In: Eibl, Maximilian; Wolff, Christian; Womser-Hacker, Christa (Hg.): Designing Information Systems. Festschrift für Jürgen Krause. Konstanz: UVK Verlagsgesellschaft mbH, 2005. S. 47 – 57.

Information Retrieval - From Information Access to Contextual Retrieval

Norbert Fuhr

1 Introduction

Information Retrieval (IR) deals with uncertainty and vagueness in information systems. Uncertainty is caused by the problem of representing the semantics of text and other media, which cannot be done in a perfect way. On the other hands, information needs to be answered by IR systems are often vague and cannot be specified precisely, thus leading to iterative query formulation.

The generic IR task can be specified as "Retrieve that amount of knowledge which a user needs in a specific situation for solving his / her current problem" (Kuhlen 1991). This definition implies two major research issues:

1. IR should consider the specific user, the situation and the problem to be solved. This view leads to the notion of contextual retrieval.

2. For retrieving the necessary knowledge, all accessible knowledge sources should be exploited; which requires methods for global information access.

These two topics also were the key issues describe in the final report of the workshop "Challenges in Information Retrieval and Language Modeling" (Allan & Croft 2003) which brought together 38 top-level researchers from the area of IR in fall 2002.

In the remainder of this paper, we will first describe current research in the area of global information access (section 2), followed by the discussion of work on contextual retrieval (Section 3). Then we will discuss issues for further research, before coming to the final conclusions.

2 Global Information Access

The workshop report (Allan & Croft 2003) defines global information access as follows:

Satisfy human information needs through natural, efficient interaction with an automated system that leverages worldwide structured and unstructured data in any language.
In order to deal with this topic, two major problems have to be addressed, namely appropriate access methods have to be developed, and the properties of the information to be accessed must be taken care of. These two issues are discussed in the following two subsections.

3 Information Properties

Important information properties investigated in current IR research are media, structure, and heterogeneity; besides, there are some minor issues worth mentioning.

Information media for which IR methods are developed are text, facts (for vague queries), 2D data like graphics and images, speech, video and 3D data, besides application specific data like e.g. in gene data banks. Most research in this area still focuses on texts. Although there is also substantial work on retrieval methods for other media, a major obstacle is still the lack of methods for automatic indexing of these media, i.e. constructing a representation of the semantics of such an object. Moreover, many applications require retrieval at the pragmatic level (e.g. in a newspaper photo archive, for illustrating an article), for which no automatic indexing methods available at the moment. Besides similarity searching (mainly at the syntactical level), multimedia indexing is feasible only in very limited domains (e.g. face recognition).

The *information structure* of objects to be retrieved can be classified into unstructured (like in classical text retrieval), semi structured (e.g. XML documents) or fully structured (as in standard databases). Furthermore, there may be hyperlinks between the objects (as e.g. on the Web). Although a lot of research has been dealing with Web retrieval recently, methods for more regular hyperlinked, semi structured data (which can be found in the ‘Hidden Web’ or in intranets) are still at an early stage.

Heterogeneity is a major problem when accessing several sources, which may differ with respect to language (addressed by multilingual retrieval), media (multimedia retrieval), structure (schema, ontology) and service functionality. For the latter two issues, standardization may help solving some problems, but the current standards (like Dublin Core for structures or the XQuery text retrieval proposal) address only very basic issues. The second strategy for dealing with heterogeneity is integration. Approaches from the area of databases aim at perfect mappings between different database systems. Since IR

has to deal with uncertainty and vagueness anyway, also vague schema mappings could be considered in this area. Standardization is the alternative way for addressing the heterogeneity issue. So far, only trivial structures (like e.g. the Dublin Core schema[1]) have been standardized; a similar statement holds for the services, where e.g. the XQuery text retrieval extension (Buxton & Rys 2003) is far from the current state of the art in IR.

Other information properties considered by some researchers are mostly related to mobile computing, dealing with location dependence ("find a good Italian restaurant nearby"), considering access bandwidth restrictions (based e.g. on GPRS or UMTS access) as well as I/O device properties like the display quality or I/O media. Closely related, time dependence of information plays an important role in certain applications (e.g. retrieval of business news).

