Conceptual challenges for the evaluation of digital repositories with multiple access options. A case study

Lydia Bauer, Josef Herget and Sonja Hierl

University of Chur (HTW)
Switzerland
{Lydia.Bauer, Josef.Herget, Sonja.Hierl}@fh-htwchur.ch

Abstract. The evaluation of digital repositories with multiple access options poses various questions concerning the choice of an appropriate methodology and an adequate evaluation setting. Especially when semantic concepts are represented by visual components a multi-methodology approach needs to be taken. This paper discusses such an approach for coping with the various challenges. This work is based on a digital image repository that allows multiple access via semantic concepts such as topic maps, classical full-text search and content-based image retrieval. This system is the central part of a research project known as Living Memory.

Keywords: Evaluation Design, Methodology, Digital Image Repository, Image Retrieval, Semantic Approach, Visual Information Retrieval System

1   Introduction

Although the evaluation of information retrieval systems and digital repositories is being widely discussed, several challenges still can be identified when visual representation of the interface is based on semantic concepts like topic maps.

The main challenge is the fact that the causal connection between semantic and visual components and the search results of the user interaction with the information system is quite complex and needs to be examined through various approaches and methods. During work on the evaluation of an information system called *Living Memory,* a case study was implemented with a multi-methodology and multiple-setting approach. This approach is presented in the present paper.

2   A framework for an evaluation setting

2.1   Current deficiencies

While the evaluation of common information retrieval systems (IRS) and search engines is already in progress and is based on standard procedures and measurements, no such standards have yet been established for IRS based on semantic and visual concepts. Vaughan (2004) states that the design of valid evaluation settings and techniques is not keeping up with the rapid development of visual IRS.

Therefore most authors either use standard information retrieval evaluation measurements and settings or apply usability measurements for evaluating their visual IRS (Koshman, 2005; Reiterer, Tullius & Mann, 2005; Zwol & Oostendorp, 2004; Reiterer, 2004; Mann, 2002; Sebrechts, Vasilakis, Miller et al., 1999 or Veeresamy & Belkin, 1996).

Only in a few cases do the authors motivate their choice and apply a combination of methods from those two areas. The different evaluation settings produce incompatible results and do not allow overall findings to be deduced (Cugini, Laskowski & Sebrechts, 2000). As an illustration, Chen and Yu (2000) performed a meta-analysis of 35 evaluation case studies, but in the end they only could compare 6 of them because the preconditions and starting points of the remaining studies were too diverse and were based on the varying methodologies.

However, the authors state that such evaluation is decisive, especially in the area of IRS with visualization and semantic approaches:

"One of the lessons of our experience is that no matter how much intuitive appeal a given interface might have, without some systematic testing, its real value remains unknown. Especially in the field of visualization, it is all too common for technical wizardry to be unaccompanied by any gain in efficiency" (Chen & Yu, 2000).

2.2   Interdependences between retrieval, usability and the visual representation of semantic concepts

A further fact that has to be considered when selecting suitable evaluation methods is the interdependences that can be identified between the retrieval algorithms in an IRS and the semantic concepts that are represented by a visual component as shown in figure 1.
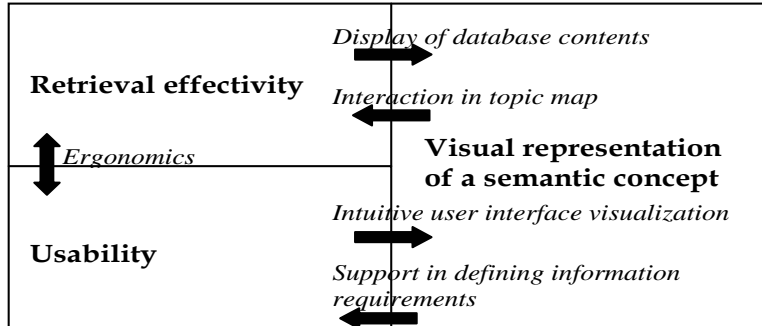
Figure 1: Interdependences between retrieval algorithms, usability and the visual representation of a semantic concept in IRS

2.1.1 Interdependences between retrieval efficiency and semantic and visual components

When conducting an evaluation it has to be considered that the visualization is affected by the retrieval algorithms used. When, for example, an IRS uses a non-hierarchical visualization (which is the case in a graphical display of a topic map), numerical ranking that has been calculated by the implemented retrieval algorithms, cannot be displayed.

