Workshop on Social and Collaborative Construction of Structured Knowledge,
16th International World Wide Web Conference, Banff, Canada, May 8, 2007

Formalization, User Strategy and Interaction Design: Users' Behaviour with Discourse Tagging Semantics

Bertrand Sereno*, Simon Buckingham Shum & Enrico Motta
Knowledge Media Institute, The Open University, Milton Keynes, UK

* Now at: Centre for Advanced Learning Technologies, INSEAD, Fontainebleau, France


Licensed under Creative Commons
Attribution-ShareAlike 2.0 License

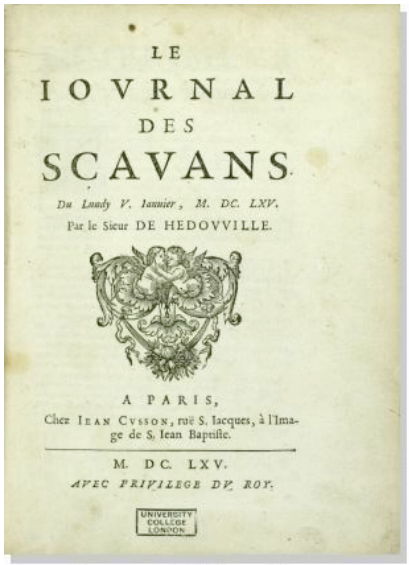
Acknowledgements:

Overview

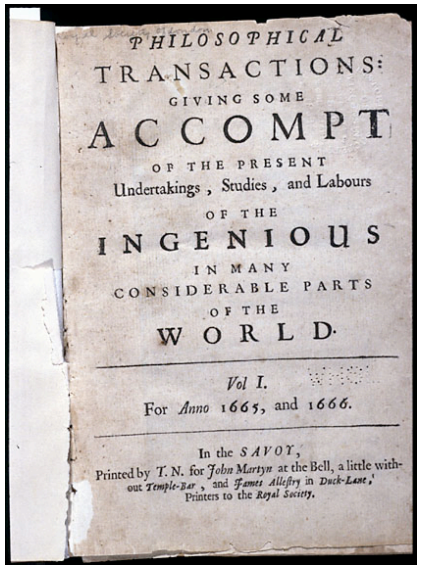
- **The problem:** collaboration semantics in contested domains — **no consensus assumed; possibly not even possible**
- **Previous work:** Scholarly discourse as Collaborative Knowledge Structuring (CKS)
  - *Modelling and querying Web argument structures*
- **How to** help users engage in CKS**?**
  - *Evaluating the ClaimSpotter tool*
- **Summary of** evaluation **results and** design principles
  - *Formalization / User Strategy / Interaction Design*

Scholarly discourse as CKS…

From:

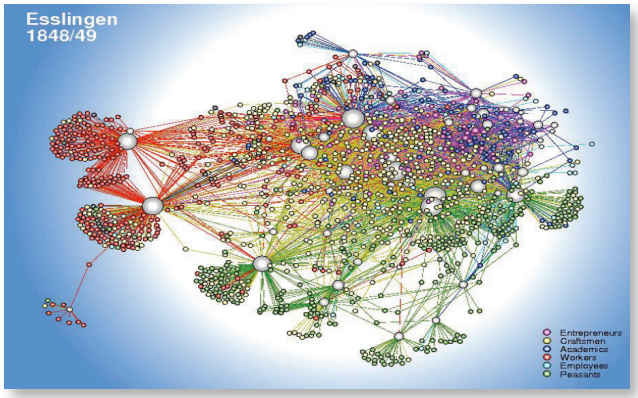

LE
IOVRNAL
DES
SCAVANS
Du Lundy V. Ianuier, M. DC. LXV.
Par le Sieur DE HEDOVVILLE.
A PARIS,
Chez IEAN CVSSON, ruë S. Iacques, à l'Image de S. Iean Baptiste.
M. DC. LXV.
AVEC PRIVILEGE DV ROY.
UNIVERSITY COLLEGE LONDON

Le Journal des Sçavans
January 1665

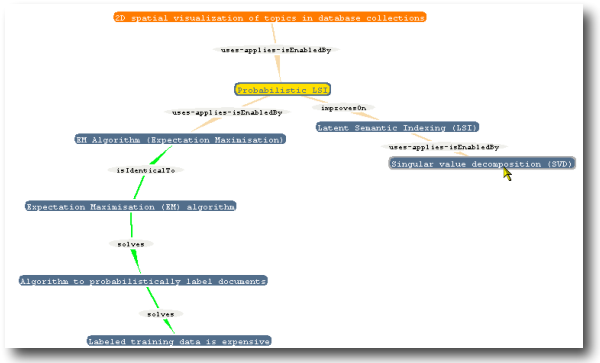

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

Philosophical Transactions of the Royal Society of London
March 1665

To:

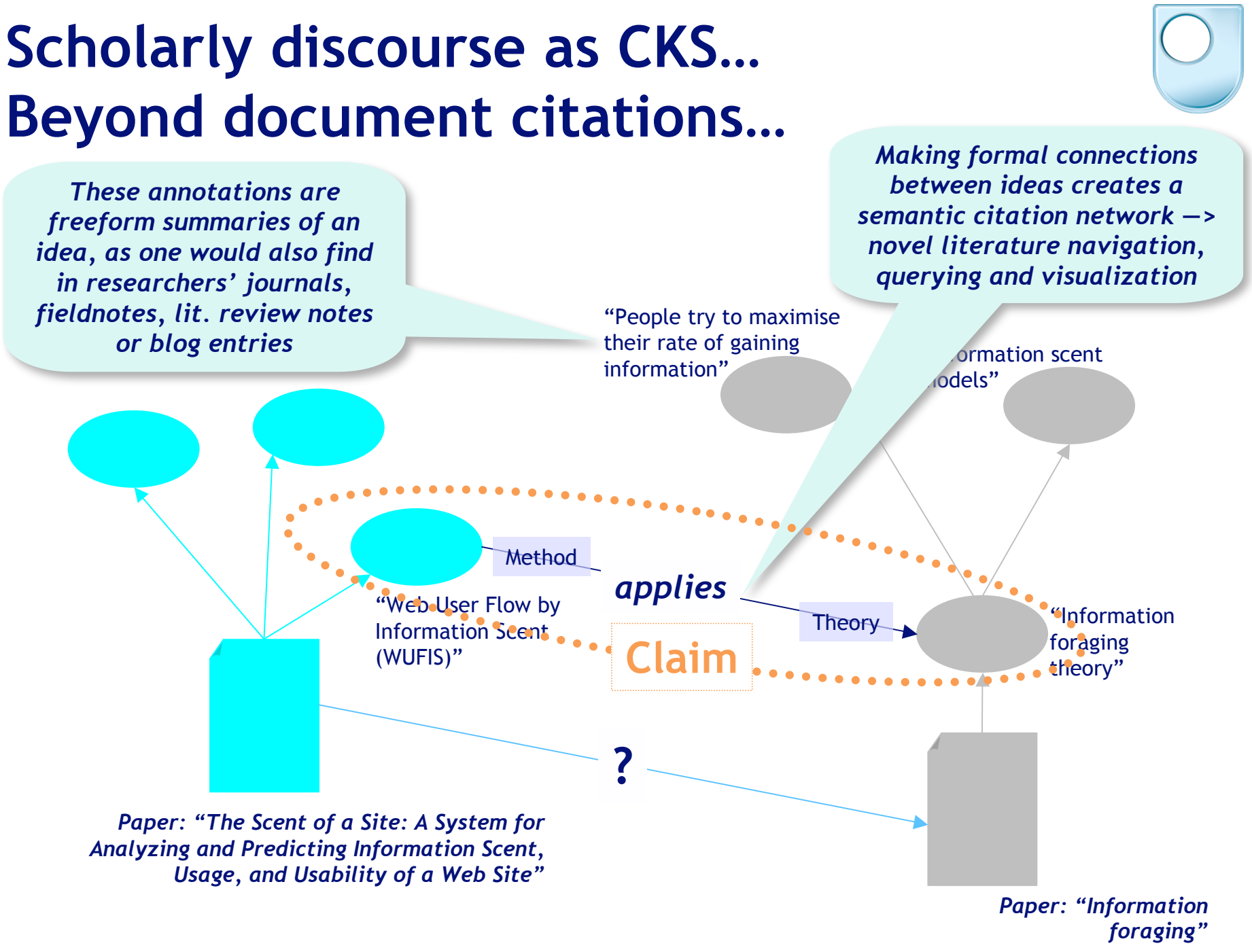


Chaomei Chen, 2006: Citation analysis



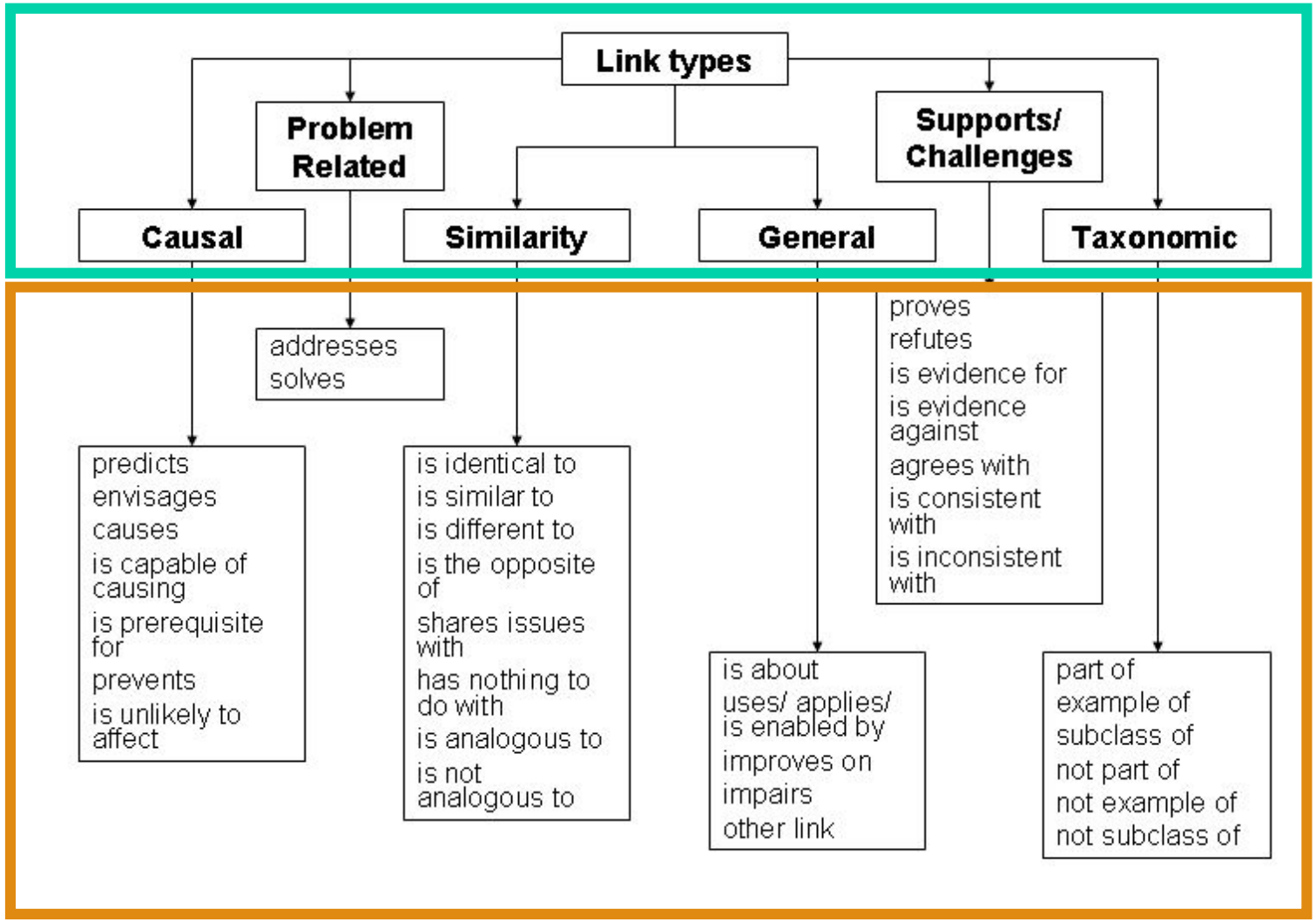
Buckingham Shum et al, 2003: lineage analysis