3.1 Information Access Methods

In current IR research, a number of information access methods are investigated:

- *Ad-hoc retrieval* deals with onetime queries, like e.g. in web retrieval.
- *Filtering and Routing* use a constant search profile for filtering relevant documents out of a message stream.
- *Categorization and clustering* group a collection of documents into classes, which are either predefined (categorization) or adaptive (clustering).
- *Topic detection and tracking* cluster messages from an incoming stream.
- *Summarization* generates short summaries of single or multiple documents, either query specific or query independent.
- *Question answering* aims at retrieving short text passages for answering fact queries.
- *Information extraction* fills template records with facts from texts.

Although most of these methods are in practical use, it is still unclear whether or not these are the most important methods needed for applications. Also, often information searches need more than one method (see below), but the integration of these methods is yet to be addressed.

[1] Dublin Core Metadata Initiative (ed.) (2004). Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://dublincore.org/documents/dces/ [Zugriff September 2004].

3.2 Contextual Retrieval

The workshop report (Allan & Croft 2003) defines contextual retrieval as follows:

- Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs.

Here we will address three major context dimensions: Time, social and work context.

3.3 Time

Most IR research still deals with batch like retrieval (given a fixed query, find the best possible answer), where evolving information needs are not taken into account. In a similar way, typical filtering and routing methods are based on the assumption of a constant information need, but try to adapt to this need over time. IR methods for *personalization* try to detect and consider (thematic) personal preferences over time, but interest shifts are yet to be considered.

Interactive retrieval may be the most important context issue, but plays only a minor role in IR research so far. Evaluation studies have shown that quality differences between batch retrieval methods play no role when these methods are used interactively (Turpin & Hersh 2001). Also, the assumption of a constant search request (forming e.g. the basis of relevance feedback methods) does not hold, since empirical studies have shown that interactive retrieval consists of a sequence of interconnected but diverse searches (O'Day & Jeffries 1993).

As a new paradigm for interactive search Bates 1989 proposes the 'berrypicking' technique, where a user collects relevant documents from different searches; an (electronic) personal library can support this strategy and also provide long-term storage of search results.

Another weakness of most of today's systems is the lack of high level search functions. Based on empirical studies of the information seeking behaviour of experienced library users, Bates 1990 distinguishes four levels of search activities. Whereas typical information systems only support low-level search functions (so-called moves), Bates introduced three additional levels of strategic search functions:

- A *tactic* is one or a handful of moves made to further a search. For example, breaking down a complex information need into sub problems, broadening or narrowing a query are tactics applied frequently.
- A *stratagem* is a complex set of actions (comprising different moves and / or tactics) exercised on a single domain (e. g. citation database, tables of contents of journals). Examples for stratagems are subject search (searching for all documents referring to this subject), citation search (find all documents citing / cited by a given article) or journal run (browse through issues or complete volumes of a relevant journal).
- A *strategy* comprises a complete plan for satisfying an information need. Thus, it typically consists of more than one stratagem (e. g. perform a subject search, browse through relevant journals and then find the documents cited by the most important articles).

For offering these functions to the user, Bates distinguishes several levels of system support. Besides rigid system behavior (where the system only executes activities on command), the system may also act in a proactive way, either by recommending possible actions or by executing them automatically and presenting the result to the user. For example when a query returns no results, the system may try different methods for broadening the query(e.g. spelling correction, related terms, modification of query logic).

Advanced IR systems also should support long-term search activities (which are typically higher level activities). For this purpose, a system also should personalization functionality, e.g. for setting personal preferences, for keeping track of seen items, but also for following evolving interests.

3.4 Work Context

Most IR methods do not consider any work context at all – which may be due to the fact that IR still is regarded as isolated task, without considering the integration in application systems.

As an example of a work context, Allen 1996 describes a generic problem solving scheme, where he distinguishes three phases:

- Problem understanding, which could be supported by a hypermedia system with introductory or survey articles.
- Identification of possible solutions, for which hierarchical hypermedia system would be useful.
- Selection of the optimum solution, which is the only step which can benefit from an IR system.

So a system supporting problem solving should integrate the functionality of the different system types.

An alternative taxonomy of information seeking goals is described in Shneiderman 1998:

- specific fact finding (like finding a book record from its ISBN),
- extended fact finding (slightly more open questions, like finding in what area a researcher is working),
- open ended browsing (like finding what types of information seeking strategies are described in the literature) and
- exploration of availability (like getting an overview of what kind of information about Harry Potter is available on a web site).