On the contrary, the interaction steps performed on the visualized results by a user need to be executable through the implemented retrieval algorithms in the IRS.

2.1.2 Interdependences between usability and visual and semantic components

Furthermore, intuitively usable visualizations need to be applied in order to assist the user during the retrieval process. Especially when relations are displayed, for example, in a topic map, the user must be able to understand what is represented by lines (associations) and dots (topics) etc. It is not easy to evaluate whether a visualization or semantic display does actually help users in defining their information requirements in a useful way as cause and effect cannot be clearly determined.

2.1.3 Interdependences between retrieval efficiency and usability

The effect of ergonomics is another difficulty: New semantic concepts might be very convincing and do theoretically improve retrieval efficiency. However, if these approaches are not applied in an ergonomic way, they do not help users to satisfy their information needs. Here again, as mentioned above, it is very challenging to measure cause and effects of the components in an evaluation process.

2.1.4   Integrated mix of methods

As a conclusion, concerning the interdependences it can be said that an integrated mix of different methods both from the area of usability and retrieval evaluation needs to be applied (Plaisant, 2004; Shneiderman & Plaisant, 2006). Whereas common IRS can only be measured by using classical retrieval efficiency measurements like recall and precision, visual and semantic IRS need to be evaluated with an integrated mix of methods that allows the measurement of the same facets from another point of view, e.g. users need to fill in a questionnaire concerning their subjective impression of the quality of results in a comparative evaluation. In interpreting the evaluation, these subjective statements can be compared with the factual results of a retrieval efficiency measurement, which makes it easier to decide whether semantic and visual components have an impact on the subjectively felt and objectively measured retrieval efficiency (Hierl, 2006).

The laboratory character of retrieval efficiency evaluation methods needs to be balanced by applying methods in a controlled field, where user opinions and conditions of everyday applications are considered.

3   The Living Memory project

Living Memory is a cooperative applied research project running for two years and is still in progress.

The aims are to set up an information system of visual resources and to explore new paths of image cataloguing and retrieval, including the investigation of how topic maps can be usefully applied for the image domain. A topic map representing index terms will be used both as a navigation tool for users, allowing them to browse the image collection, and as a means of enabling semantic searches, allowing the user to choose between precise and fuzzy results. Special emphasis is given to the combination of different access options.

The visual resources document a major project in urban planning – the structural alteration of an industrial area into a research site. In order to create a digital "living memory" of the site, some hundred visual resources in different media – photographs, drawings, graphics and videos – are added to the image database per year. The images are grouped according to their creation date and topic.

3.1   The semantic structure of Living Memory

The semantic structure of the Living Memory information system is formed by three interlocking modules: metadata schema, thesaurus and topic map.

The basis for image description is a metadata schema especially designed for Living Memory. For that purpose, existing schemas were consulted such as the Dublin Core Metadata Element Set (Dublin Core, 2006), the Categories

for the Description of Works of Art (Guidelines Works of Art, 2006) and the SEPIA Data Element Set (Sepia, 2006). Since we expect users of a Living Memory information system – mostly image professionals – to search images according to a variety of criteria, the schema combines formal metadata, index terms and visual properties. Formal metadata (such as author or medium) and index terms have to be assigned intellectually, while visual features (such as contrast or luminance) can be extracted automatically.

