A Grounded and Participatory Approach to Collaborative Information Exploration and Management

Keiichi Nakata
Institute of Environmental Studies, The University of Tokyo
nakata@k.u-tokyo.ac.jp

Abstract

Classification of concepts and terms is an effective way of organising information for managing large information sources. In many cases, the classification schemes are devised through a painstakingly time-consuming process without user participation, and impose standardised terms as a result. However, in the context of group work in which individuals with diverse views and background work together, such an overhead and the inflexibility of term usage can undermine efficient information management and user satisfaction. Instead, a bottom-up approach can be combined in which users participate and construct such classification schemes as they explore the information space. The information organisation can hence be customised to the specific group needs and term usages are grounded in the information containers themselves allowing users to interpret the meaning of terms in context. In this paper, we describe a system that constructs concept indexes, which supports such a bottom-up process of information exploration and management.

1. Introduction

The expansion of broadband networks has brought about the demand for more digitised information not only from specialists but also from ordinary users. An increasing amount of information is provided on-line, and it has been constantly pointed out that searching and navigating the Web for necessary information becomes dire with this trend. Some search engines, including *Yahoo!* and *Altavista*, incorporate a hierarchical classification of topics to assist navigation, providing entry points for Web browsing. Needless to say, this is built upon the legacy of classification-based information management. It enables information seekers to search from the broadest of categories to the exact information sought. Once the optimal sub-category is reached, there is the expectation of finding the collection of books, documents, media, etc., that contain the information required. An appropriate classification scheme is thus powerful.

The question is how can we arrive at such a classification scheme. In many cases, classification schemes are devised through a painstakingly time-consuming process by so-called experts without the participation of users. Since there is no user participation, such schemes need to be general, and ideally should cover all kinds of category, and as the result imposes "standardised" terms. These aspects turn out to be restrictions, for instance, in the context of groupwork in which individuals with diverse views and background work together. The overhead of devising an all-encompassing general classification scheme can be prohibitive and the inflexibility of term usage can undermine efficient information exploration and user satisfaction. We observe that these problems arise as consequences of the top-down approach of information management. In this paper, we argue that a considerable part of the management of large and contextually ill-defined information is suited to be performed bottom-up, and describe a tool that is designed to enable combination of top-down and bottom-up information management

2. Top-down versus bottom-up information management

There are two aspects in which top-down and bottom-up approaches can be contrasted (Figure 1). The first is in the classification of concepts itself, i.e., how the concepts are identified and incorporated into the classification scheme. The second is in the process of developing such a classification scheme, i.e., who does the work in what way.

In Figure 1, the "editorial board" refers to a person or a group of people who are responsible for constructing the classification scheme. In the following, we compare the two approaches according to this matrix.

Classification development	Concept classification: Top-down	Concept classification: Bottom-up
Top-down	Development of classification schemes for pre-conceived (*a priori*) concepts by an editorial board	Development of classification schemes for *grounded* concepts by an editorial board
Bottom-up	Development of classification schemes for pre-conceived (*a priori*) concepts through individual participation	Development of classification schemes for *grounded* concepts through individual participation

Figure 1. A matrix of top-down and bottom-up approaches.

3. Concept classification

We refer to concept classification here as a general term for a family of hierarchical conceptual structures, including term classifications, topic categorisation, conceptual graphs, semantic nets, and ontologies. These are typically described by a tree or graph, with nodes denoting concepts and arcs denoting semantic relations such as the generalisation relation (class/subclass, is-a, ako, etc.) and the compositional relation (part-of etc.).