Again, different kinds of system functionality are required for supporting the various tasks.

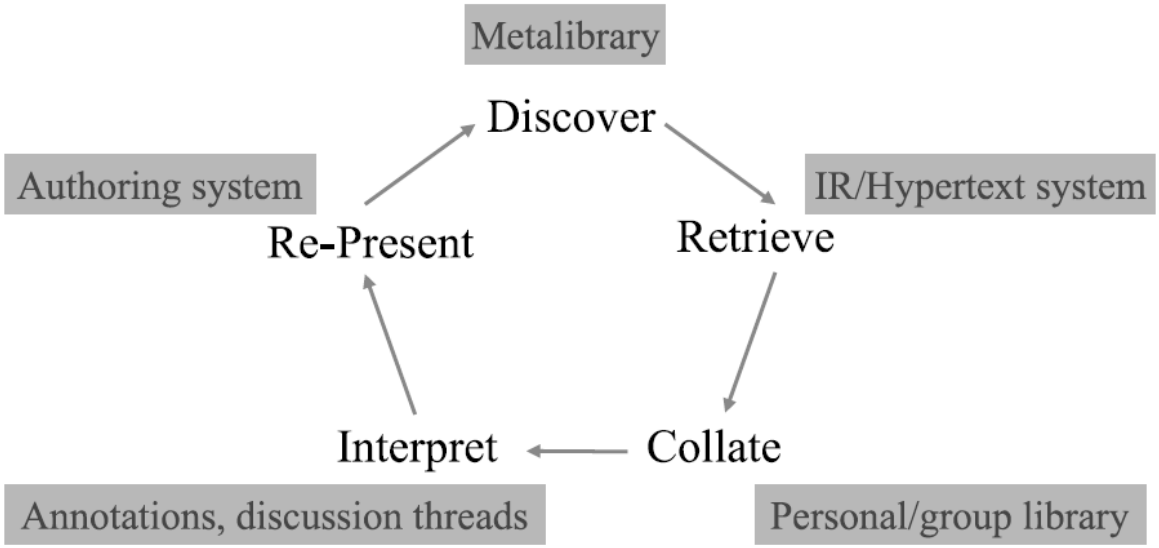

Figure 1: Digital library life cycle

Weibel & Miller 1997 describe the digital library (DL) life cycle, which can be seen as a generic workflow for knowledge workers (see Figure 1). First, a user has to discover potentially relevant sources, from which s/he can retrieve the documents s/he is looking for. In the collate step, the retrieved material is organized (e.g. by clustering related objects together). Then the documents are analyzed in the interpret step. Together with own research findings, the user represents the available material by writing new documents; once a document is completed, it is stored in a digital library, and the whole process starts all over. Figure 1 also lists the corresponding type of system (functionality) that is required for supporting the different steps. So isolated IR systems support users only in a very limited way.

A more general approach for considering the work context is described in Pejtersen & Fidel 1998, where a layered model for work centered evaluation and design is described (see figure 2). Based on the model of cognitive work analysis (Rasmussen et al. 90), work context description starts with the analysis of the work domain and the organizational context, then describes the activities in different terms (mental strategies, decision making, work domain), and finally characterizes the user and analyzes the ergonomics.