A thesaurus drawing mainly on the Art and Architecture Thesaurus (AAT, 2006) was designed and will serve as a controlled vocabulary for image indexing.

The thesaurus also served as a basis for the construction of the topic map (Rath, 2003).

The topics of the topic map cover the index terms of the thesaurus. The topic map will serve both as a navigation tool for the user (see fig. 2) and also as an instrument for semantic searches (see sect. 3.2).
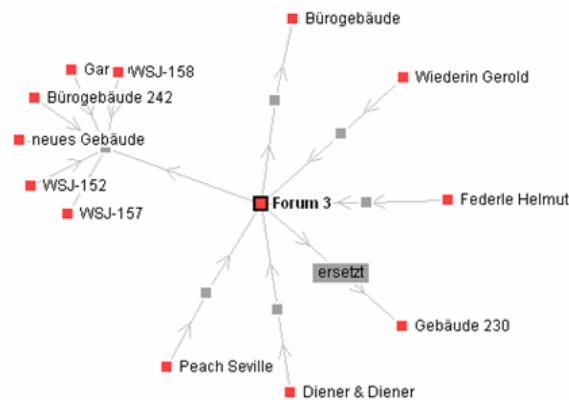

Figure 2: Extract from the *Living Memory* topic map

Since the occurrences are stored in the database, every topic will be defined as a database query. This query may be simple. The topic ”tree” , for example, will initiate a query for ”tree” in the subject term data field of the database. Consequently, every relevant image of a tree will be an occurrence, provided the images have been properly indexed. But the query may also be combined and, in fact, topics for image expression are defined in this way. The question "What makes an image evoke an idyllic scene?" may lead to answers like the following: colour feature A plus colour feature B plus subject term C. This mechanism connects the topic map to the database without redundancy.

3.2   Access options

Users will be able to combine different access options, which are assumed to be the most successful approaches in image retrieval (Rath, 2003). We, therefore, focused on the combination of different levels of image description and methods of image retrieval, rather than pushing one method to extremes. The user is presented with an interface with a menu listing the different access options: a full-text search option, the thesaurus of index terms, a list of specific objects, the lists of authors and media, the groupings by title, a time axis, a map of the area and several colour features. The user may thus make a choice from this menu and compile the search by arbitrary combinations.

The following access scenarios are expected to correspond to user needs:

- *Traditional search* can be affected as a text-based search (for formal metadata or index terms) or as a visual search (for colour features, with the possibility of adjusting a value on a scale). Given a set of images as a result, users may then want to refine their search results, or they may continue the search based on an image they have found by selecting and combining the allocated metadata of the image, including visual properties (see figure 3).
- *Semantic search* and browsing is affected with the aid of the topic map. Users can select the degree of precision (e. g. no related terms, exact match) or fuzziness (e. g. all related terms) of their search. However, they may also browse the topic map guided by subject interests. The latter has several advantages for users. They do not have to be familiar with the logic of the database or description language. Moreover, by navigating the topic map users will learn about the semantic context in which a collection and its single items are embedded, and may find unexpectedly useful items.
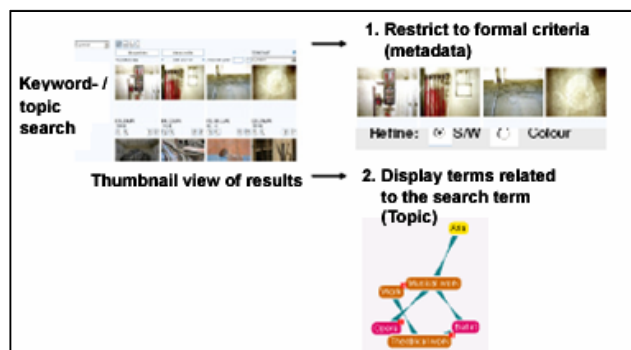

Figure 3: Schema for refining a set of search results

4   The showcase evaluation environment TERS

The *Testbed for Evaluation of Information Retrieval Systems* (TERS) shown in figure 4 is a web-based evaluation environment, developed by members of the Centre for Knowledge and Content Engineering at the University of Utrecht (Netherlands) (Zwol & Oostendorp, 2004). TERS automates most of the steps in comparative evaluation studies of IRS and to some extent combines usability methods and retrieval efficiency measurements. The management of questionnaires and user surveys is possible, as is the automated calculation of the retrieval evaluation measures of recall and precision. The testbed can be freely chosen as well as the questionnaires can be used.