3.1. Abstract or *a priori* concepts

Abstract or *a priori* concepts are typically obtained through a systematic enumeration of conceptual entities. Here, names or labels that represent concepts are already identified and extracted, i.e., have undergone the process of symbolisation, through abstraction and internalisation of objects, ideas, entities, etc. For example, when a person is asked to list the species of a dog, she can retrieve from her memory and knowledge the conceptual labels such as poodle, German shepherd and golden retriever. These are preconceived, already conceptualised, and typically are already classified—as sub-concepts of dog. This is a classical example of the top-down approach—devising a classification, based on a preconceived conceptual structure. The task of constructing the classification is almost equivalent to knowledge elicitation. Asking an expert to build a taxonomy of possible failure modes of an automobile falls into this category. Generation of ontologies [8], in particular top-level ontologies, poses the same characteristics. For a rather different flavour of applications, the CoMeMo system that captures and brings together individual's everyday memory helps elicit and organise *a priori* concepts [9].

The main advantage of this approach is the completeness. A classification can be made complete by enumerating all possible sub-concepts of a concept, iterating the same process down to the instance level. However, we can immediately identify four major disadvantages. Firstly, depending on the level of detail of the top-level concept in the hierarchy and the richness of the diversity in the conceptual structure, this task can be rather demanding. If there were theoretically a large number of sub-concepts of the top-level concept and each subconcept had further as many sub-concepts, iterating this process down to the instance level can be a grand task. Secondly, even if the classification is complete, when applied to search for documents and data that would contain one of the (sub)concepts, there is no guarantee that any information can be retrieved. Finding out the lack of any information for a concept can be a useful indication when the task is to author an encyclopaedia, but it is an undesirable feature when looking for information about the concept; if no information exists for that concept, it gives a false expectation. Thirdly, when a classification is devised, for its inherent expectation of completeness and consistency, many decisions must be made in the process to "forget" some of the attributes or properties of a concept in order to fit it into a place in the conceptual hierarchy. Bowker [3] calls it *institutional forgetting*, and his case study concerning an attempt to devise a classification scheme for the activities in nursing profession illustrates this issue. The fourth point follows from the third, that a classification devised in this manner inevitably imposes a particular view since there is no concrete context. This problem arises later in this paper again concerning the process of classification development.

3.2. Grounded concepts

The drawbacks of the top-down approach to devising concept classifications described in the previous section arise from the fact the concepts themselves are abstracted and this gives an illusion that we can devise a complete and consistent classification. While this might be true in a scientific community in which the concepts are

systematically categorised such as in botany (but note the "institutional forgetting" that could have accompanied in the process), in more pragmatic settings such as newly emerging communities of individuals, this is unlikely. There must be a context in which the classification is devised (such as a group working together to tackle a particular problem) and a scope or boundary to which it is applied (such as a document collection that contains only a limited number of concepts). For such purposes, a bottom-up approach is suitable. For concepts to be placed in a context with a scope, they should be "grounded". We use this notion from the grounded theory approach in sociology [15]. A grounded theory is a theory that was derived as a result of the qualitative analysis of data such as documents, interviews, etc., and its process of theory construction guarantees that every concept has reference(s) to existing data—hence grounded. It is also important to exclude any preconceptions and prejudices at the initial phase to ensure objectivity. The process of grounded theory approach, which involves a fair amount of document management and indexing, has been made simpler by software tools such as NUD*IST [13], Atlas.ti [1], and win-MAX [16].

According to [15], the process of grounded theory technique begins by coding, i.e., analysis and interpretation of data by identifying concepts and categories (classes) from documents. Terms, phrases, and segments of text are collected, each interpreted and assigned a code (concept). Collected concepts can be grouped or placed in a hierarchy, from which a theory may begin to form. This projects a bottom-up approach. Further documents can be analysed to modify, reject, or re-enforce the theory. It is not the intension of this paper to discuss the validity or strength of the grounded theory approach. What is attractive here is that the concepts identified in this way are always attached to the original data. This means, there is no more concepts than there is evidence of its occurrence (scope) and the tacit assumptions employed, which are in many cases the context of work, influence the interpretation of data. Classifications (categories) are initially built bottom-up, since the task of coding involves identifying the relations between the concepts, which, in the grounded theory, are the essence of the resulting theory. These features complement the drawbacks of the top-down approach that deals with *a priori* concepts. One obvious deficiency is that classifications developed in this manner are often incomplete (or, more incomplete than the top-down approach), and thus fail to spot the lack of certain concepts which should be present. However, in case of information exploration, whose primary task is to seek for information rather than the lack of it, this is not an issue.