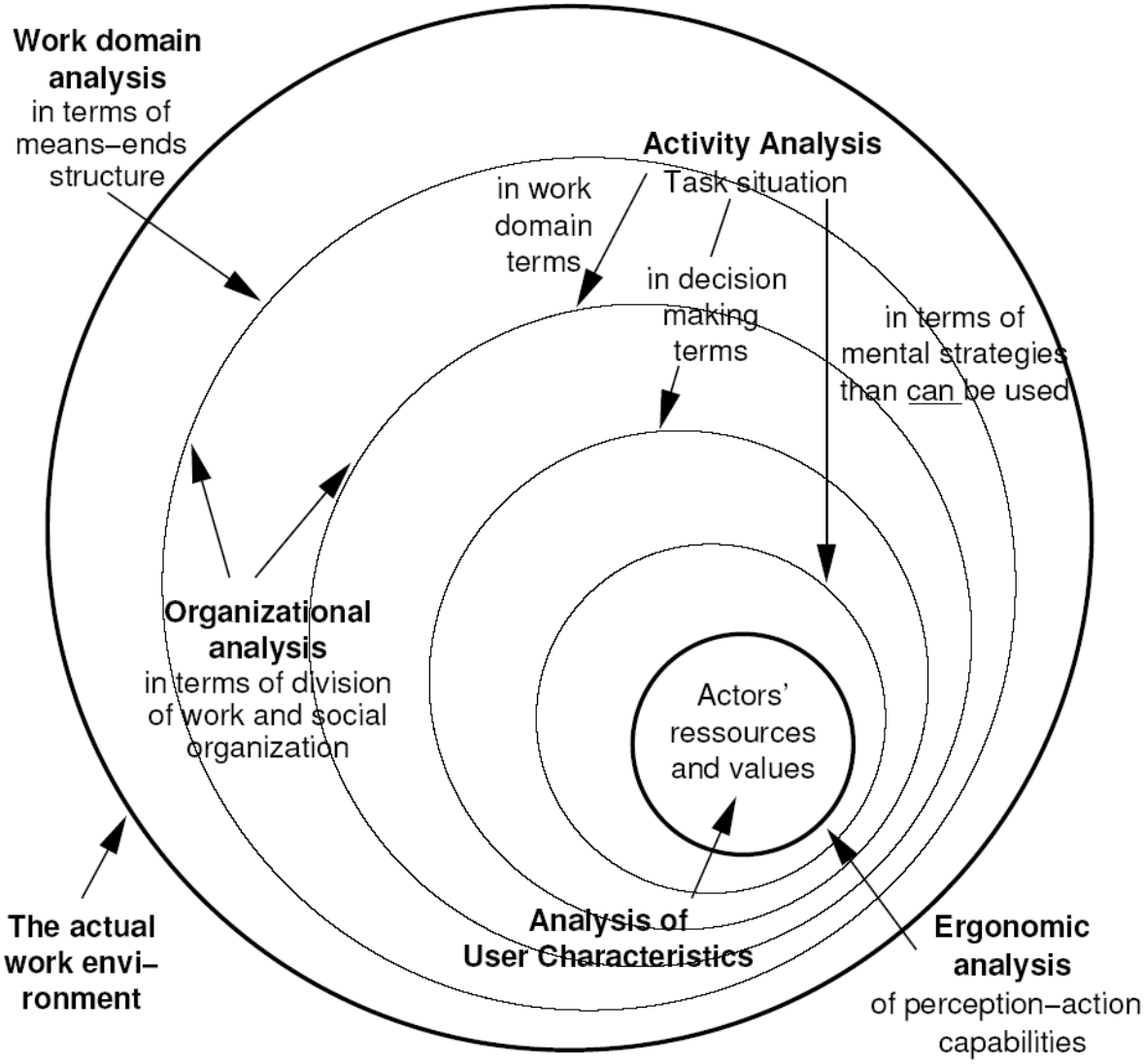

Figure 2: Basic model for work centered design and evaluation

3.5 Social context

Most IR systems are targeted at single users, thus ignoring the fact that many of them are working in teams where the members have similar interests. Thus, there is a need for supporting cooperative work. In fact, there are cooperative versions of all steps of the DL life cycle: For discover and retrieve, there are recommendation and collaborative filtering methods. By sharing folders

among the members of the group, the collate step can be supported, and a groupware system supporting discussion threads attached to stored objects (like e.g. Bentley et al. 1997) implements a group version of the interpret step. Finally, cooperative authoring systems implement the represent step. An important feature of groupware systems is awareness, e.g. for notifying other members of the group when a user has filed a new document.

In contrast to closed groups, open communities may be joined by anyone (like e.g. newsgroups on the Web or peer-to-peer filesharing systems), but high quality information access functions are still to be developed for this type of collaboration.

