Actors, Domains, and Constraints in the Design and Construction of Controlled Vocabularies

Jens-Erik Mai

University of Toronto, Faculty of Information Studies,
140 St. George St., Toronto ON Canada M5S 3G6,
<je.mai@utoronto.ca>

Jens-Erik Mai is a member of the faculty at the Faculty of Information Studies, University of Toronto. He is interested in the intersection between classification, information, and human activities; especially in the networked society and for context specific goals. Jens-Erik holds a Ph.D. in Library and Information Science from the University of Texas at Austin and a Master's and a Bachelor's degree from the Royal School of Library and Information Science, Denmark.

Mai, Jens-Erik. **Design and Construction of Controlled Vocabularies: Analysis of Actors, Domain, and Constraints.** *Knowledge Organization, 35(1)*, 16-29. 48 references.

ABSTRACT: Classification schemes, thesauri, taxonomies, and other controlled vocabularies play important roles in the organization and retrieval of information in many different environments. While the design and construction of controlled vocabularies have been prescribed at the technical level in great detail over the past decades, the methodological level has been somewhat neglected. However, classification research has in recent years focused on developing approaches to the analysis of users, domains, and activities that could produce requirements for the design of controlled vocabularies. Researchers have often argued that the design, construction, and use of controlled vocabularies need to be based on analyses and understandings of the contexts in which these controlled vocabularies function. While one would assume that the growing body of research on human information behavior might help guide the development of controlled vocabularies shed light on these contexts, unfortunately, much of the research in this area is descriptive in nature and of little use for systems design. This paper discusses these trends and outlines a holistic approach that demonstrates how the design of controlled vocabularies can be informed by investigations of people's interactions with information. This approach is based on the Cognitive Work Analysis framework and outlines several dimensions of human-information interactions. Application of this approach will result is a comprehensive understanding of the contexts in which the controlled vocabulary will function and which can be used for the development of for the development of controlled vocabularies.

1. Introduction

Classification schemes, thesauri, taxonomies, and other controlled vocabularies play important roles in the organization and retrieval of information in many different environments. While the design and construction of controlled vocabularies have been prescribed at the technical level in great detail over the past decades, the methodological level has been somewhat neglected. Classification research has in recent years focused on developing approaches to the analysis of users, domains, and activities that could produce requirements for the design of controlled vocabularies (Hjørland & Albrechtsen 1995; Pejtersen & Albrechtsen 2000; Nielsen 2001). Researchers have argued that the design, construction, and use of controlled vocabularies need to be based on analyses and understandings of the contexts in which these controlled vocabularies function.

While one would assume that the growing body of research on human information behavior might help guide the development of controlled vocabularies, unfortunately, much of the research in this area is descriptive in nature and of little use for systems design (Fidel et al. 2004). In this paper, I will outline a holistic approach that demonstrates how the design of controlled vocabularies can be informed by investigations of actors' interactions with information, the work they do, and the domain in which they are located. This approach is based on the Cognitive

Work Analysis framework (Rasmussen, Pejtersen, & Goodstein 1994; Vicente 1999) and outlines several dimensions of human-information interactions that could be investigated to determine factors that shape actors' information needs and use. Application of this approach will result in a comprehensive understanding of the contexts in which the controlled vocabulary will function, and can be used for the development of systems requirements.

2. Methods for the design and construction of controlled vocabularies

A controlled vocabulary can be defined as "a list of terms that have been enumerated explicitly" (ANSI/NISO 2005, 5) for the purpose of organizing and representing information to facilitate information retrieval. Controlled vocabularies vary in complexity from simple alphabetic lists of terms to classification schemes and taxonomies that show semantic relationships, and to complex thesauri that furthermore show associative relationships between terms (ANSI/NISO 2005; Rosenfeld & Morville 2007). The steps that a developer of controlled vocabularies can take have been well described in the literature (cf. e.g. Aitchison, Gilchrist & Bawden 2000; Lancaster, 1986; Soergel, 1974). These steps are often represented as some version of the following:

1. Analyze literature, needs, actors, tasks, domains, activities, etc.;
2. Collect, sort, and merge terms;
3. Select descriptors and establish relationships;
4. Construct the classified schedules; and,
5. Prepare the final product.

The latter steps—steps 2 through 5—are well prescribed and worked out in great detail in several standards, well-established textbooks, and best practices. These steps deal with *technical* aspects of the design and construction of controlled vocabularies, including guidelines and rules-of-thumb for how to determine appropriate form of the terms, clarify the meaning of terms, factor compound terms, determine the relationship between terms, etc. While these are important aspects and techniques that must be mastered by developers of controlled vocabularies, design decisions throughout these steps must be guided by the outcome of the first step. However, the first step—analysis of literature, needs, actors, tasks, domains, activities, etc—has been somewhat neglected in the literature. The advice given for the first step is often limited to either simply mentioning that the designer needs knowledge about the context of the controlled vocabulary (cf. e.g. Aitchison, Gilchrist, & Bawden 2000, 7-10) or to suggest that a list of potential terms is drawn up by subject experts or is selected or extracted from the content of the objects (cf. e.g. ANSI/NISO 2005, 91-92).