Like the well-known IRS evaluation initiatives Text Retrieval Conference (TREC, 2006) and the Cross Language Evaluation Forum (CLEF, 2006), TERS applies the idea of blind-review pooling by using given topics for an evaluation based on a test collection that is known to the conductors of the study.

TERS allows subjective user opinions to be collected about the retrieval quality and usability of a system. On the other hand, the objective efficiency measures of recall and precision are calculated and can be related to the other measures.

TERS is therefore a promising testbed for the evaluation of complex IRS with visual and semantic components.

Nevertheless, several aspects such as the survey of test persons during a long-term study or the arrangement of screen capture during usability tests need to be covered in a valid evaluation setting of visual and semantic IRS. Thus it is still necessary to further develop the system based on the needs of complex evaluations.

In order to adapt the system we have started conducting a case study where we evaluate the Living Memory information system. The following chapter shows the proposed evaluation environment, which is the starting point for the further development of TERS.
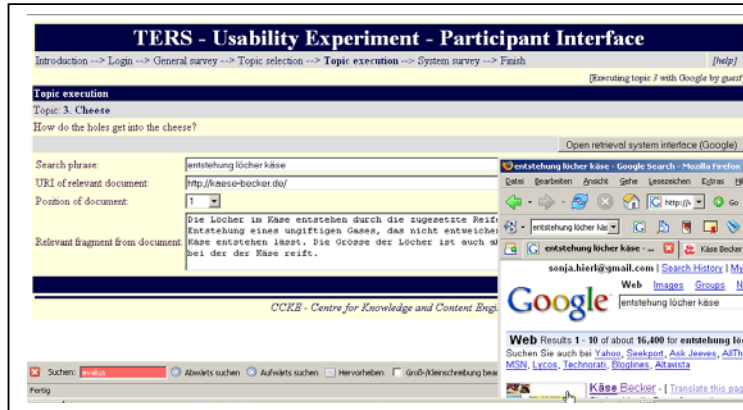
Figure 4: Graphical user interface of the TERS system

5   Proposed evaluation setting for the Living Memory project

5.1   Need for a multi-methodological and a multi-setting approach

To overcome the difficulties in evaluating IRS with visual and semantic components as described above, both evaluation dimensions will be unified in an integrated approach using a multi-methodological and a multi-setting concept combining methods from usability testing with retrieval efficiency measurements. The objective is to find out if the system is able to bring up exactly the information the user wishes to find and how satisfactory the usability of the system is.

5.2   Need for long-term studies

As Shneiderman & Plaisant (2006; Plaisant, 2004) emphasize, there is, furthermore, an urgent need for long-term studies which will take into account the fact that users need practice and especially time to get used to a system with new approaches like visualization and semantics based components.

It has been shown in several evaluations that authors qualify the measured results in the conclusions of their comparative evaluation studies by arguing that classical IRS are well known among test persons thus resulting in a bias. As a consequence, they are used to handling them and achieve better results than with new systems that apply unfamiliar approaches (Arnold, 2004).

Long-term studies have proved to be a good approach for balancing this weakness in current evaluation approaches.

Therefore, the expression "long-term study" is defined here as an evaluation process which extends over several weeks, while the test is not performed in the laboratory, but rather in the users' own, authentic working environment where they face real work processes and problems over a longer period of time. This methodology of performing research in authentic settings is

currently one attempt to implement the strategies for evaluating information visualization tools by Ben Shneiderman (Shneiderman & Plaisant, 2006).

