Knowledge Media Institute

Evolving the Web for Scientific Knowledge:
First Steps Towards an "HCI Knowledge Web"

Simon Buckingham Shum

KMI-TR-68

December, 1998

Interfaces, British HCI Group Magazine, No. 39 (Dec., 1998), pp.16-21.

Interfaces, British HCI Group Magazine, No. 39 (Dec., 1998), pp.16-21.

Evolving the Web for Scientific Knowledge:
First Steps Towards an "HCI Knowledge Web"

Simon Buckingham Shum

Knowledge Media Institute, Open University, Milton Keynes, MK7 6AA, UK
Email: S.Buckingham.Shum@open.ac.uk
WWW: kmi.open.ac.uk/sbs

Summary: In this article, I consider the challenge of building a Web-based infrastructure for scholarly research which moves beyond the basic dissemination and linking of documents, to support more powerful searching and analysis of the cumulative knowledge in the literature's documents. Taking the HCI research community as an example, the goal would be to enable HCI researchers to search for interesting documents and phenomena, and discover previously unknown but conceptually related research, for instance, other groups addressing persistent problems in the field, the structure of debates, or when and how new theoretical perspectives began to make an impact. I propose that focusing on the scientific relationships between documents is important, and has advantages as the basis for a Web metadata scheme to enrich the HCI community's Web.

Your desktop, in the not too distant future...

You are starting a new HCI research project, and want to find out what's been done so far. (You can hardly believe it, but 2 years ago, you would have had to search the Web, or one of the few HCI digital libraries, using basic keywords. The servers and search engines knew nothing about how HCI research is conducted and so could provide no assistance. Documents were not described in any machine-readable form other than keywords, and were not linked in any way beyond citations.)

You connect to your local server in the HCI Knowledge Web, and issue queries for the following:

- **documents/websites using or extending the StarViz software system**
- **documents analysing the applied problem of visualizing large datasets in astronomy**
- **documents building on a particular theoretical framework of interest to you, and extending the RouteFinder class of graphing algorithms**
- **documents challenging evidence that the RouteFinder class does not scale up**
- **documents problematising a methodology closely associated with StarViz.**

It's been said many times in recent years, and it's still true:

> *The Net, particularly the Web, provides an unprecedented opportunity in scientific history to locate, interconnect and analyse ideas and documents.*

But…

> *The Web is becoming a more chaotic place by the day. As the signal to noise ratio gets worse, research communities need better support for tracking developments and finding relevant documents.*

It is currently impossible for search engines (Web or otherwise) to answer complex questions commonly posed by researchers such as the following:

- are there distinct schools of thought in this field?
- what impact did this evidence have?
- who is currently tackling this applied problem?
- has anyone built a system based on this theory?
- has anyone applied this theory to other fields?

The reason these questions are impossible to answer at present is that there must be a way to abstract meaningful patterns of documents. Thus, in relation to the first question above, we need to ask: what is a 'school of thought', how might it manifest itself in the literature, and are there corresponding patterns that could be detectable by a software agent to present to researchers as potentially significant? It is possible that useful information may be extracted through intelligent analyses of texts, but often this information is not explicit in documents, but implicit in the minds of domain experts. *Metadata* (introduced shortly) is an alternative way to provide such information (if experts encode it), but there is rarely metadata about scientific documents expressed using the conceptual language of that field. The Web now makes this technically possible—thousands of users can now contribute structured information to a shared, searchable repository.

But is a lingua franca (using metadata or otherwise) possible for a scientific research community, HCI in particular, what would constitute a good scheme, and what social and technical systems would need to co-evolve to make it sustainable? Let's begin by briefly considering what the HCI community has available to it today.

Today's HCI digital library

What is the state of the "HCI digital library" accessible over the net today? Within the HCI community, the pioneering work of Gary Perlman's *HCI Bibliography Project (HCI-Bib)* has made thousands of abstracts (some linked to other digital libraries) downloadable and searchable over the Web <www.hcibib.org>. Professional and learned

societies such as the ACM <www.acm.org/dl> and IEEE <computer.org/epub> are creating digital libraries providing subscriber access to many HCI-relevant journals and conferences (perhaps we can expect a BCS digital library soon?…). Most scientific publishers are now providing subscriber access to digital copies of journals. There is work on automatically linking citations to abstracts, although these depend on inter-publisher agreements (se sidebar). And of course, many workshops and individuals provide access to full papers.