4. Process of classification development

So far we discussed and contrasted the top-down and bottom-up approaches in the concept classification task. Whether such a task is carried out by an individual, a small group or a large group was not considered. In this section, we focus on this issue. As shown in Figure 1, the process of classification development can be performed by 1) an individual or a small group dedicated for this task (an "editorial board"), or by 2) every individual who takes part in information exploration. We note here that this is an organisational issue and therefore depends on the work practice of the group or community. However, we deliberately disregard this element and focus on the pros and cons of each approach.

4.1. Editorial board

The first case in which the editorial board assumes the task of classification development has the following advantages. First, editorial board members are likely to be well informed and likely to have expertise in the task or the domain. This would steer the classification towards coherence and consistency. Second, the members are likely to be communicative and well equipped for discussions and negotiations concerning the classification scheme. There might even be a kind of organisational structure in which there is a chief editor supported by sub-editors etc., to make the process flow smoothly. Third, they can filter out noises, undesirable concepts, and provide an authoritative version.

The quality of the classification scheme described above is particularly important to ontologies, since they are intended to represent a consistent body of knowledge. As a result, ontology developers aimed at capturing knowledge of a community such as Ontobroker [7] takes this editorial board approach. For example, in their ontology development for the knowledge acquisition research community [2], they have an editorial committee, which is responsible for the integrity of the community ontology.

The editorial board approach too, however, has two main drawbacks. Firstly, the formation of the editorial board itself is an overhead. If we take this approach, the first thing we must do is to choose the members of the editorial board, and set up a sort of an organisational convention which governs how the classification

development should be carried out in the board, the protocol of communication between ordinary members and the board, and so forth. So before we will be able to enjoy the fruits of the information management there will be a rather long preparation phase. This is in fact a matter of trade-off—if the expected size and complexity of the classification scheme justifies this initial extra effort, it probably is worth it. On the other hand, much wanted information could be obtained by a less sophisticated method of exploration in less time than this initial set-up stage. The second drawback is shared with one of those we identified for the use of *a priori* concepts in Section 3.1: the classification scheme developed by an editorial board inevitably imposes, albeit implicitly, a particular view. While this would mean that there will be a conventional agreement that the view of the editorial board is an authoritative one, there will be a degree of dissatisfaction among the ordinary members on term usage, omission of terms, controversial views, etc. This is perhaps the inherent property of anything top-down.

4.2. Collaborative construction of classifications

The opposite extreme of the top-down editorial board approach is the open style, in which every member of the group can contribute to the classification scheme development. Under the assumption that individuals work together in a group to devise a classification scheme as a joint effort for a common task, this is a case for the collaborative approach. As a contrast to the editorial board approach, it can also be seen as the participatory approach. The drawbacks of the editorial board approach are the advantages of this style of development. Since everyone contributes, there will be minimal overhead at the start-up. And for the same reason, there will be no particular view imposed, and the resulting classification scheme can be seen more likely as the reflection of the group consensus, or at least ordinary members would have had an opportunity to voice their views more directly. Intuitively, this would improve user satisfaction. A major problem, however, is the lack of coherence and the high incidence of inconsistencies that follow from this rather anarchic approach. An inconsistent classification scheme would potentially result in confusion and undesirable consequences for any application that makes use of it. There are two distinct ways of dealing with this problem. The first is to maintain consistencies by supporting several "local versions" thereby guaranteeing the overall consistency. This approach is useful and necessary when the consistency maintenance is the utmost importance as in the case of ontology development. A good example is Ontolingua [6]. Ontolingua supports collaborative ontology development by taking into account the overall consistency. When a new piece of knowledge is added to the consensus ontology on the Ontolingua server, it is first checked for its consistency with the existing knowledge, and only added when there is no conflict. If there is, this new knowledge piece is qualified by some scope such as locality or even a person who contributed it. It is by such qualifications that each knowledge piece is guaranteed not to violate overall consistency [5].