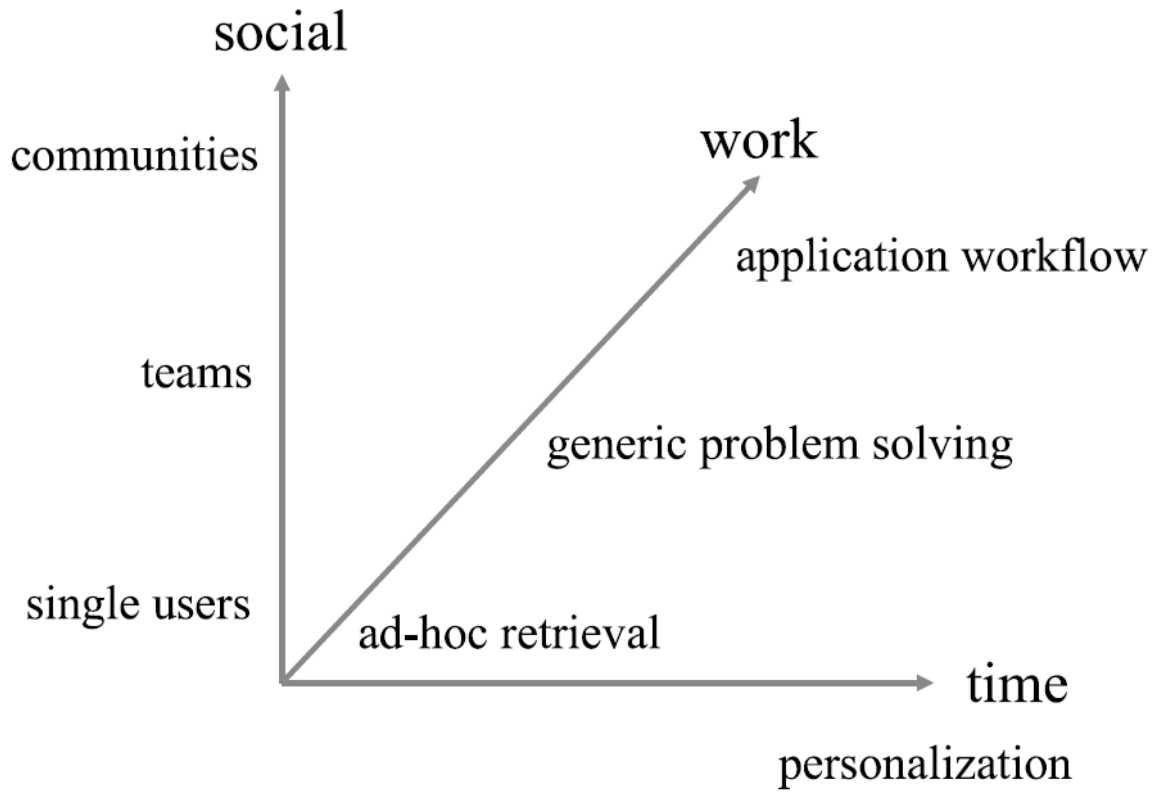

Figure 3: Context dimensions

3.6 Context Dimensions

In Figure 3, we illustrate the three major dimensions of context in IR. Classical IR approaches are located near the origin, considering almost no context. Today, there is an urgent need for systems that are positioned at higher values of the different axes.

4 Future Research

The major focus of current IR research is on models, methods and systems for information properties and access methods. A glance at the proceedings volumes of recent IR conferences shows that almost 90 % of all contributions belong to this area of global information access. However, most of this work deals with the optimization of known methods. On the other hand, there is very little research on contextual retrieval. There are several reasons for this imbalance:

- There is a lack of testbeds for this kind of research. There are several international evaluation initiatives for global information access (e.g. TREC[2], CLEF[3] , INEX[4], NTCIR[5]), but only the interactive track at TREC considers some aspects of contextual retrieval.
- In any case, the experimental effort for evaluating contextual retrieval approaches is higher. Observing real users during information seeking and analyzing the data gathered (questionnaires, logs, audio / video protocols) requires a manifold of the time needed for running and analyzing batch experiments.
- By definition, contextual retrieval is rather application specific. Thus, generalization of experimental results may be difficult. On the other hand, if the evaluation is restricted to generic activities, important facets of the application domain may be neglected.

For future research, there are only a few topics related to global information access that deserve continuing attention, like e.g. representation of the semantics of non-textual media, retrieval methods for structured documents as well as methods for copying with heterogeneous structures and services.

