extractor iteratively processes the pages for each characteristic to be mined from the datasheet and sends them to the database. The difficulties encountered in automatic mining were many due to poorly written HTML pages and inconsistency between datasheets versions.
The second paper, presented by Joseph Davis, also uses a domain centered model to support more precise access to information. The approach here is quite different though, since it uses complex metadata based on ontologies. The success of the approach is dependent on a given number of sites relevant to the domain adopting the use of the UniGuide Scheme metadata and to attach these metadata to their pages. This approach is in the line of current initiatives for metadata development, like the Dublin Core initiative, or the W3C working group on RDF schema registery.
The two presentations highlighted the need for making the underlying model of the application explicit, since it is often difficult to infer it from the structure of the site(s).

Modeling Information

The 2 papers in this session are concerned with modeling information for supporting better reuse, the first one for reusing digital videos, the second for reusing Web styles and instances.
Craig Lindley introduced FRAMES, a system for reusing digital videos through synthesised new videos. FRAMES uses a rich model for describing videos which reflects the complexity of video semantics (multiple possible interpretations) as well as physical features. Although the query engine and content models can be used directly for search and retrieval of video content, the important point, with regard to reuse of the content, is that it can also be used for a wide variety of other purposes, for example, virtual video generation. A video synthesis can be specified by a virtual document prescription which contains queries to the video database using a combination of features available in the model.
Max Mühlhäuser focussed on hyperwebs, ie. a logically correlated set of nodes and links. He showed how *typing concepts* for hyperwebs can improve finding information as well as reusing such hyperwebs in augmented hypermedia. He introduced graph grammars as a way of typing the Web in the programming language sense. With this formalism, models and instances of hyperwebs can be precisely described and reused. This work is strongly rooted in hypertext/hypermedia research and deserved to find a better place in the Web community especially for intranet design.
Both presentations demonstrated the importance of a rich description of the primary sources (content, structure) in order to support effective reuse.

Presenting Information

We did not have specific papers on reassembling/reconstructing information, although several of the participants are working on Virtual Documents, which were discussed in a separate BOF session (see below). Virtual documents were also discussed by Craig Lindley in his presentation on building video synthesis.

Petros Demetriades presented an architecture for building various views on Web Information, as well as on browsing history. Views perform the actual interpretation of the data into some visual form. This architecture can accommodate both static and dynamic search, and facilitate experimentation and side-by-side comparison of different collection, storage and visualisation methods and techniques.

Kim Mariott showed the advantages of using constraint-based techniques for the automatic generation of good document layout. Documents and Web pages are made of content, structure and layout. Both structure and layout can change according to reuse content and context. Layout should adapt itself to a wide variety of environments, medias and viewers. Constraint-based layout provides this flexibility and supports a real *negotiation* between the author and the viewer for controlling the final appearance of the document.

Some Conclusions

Conclusions drawn by Craig Lindley were:

1. clarifying the notion of reuse: making a distinction between "use" of information, "reuse", and multi-purpose information. Reuse often amounts to restructuring or adding-value to primary sources (second level information), or use it for other purposes than the one initially intended. Multi-purpose information refers to flexible information delivery to accommodate various contexts. Reuse here is anticipated.
2. One particularly interesting issue was that the context of "use" is not necessarily the "user"

context, and these should be treated separately. There is a need for evaluation of search and reuse relevance and effectiveness, according to a context or a task. Relevance regarding meaning is difficult, especially for images which may have several *meanings*. Relevance can be quite different for a vendor, or a customer, and for entertainment.
3. The conclusions also highlighted the need for research on meta models for structure and semantics and a language for virtual inclusion/transclusion.

Wrapping session

Another workshop which was running concurrently dealt with the hypertext functionality of the web and where we might like the web to be enhanced. For example, in transclusion, instead of including a link to a piece of information, that information might be included in the document on-the-fly in order to achieve some purpose. This is like the traditional "stretch-text" idea where pieces of additional text are weaved into an otherwise static document for some reason, such as the user not understanding a particular term etc. We invited the members of this workshop to join us for the final wrap-up session and a member from each workshop gave a summary of the issues which had arisen over the course of the day.
Requirements for the Web (from this point of view):

1. The Web needs to enable typed links. Examples of types of links are: annotations, transclusion (inclusion by reference), personalized links, trails and guided tours, backtracking and history-based navigation etc. Currently, the Web only offers one type of link. For the hypertext community, this is insufficient.
2. Web applications need to be evaluated more rigorously. Evaluation is likely to become a big theme, as no one does it (or does it well) and as the need is clearly there.
3. The Web needs to provide facilities that support collaboration.
4. More user-centered approaches to web design are needed.

This allowed us to see the links (pardon the pun) between the two groups more clearly and led to the formation of a "Birds of a Feather" session later in the conference.

This BoF session focussed on "virtual documents" in order to discuss at some length some of the issues we consider important in building systems which can produce documents on-the-fly from either large pieces of text or smaller fragments using a grammar etc. We spent some time discussing the meaning of the term "virtual document" but didn't reach a conclusion in one hour.

Further actions

Participants to the BoF session agreed on further collaboration about Virtual Documents, starting with be setting up a mailing list and a web page for those people who are working in this area in order to pool ideas and constraints, and to facilitate collaboration. For more information please to contact Brendan Hills: Brendan.Hills@cmis.csiro.au

Table of Content