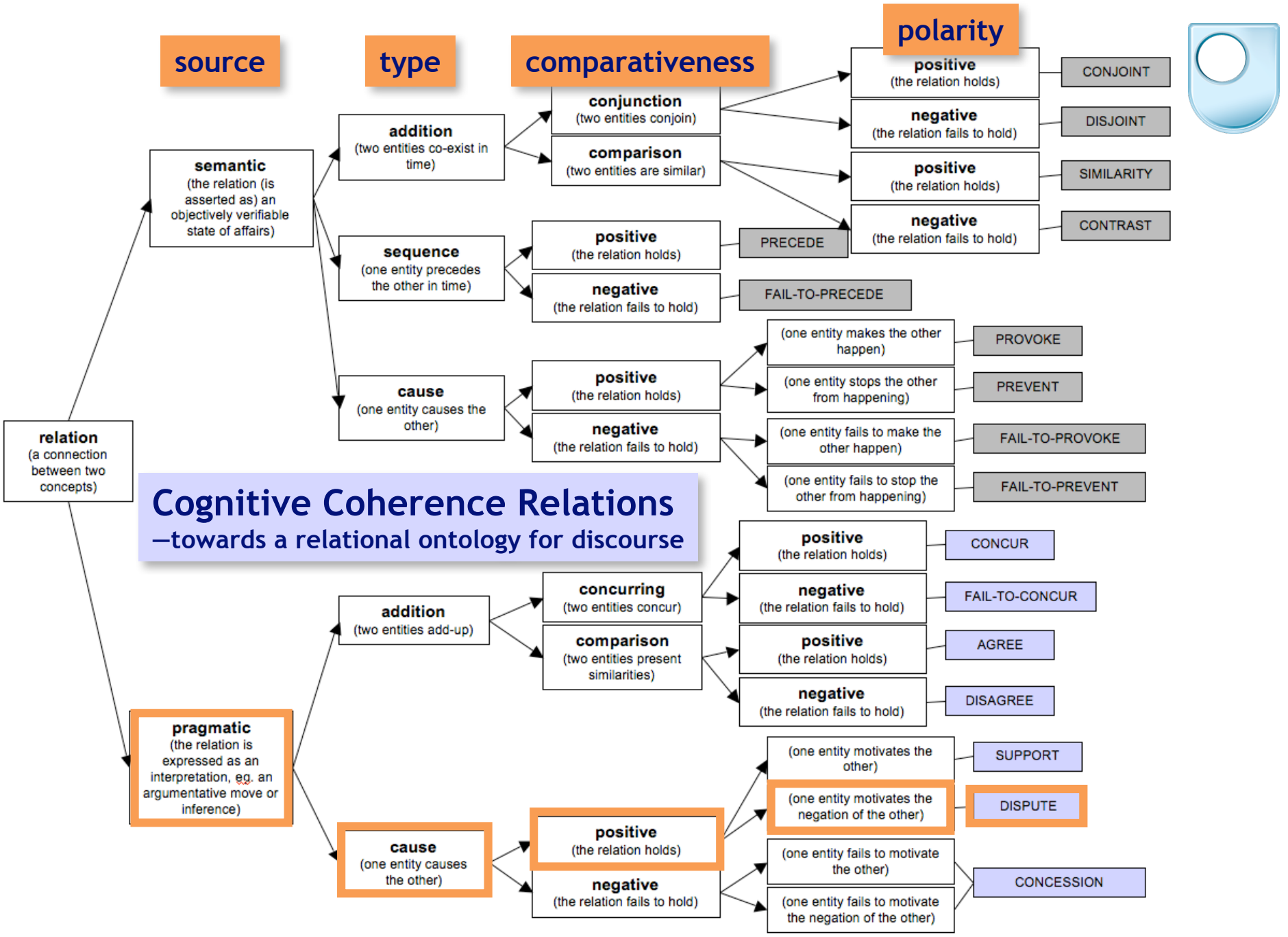
Scholarly discourse as CKS…
Beyond document citations…



Scholarly discourse as CKS…

Connecting freeform tags with naturalistic connections ("dialects")
grounded in a formal set of relations (from semiotics and coherence relations)



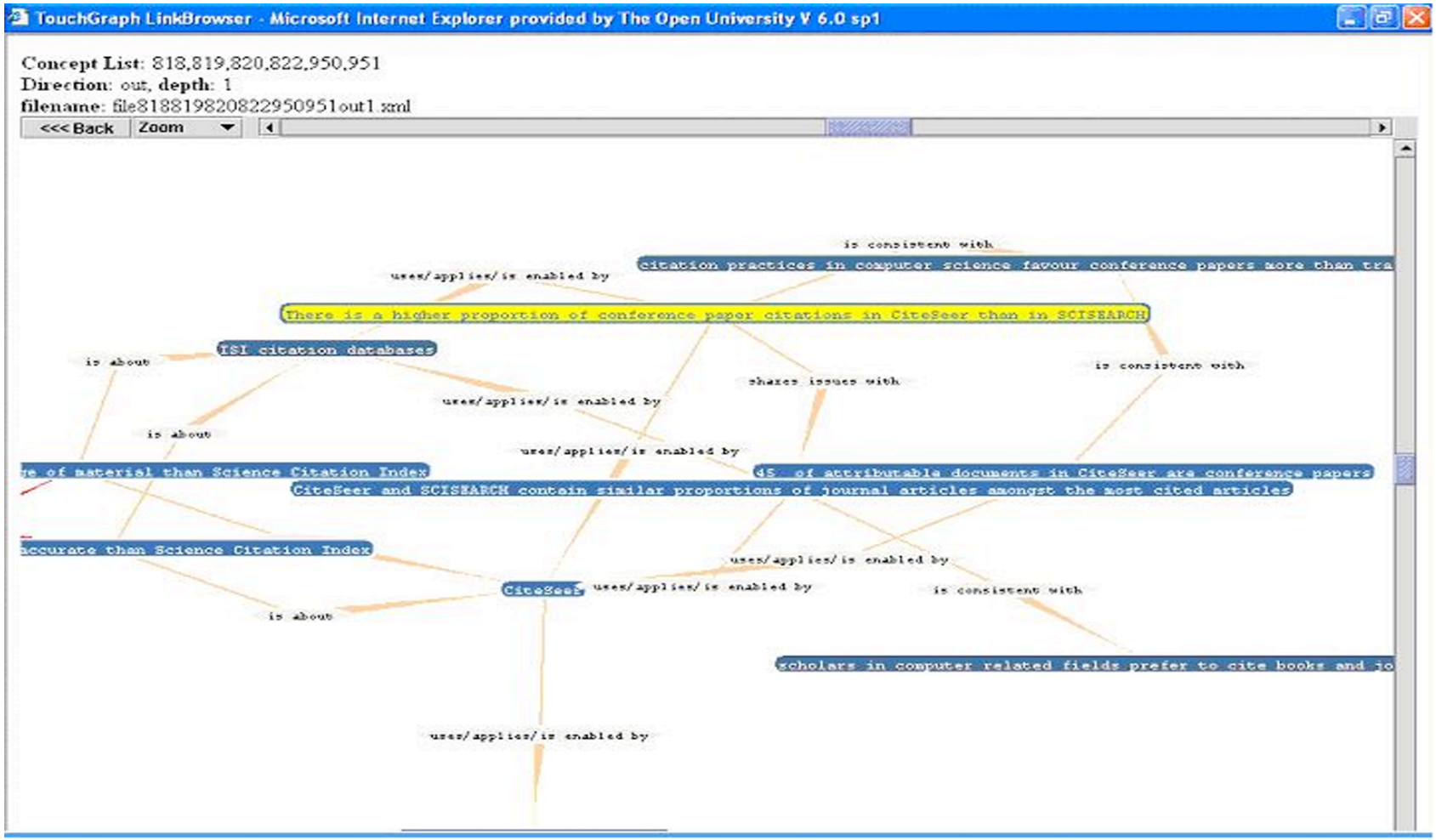


Mancini, C. and Buckingham Shum, S.J. (2006). **Modelling Discourse in Contested Domains: A Semiotic and Cognitive Framework**. International Journal of Human Computer Studies, 64, (11), pp.1154-1171