5.3 Test setting with real user requirements, authentic test cases and information retrieval measurements

To avoid artificial test cases in the long-term phase, the test questions and tasks should correspond to the test persons' everyday working life. Ideally, these test cases should be developed with the support of the test persons themselves.

This procedure prevents the usability results from being biased, which can occur because of artificial tests and unnatural environmental settings and ensures the coverage of all user requirements while working with the digital image repository (DIR).

Nevertheless, laboratory tests are still needed to evaluate the technical qualities of the digital repository and the search. The results of these tests will help to identify the real retrieval efficiency whereas the usability tests reflect the users' subjective impression, which can differ quite considerably from the results of the efficiency measurements (Al-Maskari, Clough, Sanderson, 2006). One big advantage for the efficiency retrieval tests results from the fact that the digital repository image database has already been validated according to intellectual criteria. Ontologies represented in the form of topic maps define the existing data material, so that the content is already well known and a test collection can be built up very easily.

5.4 Methodology of the test setting

The test setting uses a mix of several subjective and objective methods to set up a meaningful evaluation environment.

5.4.1 Retrieval efficiency measuring methods

The following methods are utilized to measure the objective retrieval efficiency of the DIR

- Relative Recall@n,
- Precision@n,
- Mean Average Precision@n (MAP@n)
- and the First Retrieved Document Rank (FRDR)

The relative Recall@n inspects the relative completeness of the retrieval results of the first n hits while the Precision@n yields the precision value of the first n hits regarding a special retrieval query. Additionally, over a set of queries, MAP@n calculates the weighted average value of the precision results of the first n hits while, finally, FRDR states the ranking of the first document which users have defined as relevant so that no additional search is

needed. All these tests will help to audit the retrieval performance and reveal the actual performance of the system.

5.4.1 Usability test methods

The following methods will be combined in an integrated test setting to quantify the usability of the DIR:

- diaries
- verbal reports
- screen capture
- interviews and questionnaires

The diaries will be kept continuously by test users, recording all their actions. The verbal reports will simultaneously record every verbal reaction of the test person while operating the DIR. The automated logging will be conducted by an on-screen capture tool to collect all user-computer activities. Interviews and questionnaires will finally help to reveal the subjective perspective and position of the test person.

5.5 Evaluation process

The evaluation itself will be conducted in several phases as shown in figure 5. The usability test phase combines usability and retrieval efficiency methods where subjective results from the test user will be integrated with objective results from the retrieval efficiency measurements. The usability methods employed in this phase will be screen capture, verbal records and questionnaires, whereas the retrieval efficiency methods will be relative Recall@n, Precision@n, MAP and FRDR. This phase will be conducted in TERS with approximately 25 test users and a fixed set of test cases, so that each result will be achieved under the same conditions during a short period of time.
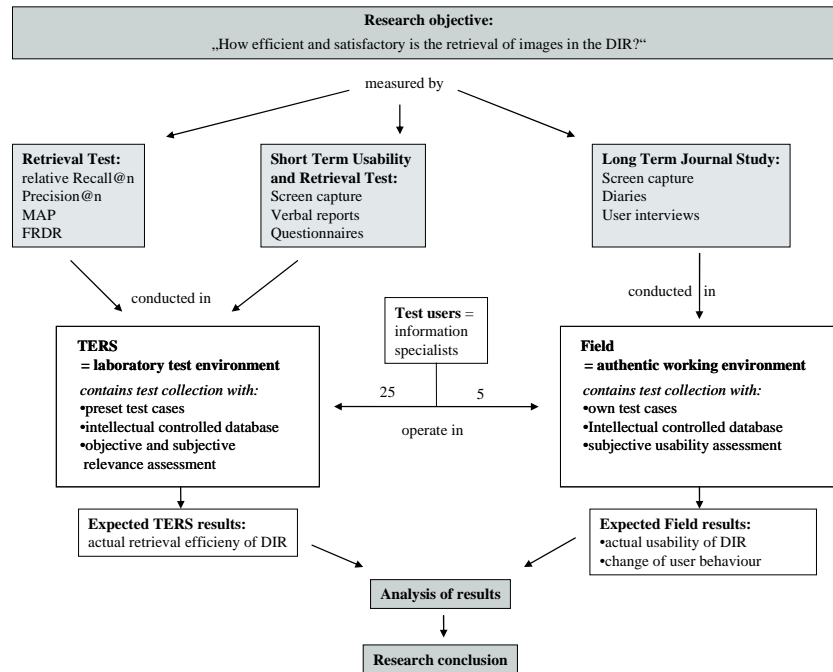
| Research objective: |
| "How efficient and satisfactory is the retrieval of images in the DIR?" |