Preprint servers provide repositories of technical reports, and if widely used within a community, are perhaps the best way to track new work (although unreviewed). The Los Alamos National Laboratory (LANL) preprint server <xxx.lanl.gov> set up by Paul Ginsparg, initially to serve the high energy physics community, has become the first place to publish new technical reports in that field (which are then replaced by the final versions when published – journals have been sidelined in this respect). Recently, a Computing Research Repository (CoRR) has been added to the LANL preprint server <xxx.lanl.gov/archive/cs/intro.html>, including an HCI subject area moderated by Terry Winograd. Preprint servers allow you to define interests using keywords, after which the server sends email alerts whenever new material is added (all of the above servers provide these). It will be interesting to see if the CoRR server achieves the same uptake as within the physics community (it may be that LANL's success derives from the premium on being the first to publish results in physics, arguably much higher than in computer science, or HCI). There used to be an HCI server at the The London & South-East Centre for High Performance Computing (SEL-HPC) but this appears to have ended (<www.lpac.ac.uk/SEL-HPC/> no longer works).

Scholarly publishers in the brave new world

The ideal of freely accessible information to all, whilst now practical at a technical and usability level with the arrival of the Web, is of course dogged by copyright restrictions. This is not the place to go into detail on this fraught topic, but suffice to note that electronic publishing has the potential to change the rules that bound researchers to publishers when they were wholly dependent on paper for dissemination. The Net provides the basis for scholars to forge new relationships with publishers, who may have to find new roles (some radical proposals and debate on this topic can be found in <cogsci.ecs.soton.ac.uk/~harnad/subvert.html>) . Nor is the simplistic equation that *the Net = unreviewed, low quality material* sustainable. It is peer review and other forms of quality control that add value and reliability, not the paper medium per se. Electronic journals are showing how the Web is well suited to scholarly publishing and peer review (e.g. <www-jime.open.ac.uk>).

These resources are a welcome alternative to having to order and wait for paper documents. But, they are just a start. Overwhelmingly, the Web as a resource for scientific knowledge is still serving as a searchable paper-publication resource, plus simple linkage. The Web's success reflects the power and attractiveness of this simple model, but it is an 'entry level hypertext system' in comparison to the power of a rich

hypertext which exploits machine-processable node and link semantics. Such semantic hypertexts were implemented in early research hypertext prototypes dating back to the mid-80s, based on cognitive science's concept of the semantic networks.

The key challenge for any effort to create a better system is deciding on the representational scheme to use (what should be the schema determining the node and link types?) and the usability of the system (how much effort is required to encode information using this scheme, and to subsequently interpret the system?). Is there a way to negotiate these inevitable overheads, in order to begin reaping the benefits of a more powerful system? I propose that metadata could be used, but in a novel way that differs from current metadata schemes.

Metadata schemes

Use of *metadata schemes* is one way to make the Web a semantically enriched hypertext. Here is some imaginary metadata for a document, using <angle brackets> to delimit each metadata field:

```
<TITLE=Unit 11: Knowledge Management Technologies>
<COURSE=B823>
<PRESENTATION=Nov1999>
<INSTITUTION=Open University UK>
<AUTHOR=Simon Buckingham Shum>
<CORE-CONCEPTS=knowledge, information, representations,
interpretation, technology, community of practice>
<BUILDS-ON=Q777, B823-Unit 2>
<PREREQUISITE-FOR=B888>
```

Note that some of the metadata tags simply describe the *content* of the document, whilst the last two actually describe particular kinds of *relationships* to others (eg. this document is not a PREREQUISITE-FOR Q777, it BUILDS-ON it). A search engine providing a query form with fields for these tags enables users to search specifically for documents which have "community of practice" as a CORE-CONCEPT, and are PREREQUISITE-FOR "B888".

We are seeing the emergence of W3C's XML scheme (a stripped down version of SGML) for adding one's own tags to text, initial work on the Resource Description Framework (RDF, for managing multiple metadata schemes), and internationally coordinated initiatives such as the Dublin Core metadata scheme, which provides a basis for communities to use or if desired, specialize their general scheme <purl.oclc.org/metadata/dublin_core>. Coupled with toolkits such as ROADS for structuring information gateways <www.ukoln.ac.uk/roads>, and protocols such as Z39.50 <lcweb.loc.gov/z3950/agency> for distributed searching of servers, we have the emerging basis for more powerful infrastructures for content discovery.

Not surprisingly, the library and information sciences are leaders and early adopters of metadata, given their interest in classification and their already large document repositories. But other research fields are initiating consortiums for resource description,

e.g. the Instructional Management System for online educational resources <http://www.imsproject.com>.

Metadata focused on scientific relationships

It is striking to note that in most metadata schemes, *relational* information (how does this document relate to others?) tends to be the poor cousin of *content* information (what's in this document?). This may be because much of the work to date has been driven by library/information scientists. However, *relationships* are critical for researchers, who invest a lot of energy in articulating and debating different claims about the significance of conceptual structures. Moreover, it is often precisely the issue of how to describe the status of a document or idea that is under debate in research—an approach which allows only one way to encode material will fail to meet the needs of a community which is constantly contesting claims.