Scholarly discourse as CKS…
Visualising claims and arguments

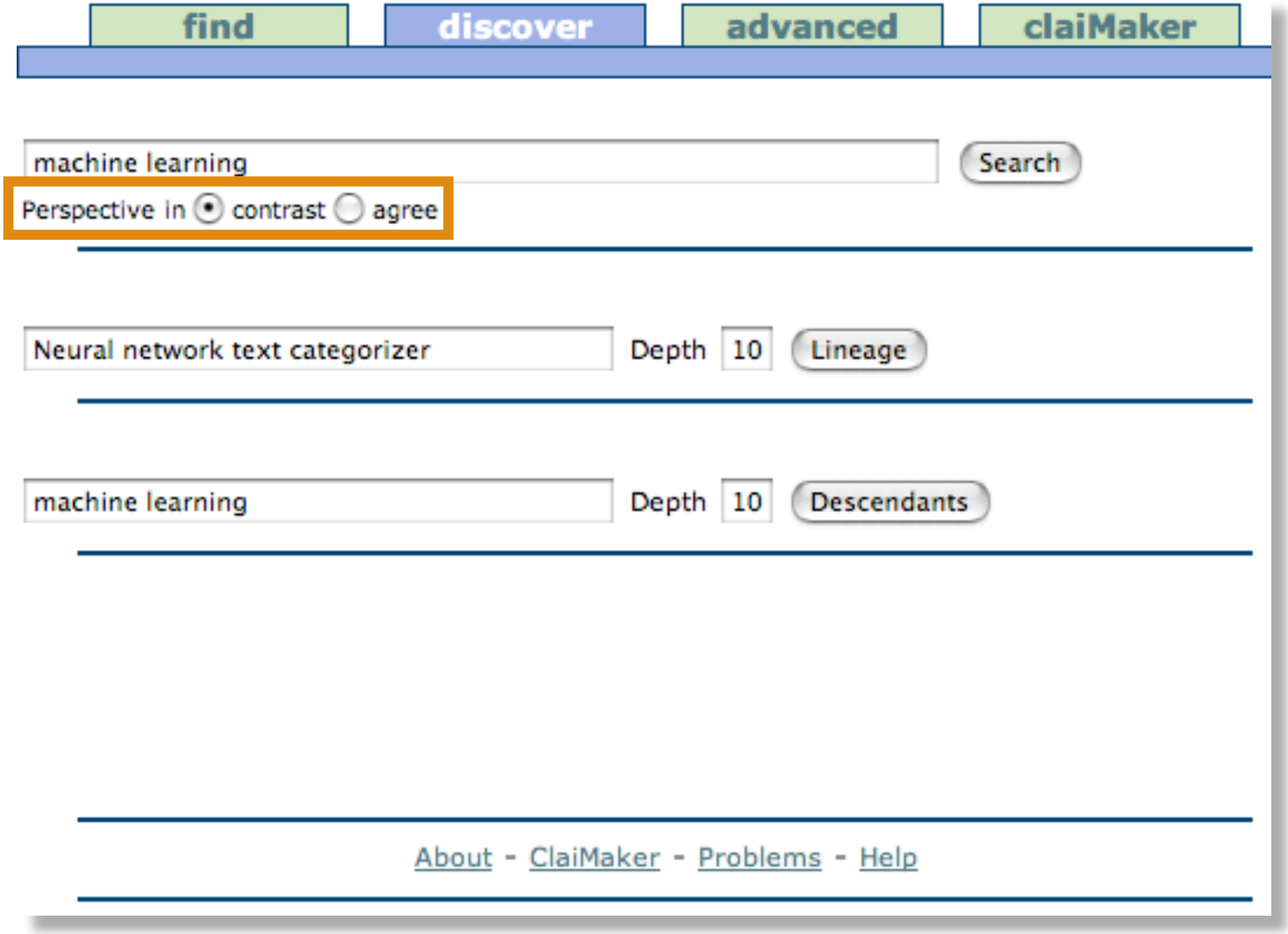
The link-tracking service allows the user to specify structures

e.g. *show tags one link out from any tag on the left hand of a claim containing the string "CiteSeer"*



Evaluated in: V. Uren, S. Buckingham Shum, G. Li, and M. Bachler. **Sensemaking Tools for Understanding Research Literatures: Design, Implementation and User Evaluation**. *International Journal of Human Computer Studies*, 64(5):420-445, 2006

Scholarly discourse as CKS…
Querying on argumentation structures



Scholarly discourse as CKS…

"What papers contrast with this paper?"

1. Extract concepts for this document
2. Trace concepts on which they build
3. Trace concepts challenging this set
4. Show root documents

The key issues you are concerned with:	
445	Decision Forest classifier
446	Decision Forest classifier improves on C4.5 and kNN

The related issues you may be concerned with:	
446	Decision Forest classifier improves on C4.5 and kNN
515	Instance based learning
511	Decision tree learning
277	decision trees and naive Bayes perform well for text categorization

The following claims disagree ...				
1	Optimised rules outperform Naive Bayes and decision trees	disagrees with	decision trees and naive Bayes perform well for text categorization	3621 2

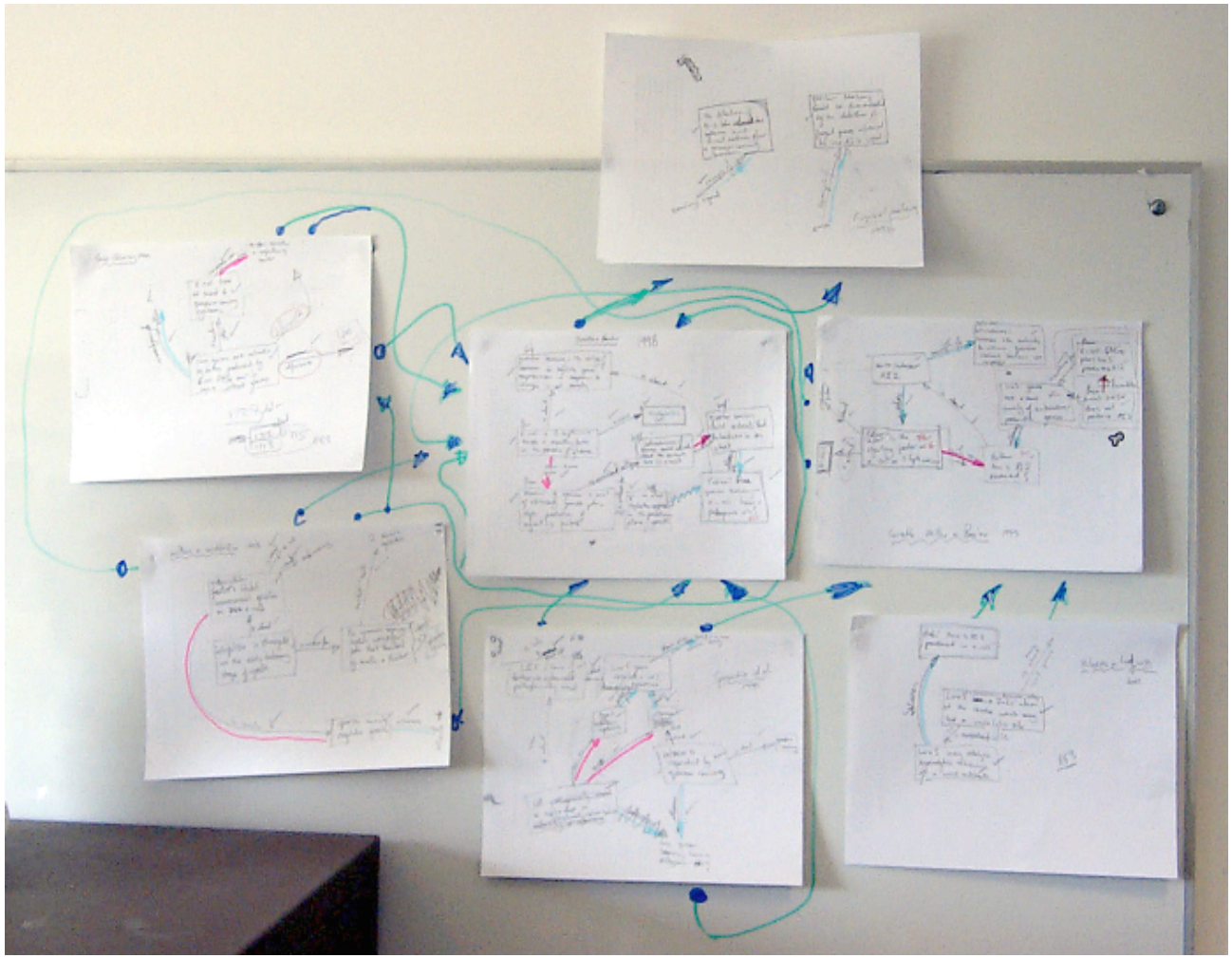
Evaluated in: V. Uren, S. Buckingham Shum, G. Li, and M. Bachler. **Sensemaking Tools for Understanding Research Literatures: Design, Implementation and User Evaluation**. *International Journal of Human Computer Studies*, 64(5):420-445, 2006

The point is… we think these kinds of structures are worth having

But can users create them?

How to help scholars engage in CKS?

Pilot study: paper-based literature modelling

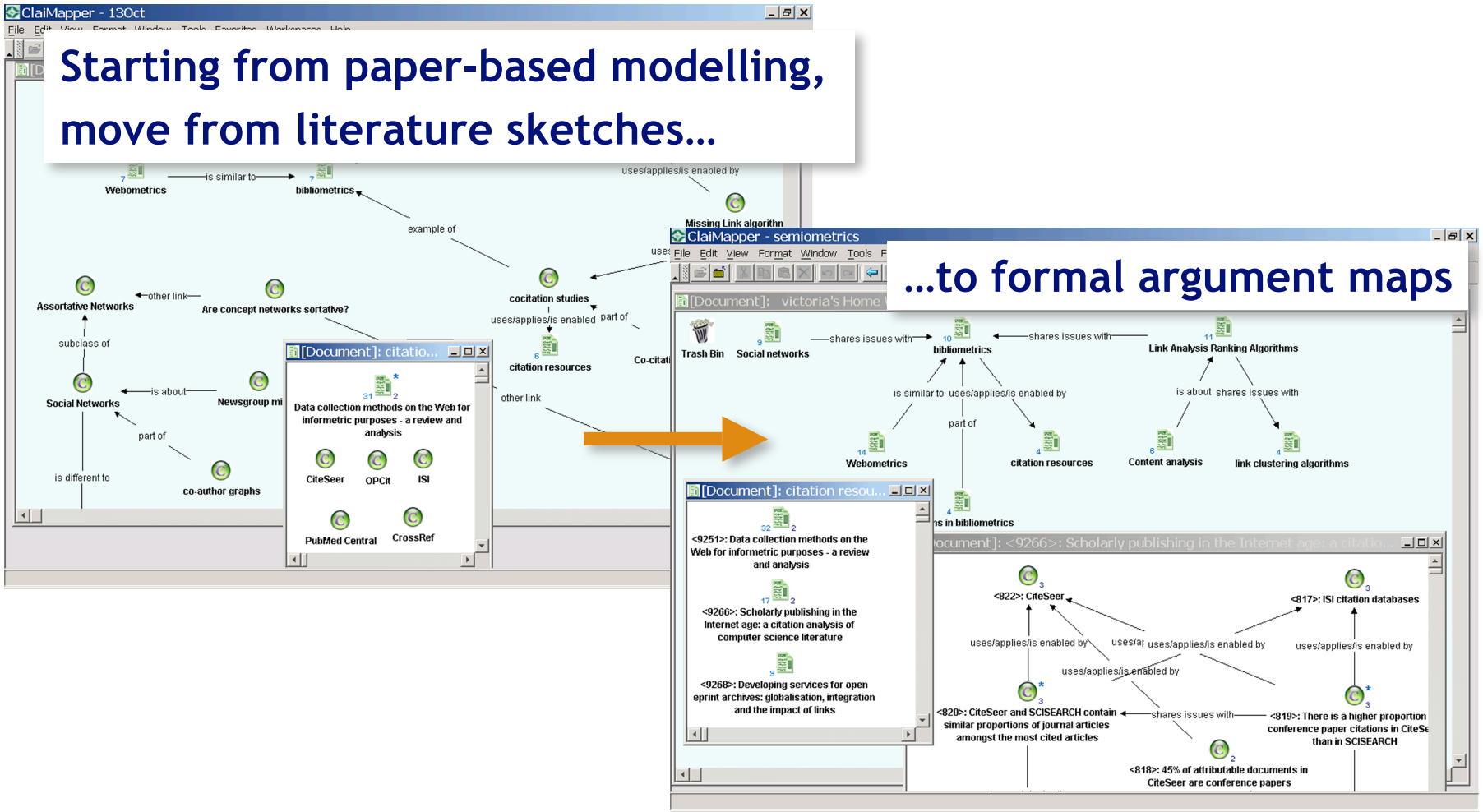


S. Buckingham Shum, V. Uren, G. Li, B. Sereno, and C. Mancini. **Computational Modelling of Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues.** *International Journal of Intelligent Systems*, 22(1):17–47, 2006

How to help scholars engage in CKS?