measured by

Retrieval Test:	**Short Term Usability**	**Long Term Journal Study:**
relative Recall@n	**and Retrieval Test:**	Screen capture
Precision@n	Screen capture	Diaries
MAP	Verbal reports	User interviews
FRDR	Questionnaires	

conducted in

conducted in

TERS	**Test users =**	**Field**
= laboratory test environment	information	**= authentic working environment**
	specialists	
contains test collection with:		*contains test collection with:*
•preset test cases	25 5	•own test cases
•intellectual controlled database		•Intellectual controlled database
•objective and subjective	operate in	•subjective usability assessment
relevance assessment		

Expected TERS results:	**Expected Field results:**
actual retrieval efficieny of DIR	•actual usability of DIR
	•change of user behaviour

| **Analysis of results** |

| **Research conclusion** |

Figure 5: Proposed evaluation process

The diary phase will lead to detailed results concerning the usability of the DIR in authentic environments and will be conducted over a fairly long period of time with a smaller number of test users, approximately 5 people. The methods employed will again be screen capture, user interviews and written diaries, but the test persons will work on their own test cases and authentic, everyday situations.

The analysis of the results of the diaries, verbal reports and screen capture methods will show what parts of the system have actually been used, which search functionalities were utilized to find images, and which images have actually been found, and perhaps what problems occurred while the system was operating. Additionally, it will be possible to see whether the users changed their interaction behaviour when using the system over a longer period of time.

The overall analysis of outcomes from both the TERS and the field evaluation will provide detailed answers for our research objective of evaluating how efficient and satisfactory the retrieval of images in the digital image repository is.

6   Conclusion and further work

In this paper we propose a focused multi-methodology and multi-setting approach for the evaluation of a digital image repository.

A systematic and thorough evaluation of the *Living Memory* information system is currently being performed. The approach described above is being applied, and the strengths and limitations of the framework will be identified. First results of the evaluation will be available in January 2007.

After adapting the system according to the consolidated findings of this case study, we hope to provide a valid evaluation framework suitable for evaluation tests addressing information retrieval interfaces based on semantic concepts represented by visual components.

7   Acknowledgement

References

[1]   AAT (2006): Art & Architecture Thesaurus Online, URL: http://www.getty.edu/research/conducting_research/vocabularies/aat.

[2]   Al-Maskari, A., Clough, P., Sanderson, M. (2006). User's Effectiveness and Satisfaction for Image Retrieval. In Schaaf, M., Althoff, K.-D. (2006) *Hildesheimer Informatik-Berichte: Lernen – Wissensentdeckung - Adaptivität, 84-88.*

[3]   Arnold, C. (2004). Visualisierung im Information Retrieval. Magisterarbeit in der Philosophischen Fakultät IV (Informationswissenschaft) der Universität Regensburg: MA Dissertation in the Philosophical Faculty of the University of Regensburg, Regensburg, 2004.

[4]   Chen, C., Yu, Y. (2000). Empirical studies of information visualization: a meta-analysis. In *International Journal of Human-Computer Studies, 53 (2000) 5*, 851-866.