In the areas of contextual retrieval, all the issues described above require substantial research efforts. Especially, the consideration of time, social and work context seems to be a major challenge. Only through successful research in this area, there is a major chance for improving IR quality in a significant way.

---

[2] National Institute of Standards and Technology (NIST) (ed.) (2004). Text Retrieval Conference Homepage. http://trec.nist.gov/ [Zugriff September 2004].
[3] Cross Language Evaluation Forum (CKEF) (2004). CLEF Homepage. http://www.clef-campaign.org [Zugriff September 2004].
[4] Initiative for the Evaluation of XML Retrieval (INEX) (2004). Inex Project Homepage. http://www.is.informatik.uni-duisburg.de/projects/inex/ [Zugriff September 2004].
[5] NTCIR (NII-NACSIS Test Collection for IR Systems) Project Homepage. http://research.nii.ac.jp/~ntcadm/index-en.html [Zugriff September 2004].

5 Conclusion

IR deals with uncertainty and vagueness, which is intrinsic to all information seeking problems that cannot build upon a well structured domain; thus, approaches currently discussed under the 'Semantic Web' framework will not be suitable for IR unless they incorporate uncertainty.

We have described two major areas of IR research in this paper, namely global information access and contextual retrieval. Research approaches dealing with the former issue can be described as a combination of information properties and access methods, and in fact 90 % of current IR research falls into this area.

In contrast, we think that contextual retrieval should be given more attention. There are many issues that have not even been addressed yet (e.g. if the quality difference between batch retrieval methods vanishes in interactive retrieval, which methods make a difference in this setting?). Also, consideration of context offers the possibility of significant quality improvements (e.g. in contrast to context free Web searches with a few query terms). A major impediment for research in this area is the higher effort for considering the context of actual applications. Only through close cooperation between industry and research, progress in this area can be achieved.

6 References

Allan, J.; Croft, B. (2003). Challenges in Information Retrieval and Language Modeling. Report of a Workshop held at the CIIR, UMass Amherst, September 2002. SIGIR Forum 37(1), 31–48.

Allen, B. L. (1996). Information Tasks. Toward a UserCentered Approach to Information Systems. San Diego: Academic Press.

Bates, M. J. (1989). "The design of browsing and berrypicking techniques for the online search interface." In: Online Review 13(5), 407–24. http://www.gseis.ucla.edu/faculty/bates/berrypicking.html [Zugriff September 2004] .

Bates, M. J. (1990). "Where Should the Person Stop and the Information Search Interface Start?" In: Information Processing and Management 26(5), 575–591.

Bentley, R.; Appelt, W.; et al. (1997). "Basic Support for Cooperative Work on the World Wide Web." In: International Journal of Human-Computer Studies 46(6), 827–46.

Buxton, S.; Rys, M. (2003). XQuery and XPath FullText Requirements. Technical report, World Wide Web Consortium. http://www.w3.org/TR/2003/WD-xquery-full-text-requirements-20030502/ [Zugriff September 2004] .

Kuhlen, R. (1991). "Information and Pragmatic Valueadding: Language Games and Information Science." In: Computers and the Humanities 25, 93–101.

O'Day, V.; Jeffries, R. (1993). "Orienting in an Information Landscape: How Information Seekers Get From Here to There." In: Proceedings of the INTERCHI '93. Amsterdam: IOS Press, 438–45.

Pejtersen, A.; Fidel, R. (1998). "A Framework for Work Centered Evaluation and Design: A Case Study of IR on the Web." Technical report, Riso National Laboratory, Denmark.

Rasmussen, J.; Pejtersen, A. M.; Schmidt, K. (1990). "Taxonomy for Cognitive Work Analysis." Technical Report 2871, Riso National Laboratory, Roskilde, Denmark.

Shneiderman, B. (1998). Designing the user interface. Boston: AddisonWesley.

Turpin, A. H.; Hersh, W. (2001). "Why batch and user evaluations do not give the same results." In: Croft, W.; Harper, D.; Kraft, D.; Zobel, J. (eds.) (2001). Proceedings of the 24th Annual International Conference on Research and development in Information Retrieval. New York: ACM Press, 225-31.

Weibel, S.; Miller, E. (1997). "A Summary of the CNI/OCLC Image Metadata Workshop." In: DLib Magazine 3(1).