From paper prototype to semiformal mapping tool

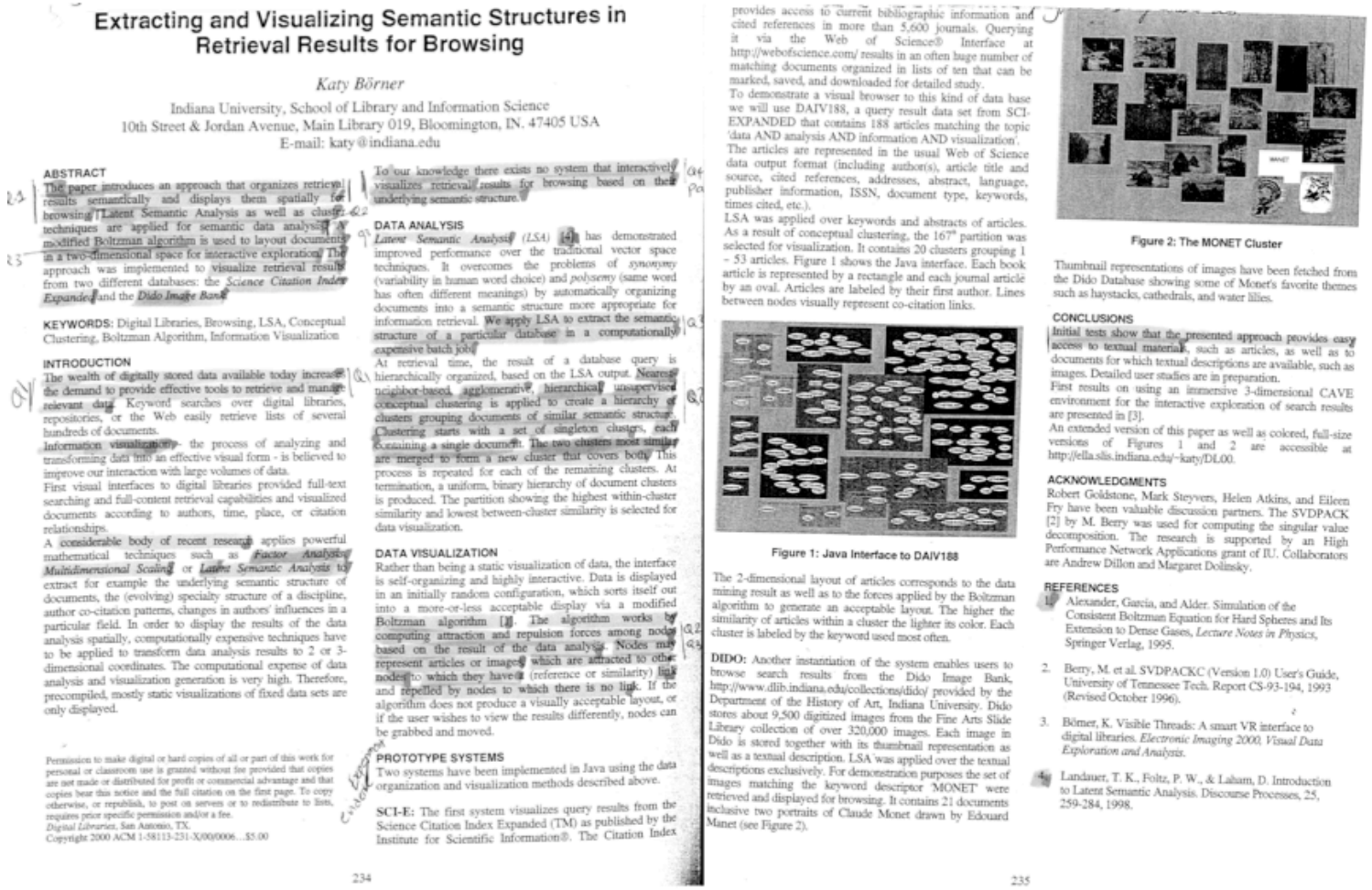
- The ClaiMapper tool



Evaluated in: V. Uren, S. Buckingham Shum, G. Li, and M. Bachler. **Sensemaking Tools for Understanding Research Literatures: Design, Implementation and User Evaluation**. *International Journal of Human Computer Studies*, 64(5):420-445, 2006

How to help scholars engage in CKS?

Pilot study: paper-based annotation



Extracting and Visualizing Semantic Structures in Retrieval Results for Browsing

Katy Börner

Indiana University, School of Library and Information Science
10th Street & Jordan Avenue, Main Library 019, Bloomington, IN. 47405 USA
E-mail: katy@indiana.edu

ABSTRACT

The paper introduces an approach that organizes retrieval results semantically and displays them spatially for browsing. Latent Semantic Analysis as well as clustering techniques are applied for semantic data analysis. A modified Boltzman algorithm is used to layout documents in a two-dimensional space for interactive exploration. The approach was implemented to visualize retrieval results from two different databases: the *Science Citation Index Expanded* and the *Dido Image Bank*.

KEYWORDS: Digital Libraries, Browsing, LSA, Conceptual Clustering, Boltzman Algorithm, Information Visualization

INTRODUCTION

The wealth of digitally stored data available today increases the demand to provide effective tools to retrieve and manage relevant data. Keyword searches over digital libraries, repositories, or the Web easily retrieve lists of several hundreds of documents.

Information visualization - the process of analyzing and transforming data into an effective visual form - is believed to improve our interaction with large volumes of data.

First visual interfaces to digital libraries provided full-text searching and full-content retrieval capabilities and visualized documents according to authors, time, place, or citation relationships.

A considerable body of recent research applies powerful mathematical techniques such as *Factor Analysis*, *Multidimensional Scaling*, or *Latent Semantic Analysis* to extract for example the underlying semantic structure of documents, the (evolving) specialty structure of a discipline, author co-citation patterns, changes in authors' influences in a particular field. In order to display the results of the data analysis spatially, computationally expensive techniques have to be applied to transform data analysis results to 2 or 3-dimensional coordinates. The computational expense of data analysis and visualization generation is very high. Therefore, precompiled, mostly static visualizations of fixed data sets are only displayed.

To our knowledge there exists no system that interactively visualizes retrieval results for browsing based on their underlying semantic structure.

DATA ANALYSIS

Latent Semantic Analysis (LSA) [4] has demonstrated improved performance over the traditional vector space techniques. It overcomes the problems of *synonymy* (variability in human word choice) and *polysemy* (same word has often different meanings) by automatically organizing documents into a semantic structure more appropriate for information retrieval. We apply LSA to extract the semantic structure of a particular database in a computationally expensive batch job.

At retrieval time, the result of a database query is hierarchically organized, based on the LSA output. Nearest-neighbor-based, agglomerative, hierarchical, unsupervised conceptual clustering is applied to create a hierarchy of clusters grouping documents of similar semantic structure. Clustering starts with a set of singleton clusters, each containing a single document. The two clusters most similar are merged to form a new cluster that covers both. This process is repeated for each of the remaining clusters. At termination, a uniform, binary hierarchy of document clusters is produced. The partition showing the highest within-cluster similarity and lowest between-cluster similarity is selected for data visualization.

DATA VISUALIZATION

Rather than being a static visualization of data, the interface is self-organizing and highly interactive. Data is displayed in an initially random configuration, which sorts itself out into a more-or-less acceptable display via a modified Boltzman algorithm [1]. The algorithm works by computing attraction and repulsion forces among nodes based on the result of the data analysis. Nodes may represent articles or images which are attracted to other nodes to which they have a (reference or similarity) link and repelled by nodes to which there is no link. If the algorithm does not produce a visually acceptable layout, or if the user wishes to view the results differently, nodes can be grabbed and moved.

PROTOTYPE SYSTEMS

Two systems have been implemented in Java using the data organization and visualization methods described above.

SCI-E: The first system visualizes query results from the Science Citation Index Expanded (TM) as published by the Institute for Scientific Information®. The Citation Index provides access to current bibliographic information and cited references in more than 5,600 journals. Querying it via the Web of Science® Interface at http://webofscience.com/ results in an often huge number of matching documents organized in lists of ten that can be marked, saved, and downloaded for detailed study.

To demonstrate a visual browser to this kind of data base we will use DAIV188, a query result data set from SCI-EXPANDED that contains 188 articles matching the topic 'data AND analysis AND information AND visualization'.

The articles are represented in the usual Web of Science data output format (including author(s), article title and source, cited references, addresses, abstract, language, publisher information, ISSN, document type, keywords, times cited, etc.).

LSA was applied over keywords and abstracts of articles. As a result of conceptual clustering, the 167th partition was selected for visualization. It contains 20 clusters grouping 1 - 53 articles. Figure 1 shows the Java interface. Each book article is represented by a rectangle and each journal article by an oval. Articles are labeled by their first author. Lines between nodes visually represent co-citation links.

Figure 1: Java Interface to DAIV188

The 2-dimensional layout of articles corresponds to the data mining result as well as to the forces applied by the Boltzman algorithm to generate an acceptable layout. The higher the similarity of articles within a cluster the lighter its color. Each cluster is labeled by the keyword used most often.

DIDO: Another instantiation of the system enables users to browse search results from the Dido Image Bank, http://www.dlib.indiana.edu/collections/dido/ provided by the Department of the History of Art, Indiana University. Dido stores about 9,500 digitized images from the Fine Arts Slide Library collection of over 320,000 images. Each image in Dido is stored together with its thumbnail representation as well as a textual description. LSA was applied over the textual descriptions exclusively. For demonstration purposes the set of images matching the keyword descriptor 'MONET' were retrieved and displayed for browsing. It contains 21 documents inclusive two portraits of Claude Monet drawn by Edouard Manet (see Figure 2).

Figure 2: The MONET Cluster

Thumbnail representations of images have been fetched from the Dido Database showing some of Monet's favorite themes such as haystacks, cathedrals, and water lilies.

CONCLUSIONS

Initial tests show that the presented approach provides easy access to textual material, such as articles, as well as to documents for which textual descriptions are available, such as images. Detailed user studies are in preparation.

First results on using an immersive 3-dimensional CAVE environment for the interactive exploration of search results are presented in [3].