The second approach addresses the problem by making it a non-problem: allow a certain degree of inconsistencies. This is possible in the application area in which inconsistencies pose no serious consequences, such as in the case of information exploration. Although it depends on the user expectation of the results of a query etc., in most cases inconsistencies such as circular hierarchical definitions ("A is a subclass of B, and B is a subclass of A") may reduce precision, but would increase recall. We can even go one step further by loosening the semantic restrictions of relations such as subclass and part-of, and reduce all relations to uni- or even bi-directional associations. This at the first sight seems to reduce the information content and hence has undesirable consequences. However, considering that people are notoriously bad at assigning the "right" relations between concepts (e.g. is "housing" a subconcept of the concept "welfare" or is it a part of it?) this is not such a bad solution. There is evidence that such a "weak semantics" in fact stimulates people to assign relations among concepts [9]. Moreover, this approach taken in conjunction with the bottom-up concept extraction approach, it would enhance the emergence of structure among dispersed, independent concepts. Such a feature can be useful for the initial stages of ontology development during which a form of brainstorming takes place (the "rapid prototyping" of an ontology).

5. Weighing the benefits

Obviously, the four modes of classification scheme development described in the matrix of Figure 1 represent four extreme cases, and in general these modes are mixed at different phases of development. The editorial board seldom works in isolation without any user evaluation and feedback; the concepts to be included are judged by the editorial board through their expertise and knowledge. However, we must emphasise that many of the existing tools and systems support primarily top-down approaches.

One argument that is worth putting forward here in favour of the collaborative approach is that with the ever-increasing amount of information on the Internet, collaboration might be one of few ways to practically manage the complexity of information management. The rate of growth in the information pool is already unmanageable by an individual. By joining efforts in a group or community, we might be able to maintain the reasonable coverage. This is possible in the case of Internet to which almost anyone can have access and initiate his information exploration. With the tool for such exploration being made universally available, so should also be tools for collaborative work that encourage such a participatory mode of information exploration and management.

6. Concept Index

We have developed a prototype of a tool that supports grounded and participatory information exploration and management. This Web-based tool supports the construction and management of concept indexes. Operationally, a concept index consists of a collection of documents that provides the scope of the index, and a set of interrelated concepts which are expressed by a set of text pieces (words, phrases, etc.) that appear in the documents and provide handles to the aggregated and shared interests of the user group that shares the same index. It is intended as groupware and incorporates the notion of users and user groups, with access rights concerns.

Three primary features of concept indexes are as follows.

- Concept-based orientation: concepts can provide a survey on the content of the document collection and enable users to navigate by the content of documents (Figure 2).
- Provide visual cues concerning the relevance of a document regarding group interests: concept occurrences are highlighted and cross-referenced (Figure 3).
- Offer a simple user interface for contribution to the index: identify and register important or interesting concepts while browsing a document, and organise them in a structure (Figure 3).

Due to the space limitation, here we only provide an overview of the concept index tool. More details on technical issues and discussions on this tool can be found elsewhere [10, 11].

6.1. Concepts and relations in a concept index

Every entity that represents an object, conceptualisation, idea, person, place etc., in a concept index is a concept. A concept can be verbally expressed in terms of words, phrases, sentences, paragraphs etc., and also by other concepts. A concept may be expressed by a number of different expressions. For instance, the concept named “agent-based system” can be expressed by, and therefore may appear in documents as, “agent-based system” (coincidence of the textual expression and concept name), “agent application”, “applications of agents” or a descriptive piece of text. Note that by using a highlighting mechanism in a browser, users can assign to the concept “agent-based system” these textual expressions from documents.