A principle from hypertext research is to avoid loading *nodes* (e.g. web documents) with semantic information (e.g. metadata encoding), and focus instead on the *links*. That way, a given node remains 'neutral' on its own, but can be referred to in many different ways by different authors; it is its place in the network which determines its role and interpretation. It may be, therefore, that the generic *relation* field in a Web metadata scheme such as Dublin Core could provide the anchor that researchers need to specialise into a set reflecting the important relationships in their field.

A metadata scheme grounded in concepts and relationships for scholarly discourse would, for instance, provide a way to 'semantically tag' keywords and references. Consider for instance, how you choose keywords for your papers. They typically reflect many different conceptual relationships. Instead of an undifferentiated list, very different keywords such as "Java" and "Situated Cognition" could be tagged to indicate (to a software agent) that they refer to *software* and a *theory/framework*, respectively. Another example: both human and software agents would be interested to know that a citation to "Smith, 1998" is not evidence of its reliability (the implicit interpretation of science citation indices), but that it is *problematising* the *method* used in that paper.

I therefore propose that a more tractable goal is a scheme which reflects *the WAY in which HCI research discourse proceeds* as a discipline, focusing not on encoding the content of documents (other techniques exist for doing this automatically), but on the *scientific relationships between documents* that are hard, if not impossible, to infer automatically. Consider familiar relationships between papers in the literature such as *modifies, describes, supports, problematises.* These verbs are commonly used in conjunction with concepts such as *applied problem, theory/framework, software, evaluation, trends.* I suggest that these are relatively uncontentious and stable—they are how we think about documents and their inter-relationships.

A scheme based on accepted scientific relationships is less brittle than classification schemes which seek to reflect key subject matter in the field, but which require regular updating (cf. the ACM 1998 Computing Classification Scheme for keywords <www.acm.org/class/1998>).

So, what might a scholarly metadata scheme—reflecting the *modes of discourse* common in HCI, not a master classification scheme—look like?

A possible HCI metadata scheme

Consider the following form which provides a way to construct common relationships between research documents— could you describe one of your publications using this scheme? It usually takes a little effort (perhaps productive) to distill the key contributions of a document, but informal testing has shown that most can be described using the constructs offered. Some examples follow the tables.

TITLE:
AUTHOR:
CITATION:
URL:
ABSTRACT:
KEYWORDS:

Key Contributions, and Relations to other work

From the following table, copy and paste a **RELATION**, add a **CONCEPT**, a **Description** of the Concept, and any **References/URLs** for the Concept**.**

Repeat until you are satisfied that you have summarised the document's key content and relationships to the existing literature.

RELATION	**CONCEPT**	**Description**	**Reference / URL**
Most Relations pair meaningfully with most Concepts. The next table summarises legitimate pairings.		*Name/ keywords for the* **CONCEPT**	*…for the* **CONCEPT**
ANALYSES	**APPLIED-PROBLEM**	…	…
SOLVES	**THEORETICAL-PROBLEM**	…	…
DESCRIBES-NEW	**METHOD**	…	…
USES/APPLIES	**LANGUAGE**	…	…
MODIFIES/EXTENDS	**SOFTWARE**	…	…
CHARACTERISES/RECASTS	**EVIDENCE**	…	…
EVALUATES	**THEORY/FRAMEWORK**	…	…
or more specifically:	**TREND**		
SUPPORTS	**SCHOOL-OF-THOUGHT**		
PROBLEMATISES			
CHALLENGES			

Most of the Relations can be sensibly combined with any of the Concepts, but the table below shows nonsensical combinations (dark cells)

CONCEPT: / **RELATION:**	PROBLEM	THEORY/ FRAME-WORK	LANGUAGE	SOFT-WARE	METHOD	EVIDENCE	TREND	SCHOOL-OF-THOUGHT
ANALYSES								
SOLVES								
DESCRIBES-NEW								
USES/APPLIES								
MODIFIES/EXTENDS								
CHARACTERISES/ RECASTS								
EVALUATES								
SUPPORTS								
PROBLEMATISES								
CHALLENGES								

Examples

Given the building blocks of the above metadata scheme, here are some fragments of metadata description to show its application.