An extended version of this paper as well as colored, full-size versions of Figures 1 and 2 are accessible at http://ella.slis.indiana.edu/~katy/DL00.

ACKNOWLEDGMENTS

Robert Goldstone, Mark Steyvers, Helen Atkins, and Eileen Fry have been valuable discussion partners. The SVDPACK [2] by M. Berry was used for computing the singular value decomposition. The research is supported by an High Performance Network Applications grant of IU. Collaborators are Andrew Dillon and Margaret Dolinsky.

REFERENCES

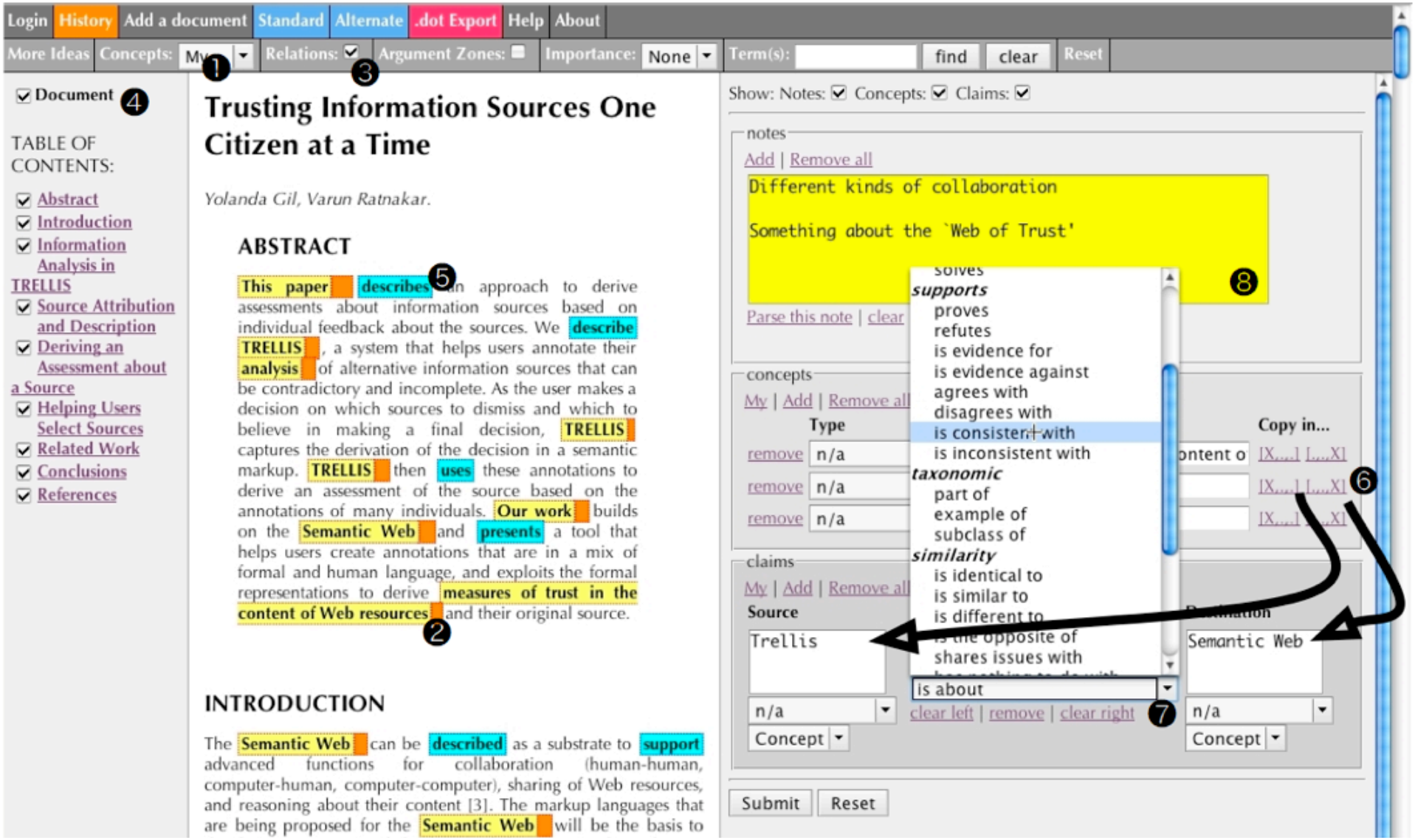
1. Alexander, Garcia, and Alder. Simulation of the Consistent Boltzman Equation for Hard Spheres and Its Extension to Dense Gases, *Lecture Notes in Physics*, Springer Verlag, 1995.
2. Berry, M. et al. SVDPACKC (Version 1.0) User's Guide, University of Tennessee Tech. Report CS-93-194, 1993 (Revised October 1996).
3. Börner, K. Visible Threads: A smart VR interface to digital libraries. *Electronic Imaging 2000, Visual Data Exploration and Analysis*.
4. Landauer, T. K., Foltz, P. W., & Laham, D. Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284, 1998.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Digital Libraries, San Antonio, TX.
Copyright 2000 ACM 1-58113-231-X/00/0006…$5.00

Pilot study reported in: B. Sereno, S. Buckingham Shum, and E. Motta. (2005). **ClaimSpotter: an Environment to Support Sensemaking with Knowledge Triples**. Proc. *Int. Conf. Intelligent User Interfaces*, pages 199–206, ACM

How to help scholars engage in CKS?

- The ClaimSpotter annotation tool: Web 2.0-style tagging with optional community/system tag recommendations



A user-centred approach to the formative evaluation of a CKS tool

- **Research question:**
  - how do annotators approach the task of using a new Web tool to semantically annotate a document with its key contributions?

- **Focus**
  - ..is on how the tool functionality and UI affordances shape tagging behaviour

- **Quantitative and qualitative analysis**

Example *claims* (tag triples) from users

- Domain ontology *is about* A hierarchy of URIs on multiple levels
- Universal physical access *is unlikely to affect* Digital divide
- Hypertext node juxtaposition *is analogous to* Cinematic shot juxtaposition
- [Evidence] In the Bristol trial, the awareness of the presence of other players was correlated with how much our participants enjoyed the game as well as with how engaged they felt *is consistent with* Presence awareness of many other people is capable of causing, feel good factor
- Magpie moves away from hypermedia towards open service-based architectures *is evidence for* [Magpie *improves on* COHSE]

User study: selected results

- See paper for details

- and the PhD for complete account

  B. Sereno. A Document-Centric Semantic Annotation Environment to Support Sense-Making. PhD Thesis, Technical Report KMI-06-13, Knowledge Media Institute, The Open University, UK, September 2005.
  [http://kmi.open.ac.uk/publications/pdf/KMI-TR-06-13.pdf]

Tag length similar for novices and experts (64% 1-3 words)

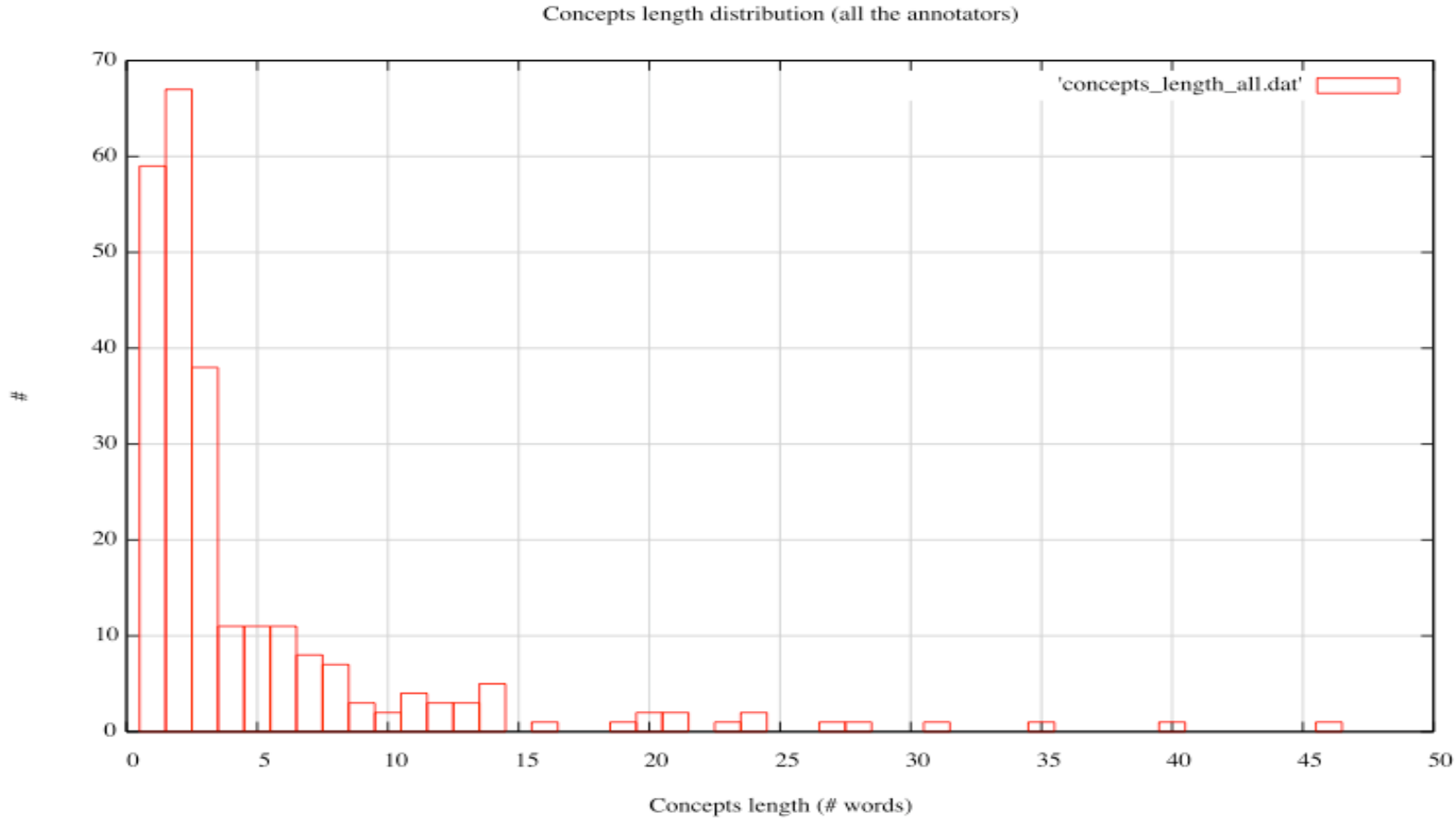


Figure 6.2: Concepts length distribution graph for all the annotators. 164 concepts out of 257 are composed of three words or less.