On registering the concept, these textual expressions are analysed so that on finding the occurrences of this concept in documents, they would match more flexibly than strict string matching. For example, “applications” should find the occurrences of application, applied, apply etc., since they are simply different forms of the same word. To achieve this, each word is stored by its “normalised” form; in this case, word lemmatisation (transformation to dictionary entry form) is used. The normalised form together with stemming can achieve flexible matches. The search for concept occurrences in the document collection is a computationally intensive task; hence it is delegated to agents that execute this mundane task persistently. In the current design, we support two types of relations, comprise and associated. Comprise relations are intended to capture the notion of collective concepts. For example, concepts “distributed system” and “FIPA agents” each can be seen as a part of, or a sub-concept of “agent-based system”. However, there is no strict semantics attached to this relation; hence, any grouping of concepts can become comprising concepts of a concept. Associated relations are intended to capture the notion of associations, or loose relations, between concepts. For example, the concept “BT Labs” can be associated with “agent-based systems”.

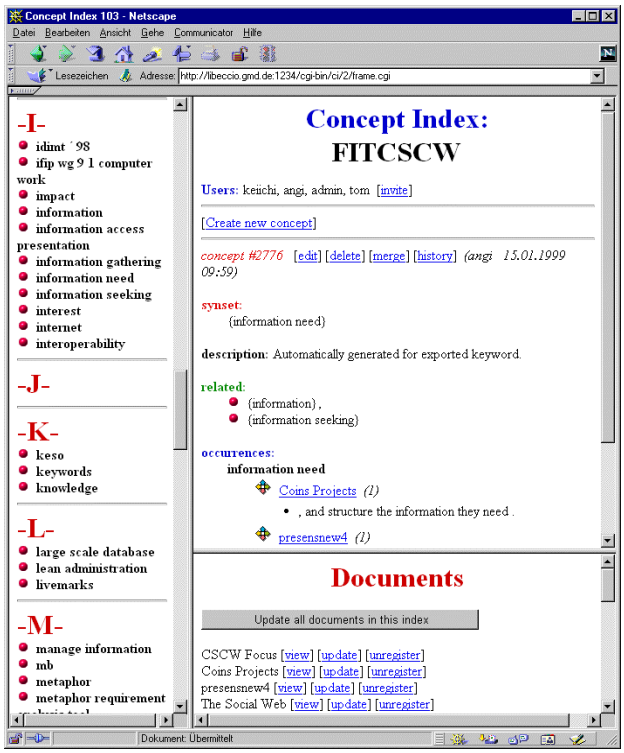

Figure 2. A screenshot from the concept index tool. The left frame lists phrases which can be selected to display the concept containing the chosen phrase in the concept frame (top right). The concept frame describes the concept in terms of a) synonymous phrases that describe the concept (*synset*), b) annotation concerning the concept, c) list of associated concepts (clickable), and d) list of documents that contain the concept. The bottom right frame lists the documents that are indexed.

This document viewer is spawned whenever a document is opened from the concept index tool (Figure 2). Occurrences of concepts that are registered in the current index is highlighted in colour (default red) followed by two icons: the "i" icon when clicked displays the selected concept in the concept index tool which displays other occurrences of the concept in the document collection; clicking the neighbouring triangle icon jumps to the next occurrence of the same concept within the document. Registering a new concept can be performed in this browser by highlighting a phrase by the mouse, the result of which is displayed in the bottom frame (this screenshot demonstrates the highlighting of, hence the addition of the phrase (concept) "social spaces" to the index).