[5]   CLEF (2006): Cross Language Evaluation Forum, URL: http://www.clef-campaign.org/.

[6]   Cugini, J., Laskowski, S, Sebrechts, M. (2000). Presenting Search Results: Design, Visualization and Evaluation. In: *Workshop: Information Doors - Where Information Search and Hypertext Link*, San Antonio

[7] Cumulus (2006): Canto Database Software Cumulus, URL: http://www.canto.de/pro/

[8] Dublincore (2006): Dublin Core Metadata Initiative, URL: http://dublincore.org/

[9] Guidelines Works of Art (2006): Categories for the Description of Works of Art, URL: http://www.getty.edu/research/conducting_research/standards/cdwa/index.html

[10] Koshman, S. (2005). Testing user interaction with a prototype visualization-based information retrieval system. In *Journal of the American Society for Information Science and Technology, 56(8) 2005*, 824-833.

[11] Mann, T. M. (2002). Visualization of Search Results from the World Wide Web, PhD thesis, University of Constance.

[12] Plaisant, C. (2004). The Challenge of Information Visualization Evaluation. In *Proceedings of Conference on Advanced Visual Interfaces AVI'04*.

[13] Rath, H. H. (2003): The Topic Maps Handbook. White Paper. URL: http://www.empolis.com/downloads/empolis_TopicMaps_Whitepaper20030206.pdf

[14] Reiterer, H., Tullius, G., Mann, T. M. (2005). INSYDER: a content-based visual-information-seeking system for the Web. In *International Journal on Digital Libraries, Volume 5, Issue 1, Mar. 2005*, 25 - 41.

[15] Reiterer, H. (2004). Visuelle Recherchesysteme zur Unterstützung der Wissensverarbeitung. In Hammwöhner, R.; Rittberger, M.; Semar, W. (Eds.): *Wissen in Aktion. Der Primat der Pragmatik als Motto der Konstanzer Informationswissenschaft. Festschrift für Rainer Kuhlen. Constance*, 1–21.

[16] Renz, M., Renz, W. (2000): Neue Verfahren im Bildretrieval. Perspektiven für die Anwendung. In *Contribution to 22th DGI-Online Conference*, May 2-4 2000, Frankfurt a.M.., URL: http://users.informatik.haw-hamburg.de/~wr/internFH/Renz_DGIonline.pdf

[17] Sebrechts, M., Vasilakis, J., Miller, M. et al. (1999): Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,. 3-10.

[18] Shneiderman, B., Plaisant, C. (2006). Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. In *Proceedings of the BELIV'06 Workshop, Venice* 61–77.

[19] SEPIA (2006): Safeguarding European Photographic Images for Access,   URL: http://www.knaw.nl/ecpa/sepia/workinggroups/wp5/cataloguing.html

[20] TREC (2006): Text Retrieval Evaluation Conference, URL: http://trec.nist.gov/.

[21] Vaughan, L. (2004): New measurements for search engine evaluation proposed and tested. In *Information Processing and Management 40 (2004),* 677-691.

[22] Veerasamy, A., Belkin, N. J. (1996). Evaluation of a tool for visualization of information retrieval results. In *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '96. ACM Press, New York, NY, 85-92.

[23] Wielinga, B. J., Schreiber, A. Th., Wielemaker, J., Sandberg, J. A. C. (2001): From thesaurus to ontology. In *Proceedings of the 1st International Conference on Knowledge Capture*, *International Conference on Knowledge Capture,* Canada. 194-201.

[24] Zwol, R. Van, Oostendorp, H. Van (1004). Google's "I'm feeling lucky", Truly a Gamble?. In Zhou, X. et al. (Eds..) (2004): *Web Information Systems - WISE 2004, Proceedings of the 5th International Conference on Web Information Systems Engineering.* Brisbane, Australia, 378-390.