B. Sereno. A Document-Centric Semantic Annotation Environment to Support Sense-Making. PhD Thesis, *Technical Report KMI-06-13*), Knowledge Media Institute, The Open University, UK, September 2005. [http://kmi.open.ac.uk/publications/pdf/KMI-TR-06-13.pdf] – **p.161**

Tag reuse

most of them used only twice in this study (1 hour)

Reused concepts
a community-based project that wired four computing centres (hubs) in a lower socio-economic urban area | *a research project aiming to explore the potential of spontaneous social behaviour and playful group interaction in public spaces* | *A set of recommendations to make the process as painless as possible* | *a tool that assists users with interpreting the web resources* | *a wireless location based multiplayer game* | *access* | *Accessing information efficiently* | *ACE* | *ACM Digital Library* | *Adding formalised knowledge to a document* | *Adding information to help sense-making* | *analysis* | *AquaLog* | *awareness* | *CitiTag* | *ClaiMaker* | *ClaimSpotter* | *cognitive overhead* | *Cognitive overload in ClaimSpotter* | *cognitive strategies* | *coherence* | *COHSE* | *Collective sensemaking* | *Data-Flow* | *data-flow model* | *Digital divide* | *Discourse ontology* | *Document annotation* | *domain hierarchy* | *don't-want-tos* | *Eprint archives* | *ePrint services* | *espotter* | *explores* | *Formalization overhead* | *GATE* | *have nots* | *Holding an internal model is troublesome* | *How people approach documents* | *impact of the social context* | *Information environments organised via digital libraries* | *Information-driven reading* | *interpretation and information gathering* | *Linking* | *Magpie* | *mobile technologies* | *navigation of web resources* | *non-users* | *OpCit* | *Point-driven reading* | *Presence awareness* | *reasons why some people choose not to compute.* | *Recognising entities likes names and organisations in a document* | *robust services required for large-scale information environments* | *ScholOnto* | *Semantic services* | *Semantic Web* | *Sensemaking* | *social experiences and group play* | *START* | *stories* | *Story-driven reading* | *subscribes to* | *survey* | *The Compendium approach* | *The Fujitsu hub wiring experiment* | *The information in there does not exist in the document* | *The Internet* | *This paper* | *universal physical access* | *use of semantic information* | *User studies* | *VIPERS*

Table C.3: Concepts reused by the annotators.

B. Sereno. A Document-Centric Semantic Annotation Environment to Support Sense-Making. PhD Thesis, *Technical Report KMI-06-13*), Knowledge Media Institute, The Open University, UK, September 2005. [http://kmi.open.ac.uk/publications/pdf/KMI-TR-06-13.pdf] — **p.259**

Transcript analysis

to derive themes, sub-categories and codes

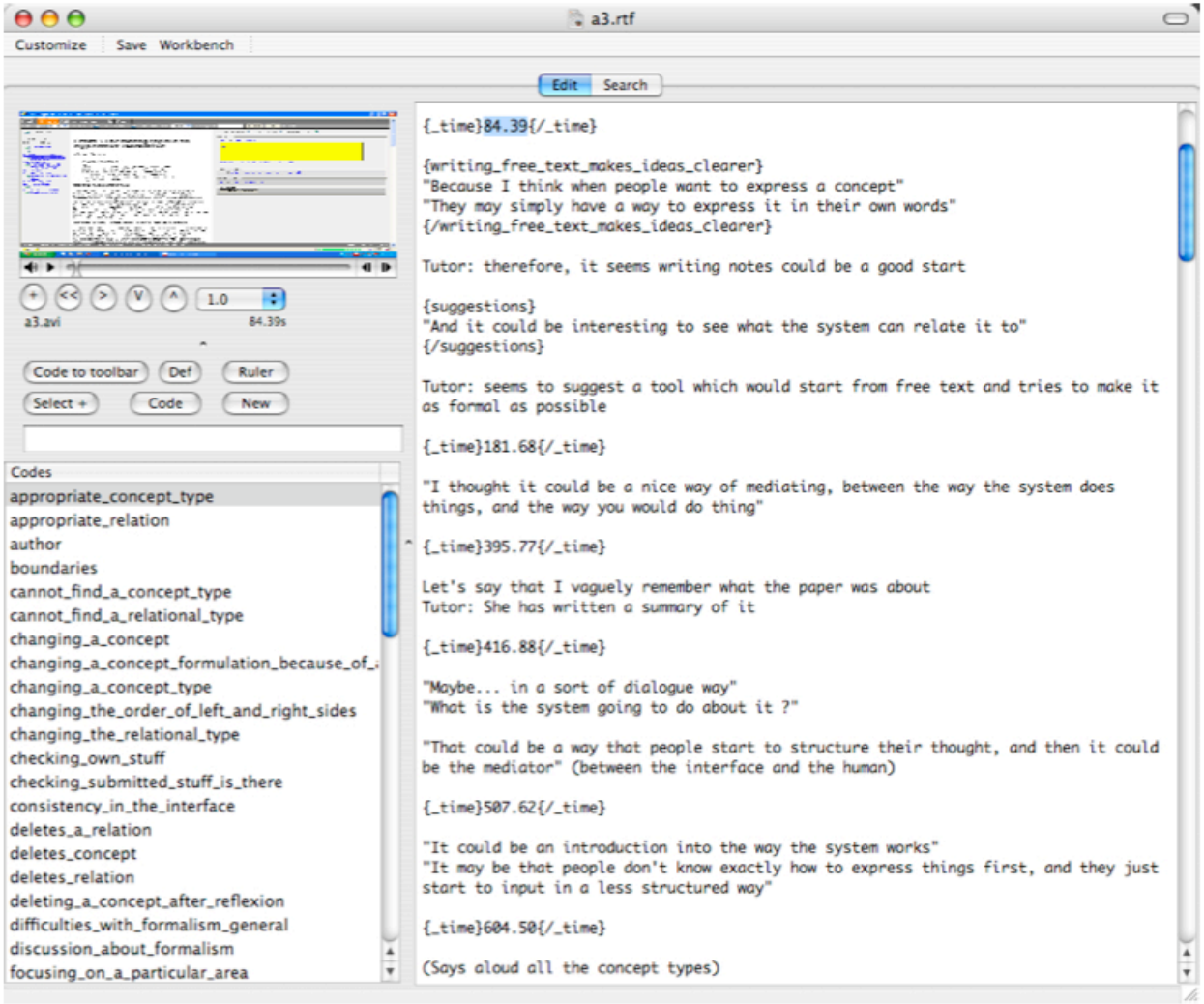


Figure 6.7: A coding session in TAMS: text chunks (main window, right side) are selected and assigned a code (selected from the bottom-left window, or created from scratch.)

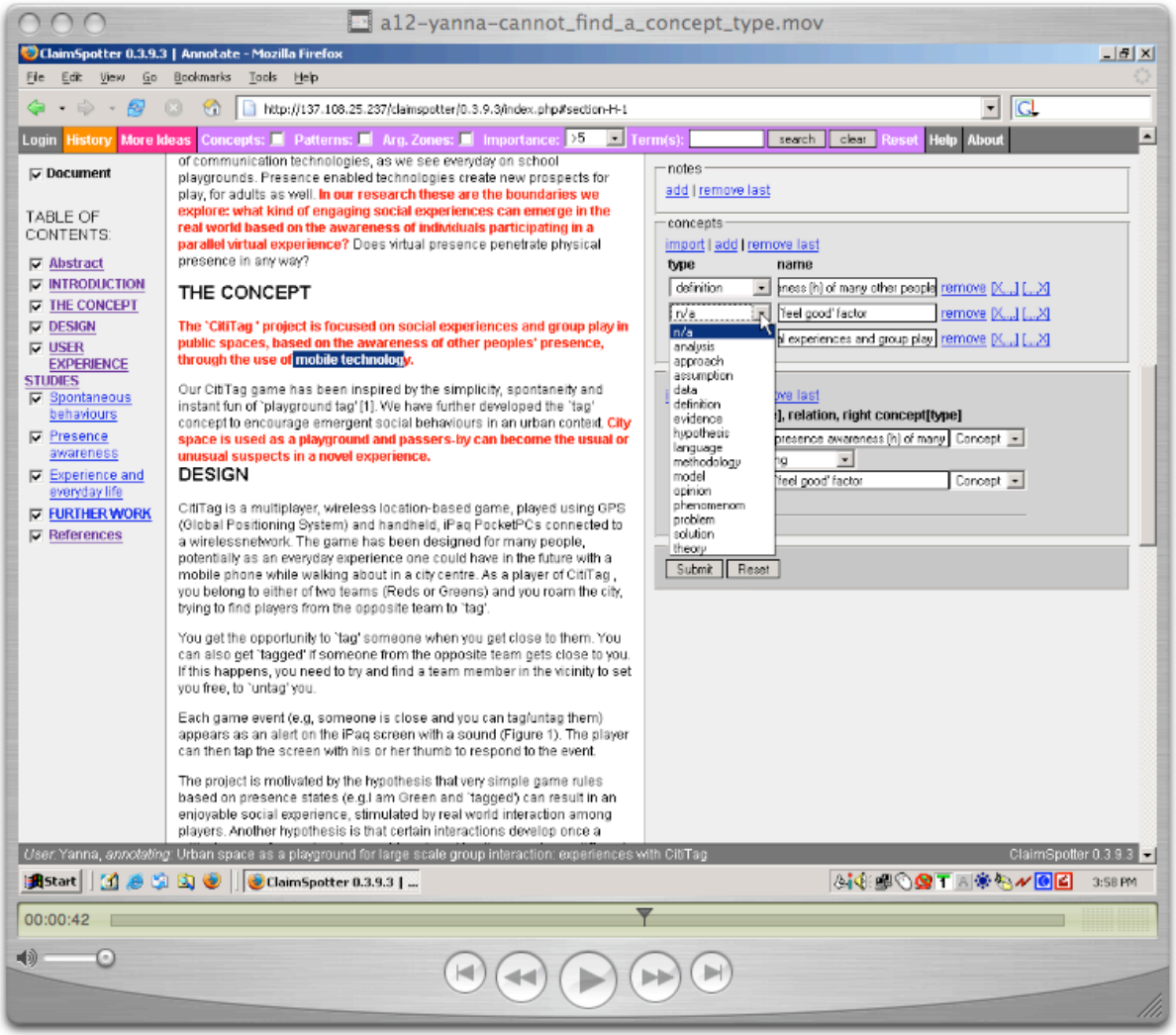
Themes from the user study

- **Formalization**
  - the cognitive task of structured tagging
- **User Strategy**
  - how users approach the semantic annotation task
- **Interaction Design**
  - how behaviour is shaped by the tool's affordances