ANALYSES APPLIED-PROBLEM Air traffic controller cognitive overhead
REF: Smith, J. (1997) ATC Overload. Journal of ATC, 3 (4), 100-150

USES/APPLIES THEORY/FRAMEWORK Situated Cognition, Activity Theory

DESCRIBES-NEW EVIDENCE use of video, undergraduate university physics, student ability

PROBLEMATISES SOFTWARE GOMS cognitive modelling tools

MODIFIES/EXTENDS LANGUAGE Knowledge Interchange Format (KIF)

CHARACTERISES/RECASTS TREND Electronic trading over the internet
REF:
REF:
REF:

CHALLENGES SCHOOL-OF-THOUGHT Postmodernism
REF:
REF:
REF:

SUPPORTS EVIDENCE multimedia, school chemistry teaching

Such fragments can be built into more complex structures. Returning to our earlier question about schools of thought, we might define a ‘structural signature’ in the document web which we would find interesting. We could therefore define and search for patterns in the literature (of encoded documents) which suggested the emergence of distinctive perspectives through a structural signature (perhaps graphically constructed) expressing the following: a ‘school of thought’ is a perspective, in contrast to at least one

other, on a common phenomenon. A perspective can be recognised by the common THEORY/FRAMEWORKS on which a group of researchers draws (size=N?), the associated METHODS and LANGUAGES which they deploy, and the body of EVIDENCE that they mutually support. Conversely, the set of THEORY/FRAMEWORKS, METHODS, LANGUAGES and EVIDENCE that they collectively CHALLENGE or PROBLEMATISE may represent a different perspective.

Within HCI, we might recognise several examples of perspectives that seem to fall into distinctive 'camps', building as they do on very different conceptual foundations. Consider ethnographic/sociological approaches as opposed to information processing approaches to studying the workplace. Or situated cognition and learning 'versus' symbolic AI perspectives on interaction. Or discount usability as opposed to cognitive modelling techniques. In summary, a school of thought may be declared by someone as shorthand in their description of their document, but such phenomena might also be detected as an emergent pattern within the literature.

Making it work: technical and social processes

As the Olde Englishe proverb goes, "a metadata scheme alone doth not a knowledge web make"—even if it does provide a successful lingua franca. There's no denying that many issues remain, which we are currently seeking resources to investigate. Interesting challenges that would quickly emerge if this initiative took off include:

User interfaces: for assisting in the construction of hypertext and metadata-based queries, interest profiles and agents, and the display of search results involving potentially large document sets and complex inter-relationships.

Managing terminological variations: the subtleties of language are important to researchers in expressing their ideas so the *relational types* in the metadata scheme need to be acceptable and document *content* indexing needs to cope with terminological variations. 'Bottom-up' information analysis techniques for analysing text corpuses (such as latent semantic indexing), and thesauri for synonym matching need to be used in synergy with 'top down' metadata (which provides the valuable relational knowledge of researchers).

Supporting emergent structures: a particular claim, or a network of ideas, may be asserted by one author, but what do others say? Mechanisms need to be worked out to enable detection of structural patterns that are widely subscribed to, and which therefore may be more credible than a 'lone voice' (e.g. a theory/framework which provides the motivation for subsequent work; a software system or design method that is widely used and extended). Authority naturally rests with more established researchers, so one would expect a facility to define filters and interest profiles which prioritise perspectives from particular people, research groups or institutions.

A lesson and guiding principle from the HCI Bibliography project is that any large scale, community-centered initiative must be realisable through a *little effort* from a *lot of people.* In scientific research, it is the *authors* who have the interest in maintaining the

'electronic visibility' of their work, to ensure accessibility and, hopefully, impact on the field. If individual HCI researchers took responsibility for enriching the descriptions of their own publications—the key content, and its linkage to other work—then a spectrum of new possibilities opens up, and we have an answer to the challenge of encoding documents. It is realistic to envisage connecting to a web server, completing a form based on a metadata scheme, and submitting it to the repository, which is mirrored around the world. Again, this is already standard practice in certain communities.

Invitation to participate in a pilot study

We can dream about the possibilities, but the first step is to pilot a metadata scheme to describe HCI publications, seeking the right balance of simplicity and expressiveness.

I therefore invite you to 'beta test' the above metadata scheme by using it to describe just one or two of your own HCI publications (initial design iterations have already been done). Please send me your form(s) and your feedback. The forms will undergo a preliminary analysis to assess the scheme's usability and the potential power of the information it generates. The more participants we have, the larger our testbed dataset for testing serverside tools and demonstration services such as alerting and visualizations (cf. interesting work by Chaomei Chen who has automatically generated VRML maps of ACM CHI and Hypertext proceedings <www.brunel.ac.uk/~cssrccc2>). We hope to combine such techniques with the approach described here, starting with the metadata from the pilot study.

This article, the metadata form and example metadata descriptions for a couple of HCI publications are on the *HCI Knowledge Web* pilot site at:

kmi.open.ac.uk/sbs/hciweb/pilot.html

From there, download the form as an RTF or HTML document and import into your wordprocessor:

kmi.open.ac.uk/sbs/hciweb/HCI-Pilot-Metadata.rtf

kmi.open.ac.uk/sbs/hciweb/HCI-Pilot-Metadata.html

Predicting the Web's evolution is a tricky business. But there's one thing we can be sure about: no-one will – or *can* – do it for the HCI community, but the HCI community.