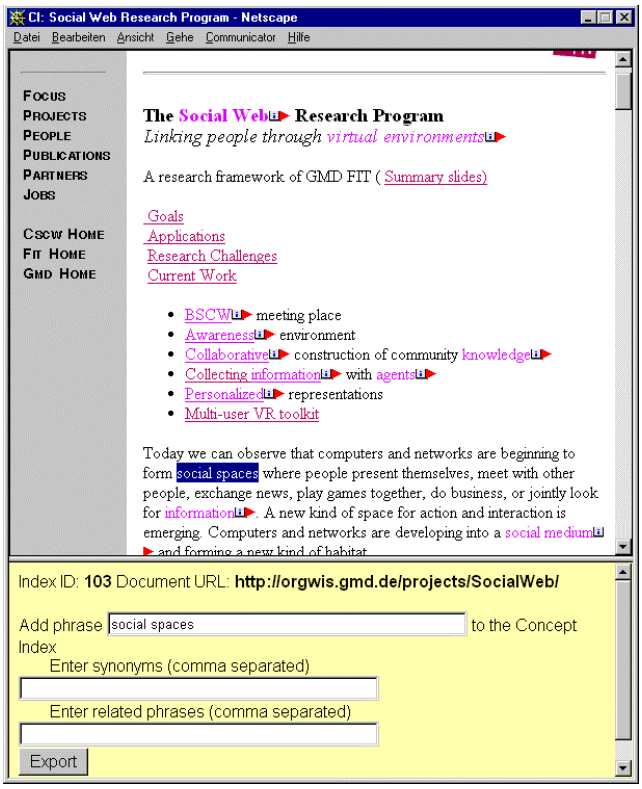

Figure 3: A screenshot of a Web page enriched with information from a concept index.

Since neither relation has strict semantics, these two relations can overlap with each other. The distinction is rather practical: users may wish to group several concepts and create a new concept out of them; this is a case of comprise relation. When two concepts are simply seen to be closely associated, but not necessarily to be grouped into another concept, associated captures such a relation.

6.2. Groundedness and collaboration in concept index

The concept index tool is intended to address the bottom-up approaches both for concept classification and its construction process. The concepts are grounded since they are identified during the users' document browsing activity and registered into the index as they find them in the text. Note that concepts are not given any definitions; this is precisely because all the concepts should be interpreted in their usage thereby always being within their context. This feature has an additional advantage that where there is a concept entry in the index, it is without exception that there is a document that contains it (although in case the document that initially contained the

concept was removed from the collection or modified, the concept occurrence may disappear). From these concepts collected from the document collection, an organisation or structure of concept relations would emerge, as a product of the same motivation that urge and necessitate people to organise files into folders and directories—for better access and ease of management.

The weak semantics of relations in a concept index should ease collaboration and encourages user participation. For the purpose of information exploration assumed in this context, inconsistencies that would inevitably arise are not too harmful. If there is a need for consistency, a straightforward inconsistency check can be carried out by the system and be left between the users to resolve them as appropriate.

Applications we envisage include construction and management of virtual libraries, distributed project management, virtual schools and remote teaching, all of which potentially benefit from grounded and participatory features of concept indexes. Although we have emphasised the bottom-up aspect in the concept index tool, there is nothing in the system that prevents the top-down approach. A concept can be registered simply by supplying the text expression without any reference to its occurrence in any of the documents. And within the user group, they may choose to name a single person responsible for constructing the concept index, allowing others only to browse documents and navigate through the concepts but not to contribute to the index. Top-down and bottom-up approaches must be interleaved to best suit the objective of the group activity.

7. Related work

The comparison between the qualitative analysis using grounded theory and the construction of faceted classifications in library and information science has been conducted by Star [14]. Following a comprehensive analysis of the parallels between these two methodologies, Star argues that they can complement each other and suggests that faceted classifications can be applied to the analysis and construction of grounded theories. Our observations and approaches support her stance, and Concept Index contributes to the implementation of her claim.

An instance of the group effort in realising participatory information source is the Open Directory Project [4]. Based on the criticism that keyword based automated search engines often contain meaningless information, it maintains the contents by a group of voluntary editors. Editors are solicited for each category of the directory and since they control the contents, it follows, irrelevant and low-quality information sources (i.e. Web pages) can be minimised. The search directories constructed in this manner has been used by *Netscape*, *Lycos* and *Hotbot*. This can be seen as a form of "editorial" approach suggested in this paper. However, the manner in which the directory headings are chosen and created is unclear, since one can only volunteer to become an editor of an existing category. From this point of view, it does not embody the bottom-up approach to information classification, but would serve as a social filtering tool. While there are positive comments about the reliability of the Open Directory, it has been criticised for the opacity of editor selection process.