Formalization

the cognitive task of structured tagging

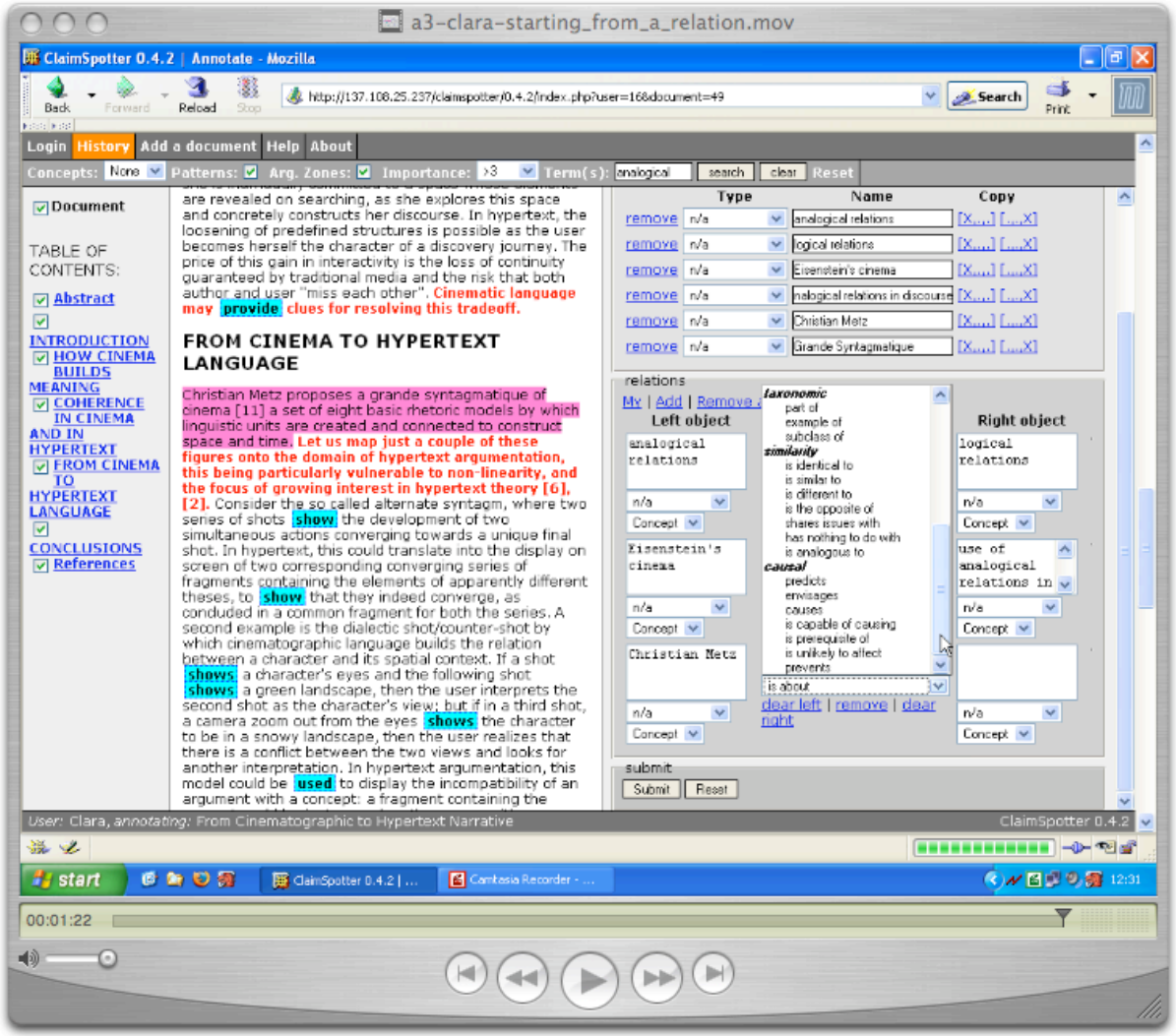
- Looking for the right tag type…



Formalization

the cognitive task of structured tagging

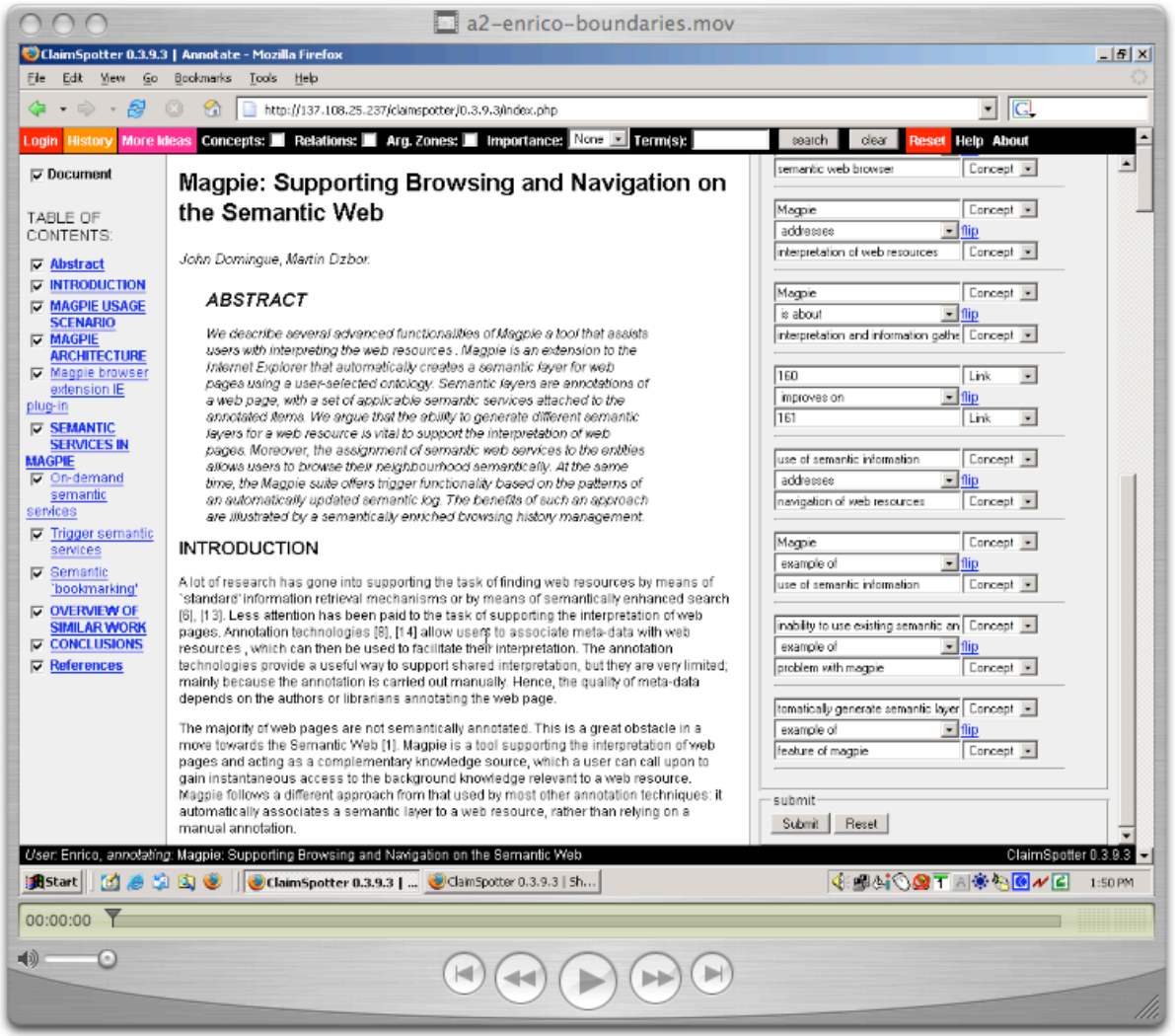
- Looking for the right link type…



User Strategy

how users approach the semantic annotation task

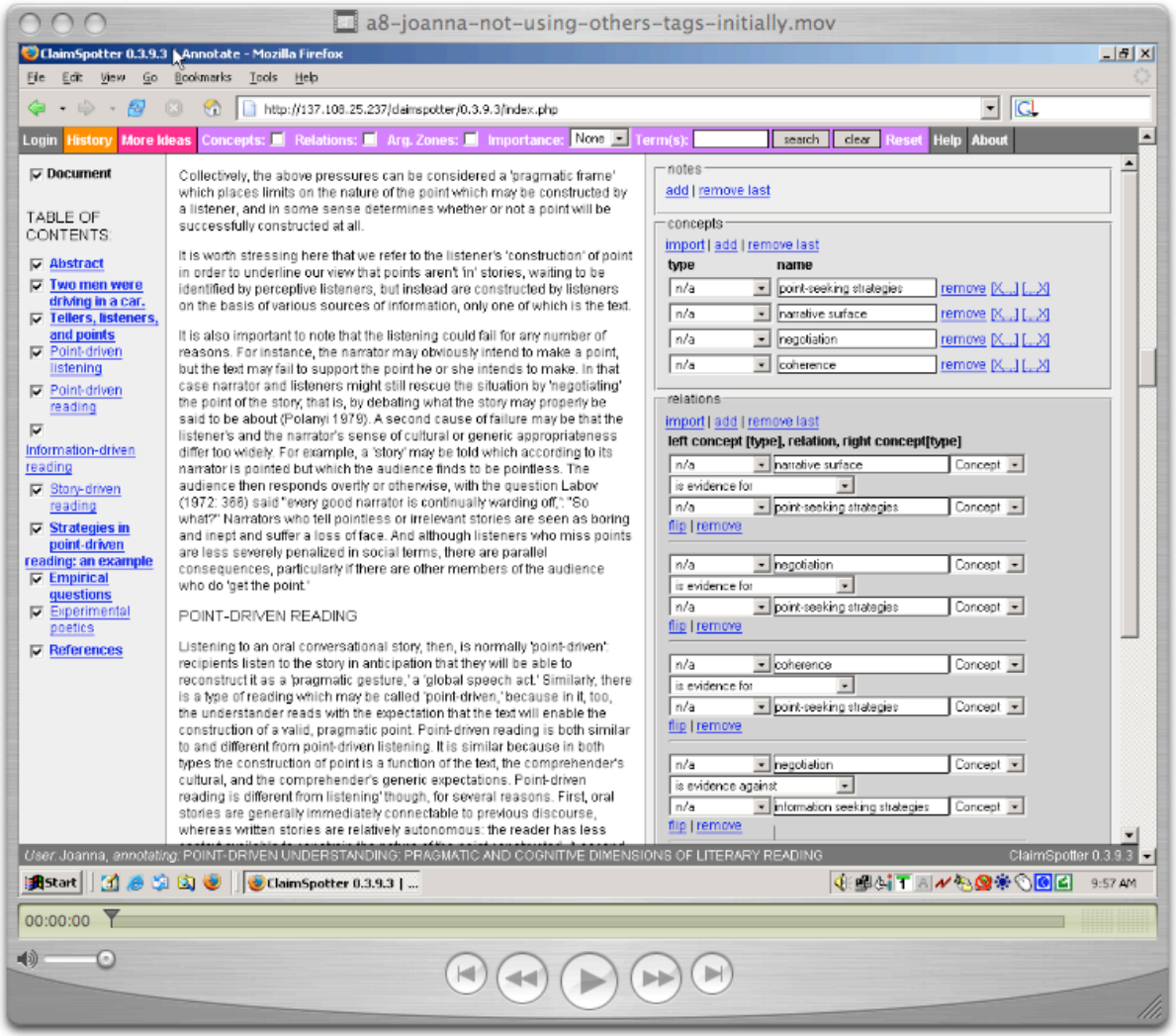
- What granularity and type of claims? When to stop?



User Strategy

how users approach the semantic annotation task

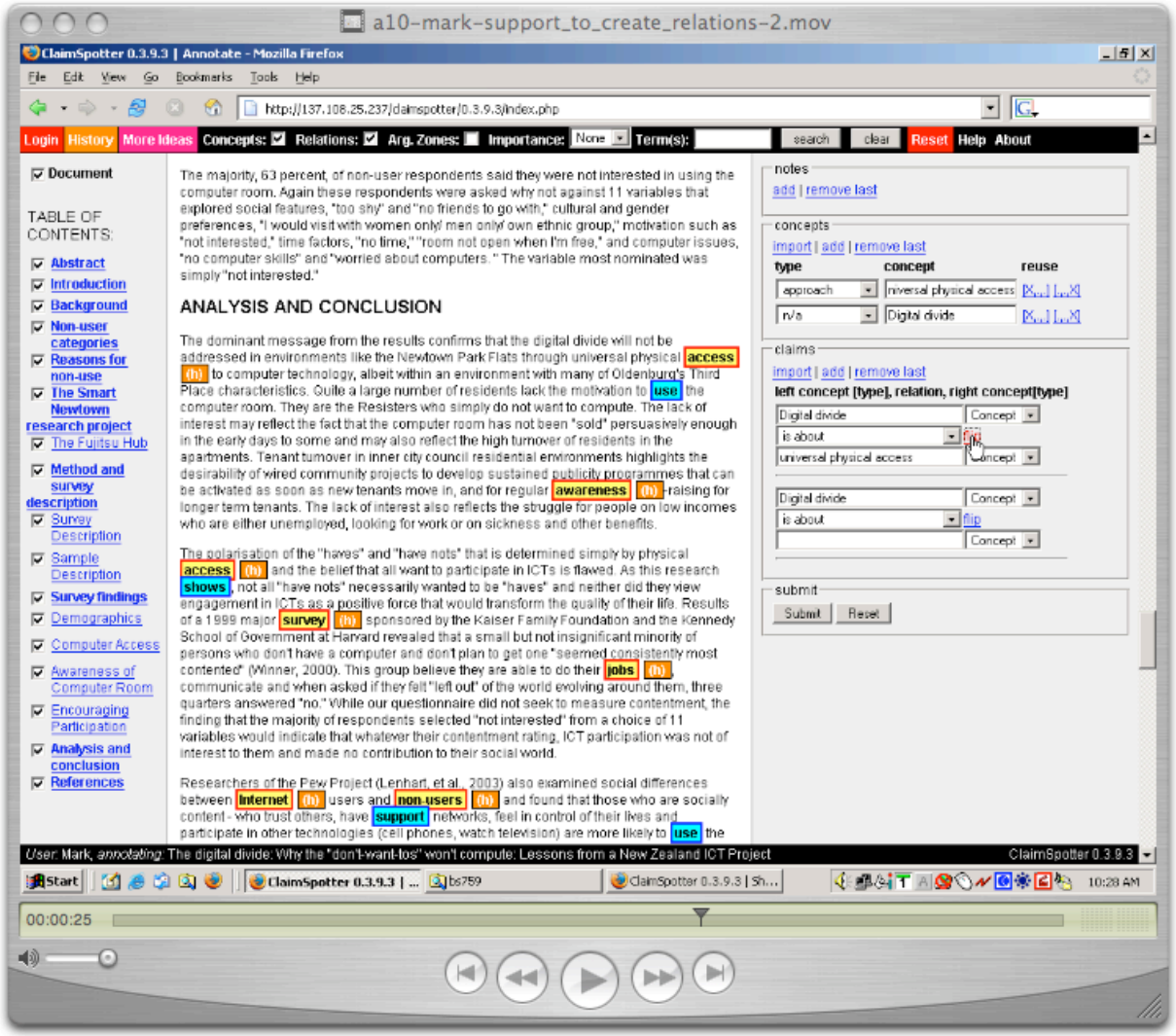
- No initial use of tagging aids — focus is on own tags



Interaction Design

how behaviour is shaped by the tool's affordances

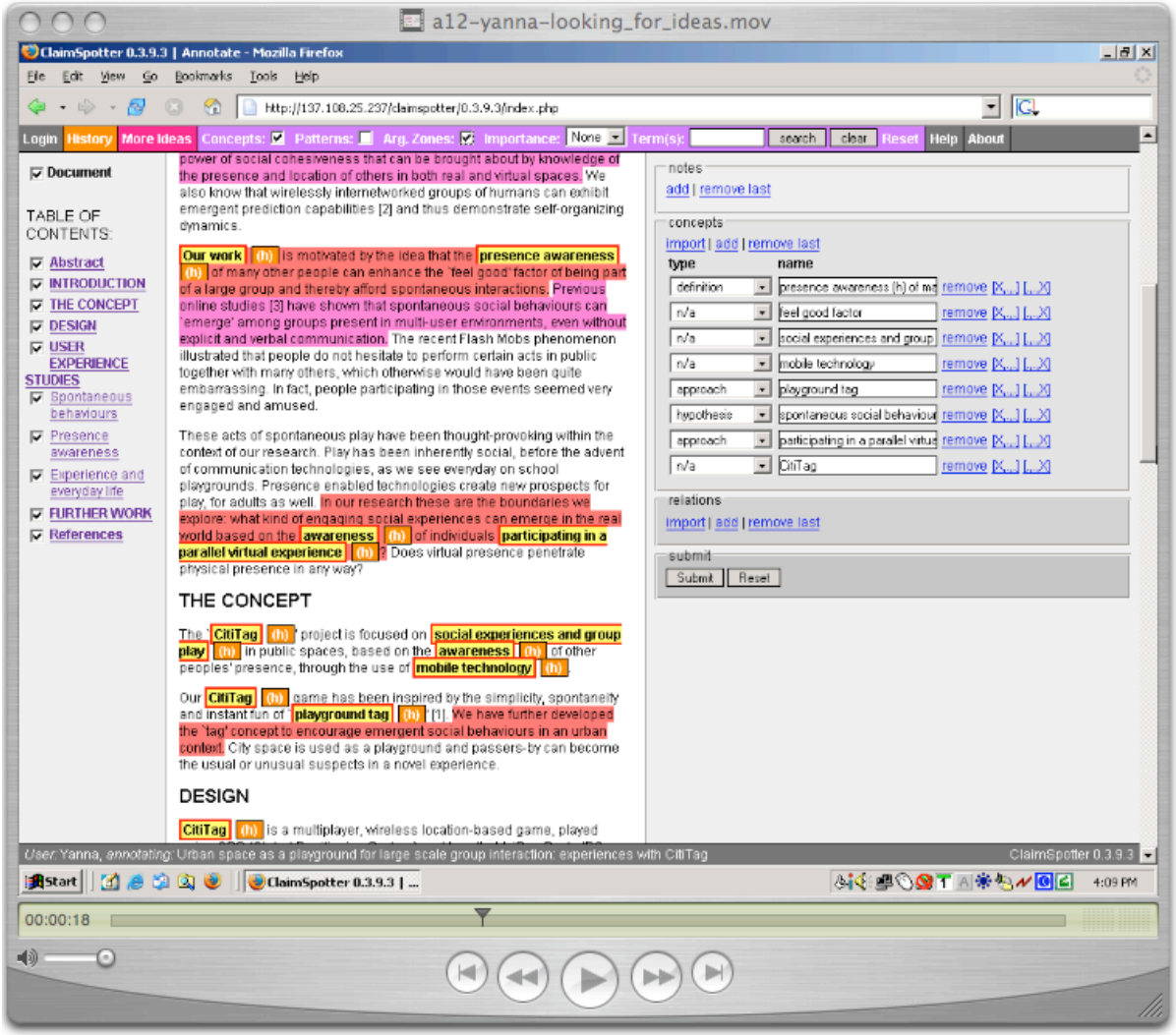
- 'Flip' left/right tags to match the link type



Interaction Design

how behaviour is shaped by the tool's affordances

- Skimming highlighted text



Lessons Learnt & Design Principles

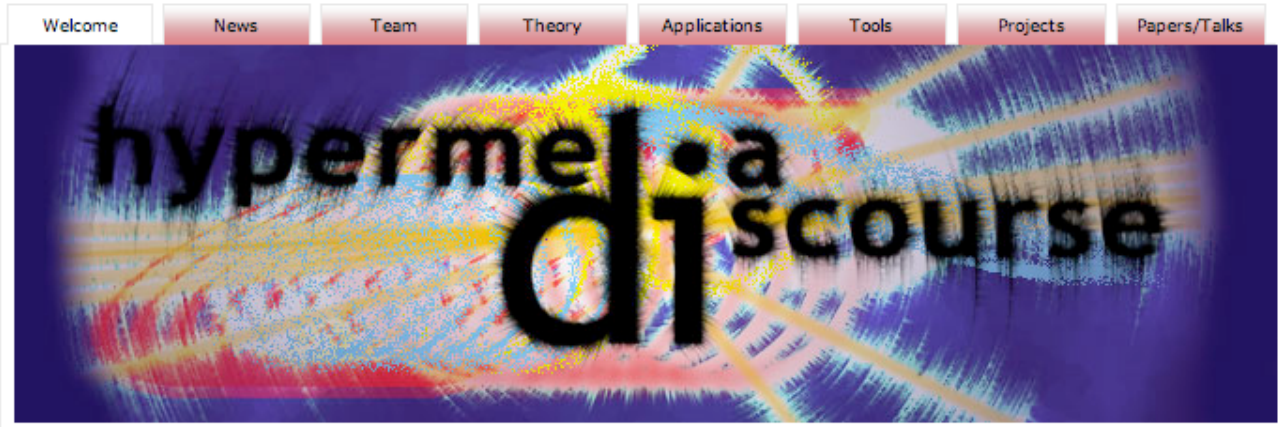
- Untrained users can do it: in their first hour they created coherent claims. UI design validated to this degree.
  —future work: longitudinal evaluation at scale

- New users attend to what is highlighted for them (matching tags; primary doct.), and generally don't click down a level
  —next version combines visualizations and document-centric features

- Support incremental formalization
  —cf. use of *is-about* as a placeholder link; provide an *Other…* category and try to map automatically to the ontology

- Users' strategies vary — don't assume a strong workflow
  a paper-based pilot study can provide insights into this

- Web 2.0 UI simplicity: good design needed to provide high functionality, walk-up-and-use CKS tools
  —we overwhelmed some users with overlaid suggestions for tags



ClaimSpotter:

papers and demos

http://kmi.open.ac.uk/projects/hyperdiscourse/tools/claimspotter

Hypermedia Discourse project:

theories / tools / case studies / user studies: face-face and asynch. interaction

http://kmi.open.ac.uk/projects/hyperdiscourse

collaboration / semantics / usability / community informatics / argumentation

http://www.PragmaticWeb.info

Short/full paper submission deadline: 14 May