8. Conclusions and further work

In this paper, we have argued for more emphasis on bottom-up approaches to concept-oriented information management systems to complement the presently dominant top-down approaches. A bottom-up approach in concept classification offers focus and context to classification schemes, and the same approach in collaborative development adds flexibility without imposing prescriptive classification schemes. The degree to which the bottom-up approach is combined with the top-down process inevitably depends on the nature of the group and the complexity and *natural structure* of the information space. However, as Ranganathan pointed out [13], it is rather unusual that any classification should be single-faceted, and by considering multiple facets we are more likely to discover alternative analyses of the information space, as exemplified by grounded theory.

We have described the concept index tool, which aims to enhance the bottom-up approaches in these domains. The current prototype is still preliminary and there is room for more improvements and extensions. They include more variety in concept index input interface in addition to the current highlight-while-you-browse mouse interface, such as pen-style scanners and tablet-style Web browsers. The improvement of performance and incorporation of more groupware features and facilities are also considered. Furthermore, attaching text pieces longer than phrases and images to concepts will enhance the expressiveness of concepts in the index. Through the usage of this tool, we plan to conduct a series of experiments to evaluate the system and assess the claim made in this paper.

Acknowledgements

The author would like to thank the members of the GMD-FIT Coins Project, Angi Voss and Marcus Juhnke with whom the concept index tool is developed, and anonymous reviewers for their valuable comments on the earlier version of this paper.

References

[1] ATLAS.ti. ATLAS.ti—The Knowledge Workbench. http://atlasti.de/

[2] V. R. Benjamins and D. Fensel. Community is Knowledge! in $(KA)^2$. In Proc. 11th Banff Knowledge Acquisition for Knowledge-Based System Workshop (KAW98), Banff, 1998.

[3] G. C. Bowker. Lest we remember: organizational forgetting and the production of knowledge. http://alexia.lis.uiuc.edu/~bowker/forget.html

[4] dmoz. The Open Directory Project. http://www.dmoz.com

[5] A. Farquhar, R. Fikes, W. Pratt, and J. Rice. Collaborative Ontology Construction for Information Integration. Technical Report KSL-95-63, Stanford Knowledge Systems Laboratory, 1995.

[6] A. Farquhar, R. Fikes, and J. Rice. The Ontolingua Server: A Tool for Collaborative Ontology Construction. Technical Report KSL-96-26, Stanford Knowledge Systems Laboratory, 1996.

[7] D. Fensel, S. Decker, M. Erdmann, and R. Studer. Ontobroker: The Very High Idea. In Proc. 11th International FLAIRS Conference (FLAIRS-98), Sanibal Island, Florida, 1998.

[8] T. Gruber. Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL 93-04, Stanford University Knowledge Systems Laboratory, 1993.

[9] H. Maeda, M. Kajihara, H. Adachi, A. Sawada, H. Takeda, and T. Nishida. Weak information structures for community information sharing. International Journal of Knowledge-Based Intelligent Engineering Systems, 1(4):225-234, 1997.

[10] K. Nakata, A. Voss, M. Juhnke, and T. Kreifelts. Collaborative Concept Extraction and Management. In U. Reimer (ed.), Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98), Basel, 1998.

[11] K. Nakata, A. Voss, M. Juhnke, and T. Kreifelts. Concept Index: Capturing Emergent Community Knowledge from Documents. In P. Marti and S. Bagnara, editors, The 7th Le Travail Humain Workshop “Designing Collective Memories”, Paris, 1998.

[12] S. R. Ranganathan 1965. The Colon Classification. In S. Artandi (ed.) The Rutgers Series on Systems for the Intellectual Organization of Information, Vol. IV, Graduate School of Library Service, Rutgers University, 1965.

[13] QSR Software. NUD.IST 4. http://www.qsr.com.au, accessed 23 March 1999.

[14] S. L. Star. Grounded Classification: Grounded Theory and Faceted Classification. 1996. Available at http://alexia.lis.uiuc.edu/~star/gt.html

[15] A. Strauss and J. Corbin. Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Sage Publications, London, 1990.

[16] winMAX. winMAX Qualitative Data Analysis. http://www.winmax.de/heade.htm