As it has been demonstrated in the literature, the selection of terms to represent the subject matter of documents is rather complex (cf. e.g. Wilson 1968, Hjørland 1992, Mai 2001; Mai 2005). The exact procedure for selecting the terms and the procedure that a controlled vocabulary designer should or could follow is debated in the literature; this debate has generated several approaches one can follow, none of which has emerged as predominate or logically most excellent. Blair (1990, 163) commented on this issue several years ago:

> Scientific taxonomies are built around *observable* differences between members of categories. These differences, though often subtle, must be objectively verifiable (a zebra *must* have stripes, a fish *must* have gills). But when we distinguish documents by subject categories, what objectively verifiable criteria can we use? None has been established.

Although there is much debate in the scientific community about the epistemological status of categorical claims about scientific objects (cf. Bryant 2000), Blair's point is well taken; the determination and selection of documents' subject matter cannot be done in a systematic way that ensures that the same subject matter will be identified independently of time, place, and person performing the determination.

The decisions a designer of controlled vocabularies needs to make at this stage are informed by the designer's epistemological stance. This epistemological stance frames the *methodological* aspect of the design and construction process. As demonstrated by Hjørland (1998) a domain can be organized in multiple equally valid ways, depending on the particular epistemological stance taken. Hjørland argues that there are four basic epidemiological approaches that one can take; each of these four approaches leads to specific methodological consequences:

1. Empiricist: Documents are clustered based on *statistical analysis* of resemblance.
2. Rationalist: Classification based on *logical division* and/or *eternal and unchangeable* categories.

3. Historicist: Classification based on a notion of natural *development or evolution*.
4. Pragmatic: Classification based on an analysis of *goals and usage.*

Each of these approaches could lead to different classifications of the domain. It would be difficult to argue that any of these approaches is more correct or better that the others–they are just different, based on different assumptions, leading to different classifications.

That being said, I would argue that consideration of goals and usage of documents and of controlled vocabularies' purposes, would lead to design and construction of controlled vocabularies that are more closely in correspondence with actors' activities, needs, and demands. This result follows from a line of thinking that argues that a controlled vocabulary "is always required for a [specific] purpose, why a consideration of that purpose is the most important part of the methodology of information science" (Hjørland & Pedersen 2005, 585). To accomplish this, a clear methodological framework for studying actors and activities for the purpose of designing controlled vocabularies is needed. This paper outlines such a framework for the design of controlled vocabularies that shows which analyses and decisions a designer have to make in the first step of the development of controlled vocabularies.

3. Users, domains, and information interaction

3.1. Users' information needs and behavior

One of the major challenges for the representation and organization of information is that information does not have any meaning in itself, but only to somebody in particular contexts. Indexers therefore have to guess the meaning and subject matter users will attribute to the information. Users, on the other hand, are in an equally difficult situation. Users must attempt to describe the content of the documents that are desired and they frequently therefore must describe something that is unknown to them (Belkin, Oddy, & Brooks 1982). This has sometimes been described as a process of uncovering the users' actual information need and expressing this in a compromised form as a request (Taylor 1968), and as discovering and developing the topics of interest through interactions with the search mechanism (Kuhlthau 1993). Users furthermore have to express information needs as search requests using appropriate search terms and they are thereby forced to guess which terms indexers might have chosen for documents that contain relevant information. Considering that the goal of information systems is to match users' information needs with the product of indexers' work, there is surprising little research that addresses these issues together.

After a review and analysis of a range of models of information behavior, Wilson notes that: "the various areas of research within the general area of information behaviour may be seen as a series of nested fields" (Wilson 1999, 262-263), where "information behavior" is the broad general area of study, "information-seeking behavior" is a sub-set of information behavior, and "information-searching behavior" is a sub-set of information-seeking behavior (and sub-sub-set of information behavior), as illustrated in figure 1.

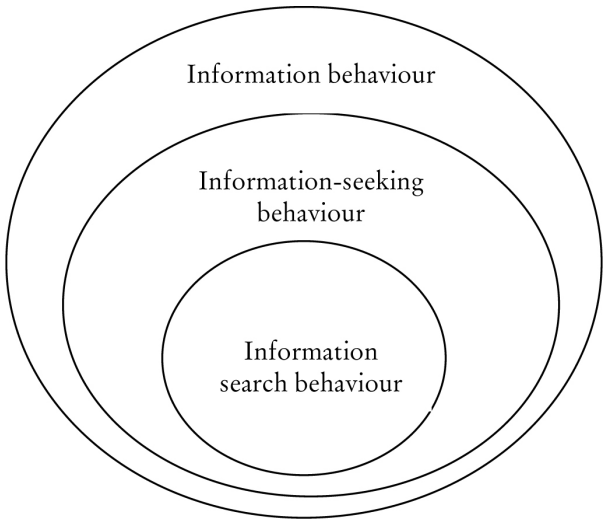

Figure 1. *A nested model of the Information Seeking and Information Searching research areas (Wilson 1999, 263)*

As research narrows from general information behavior research to information seeking, it is "concerned with the variety of methods people employ to discover, and gain access to information resources" (Wilson 1999, 263), and to information searching that is concerned with "the interactions between information users (with or without intermediary) and computer-based information systems" (Wilson 1999, 263). The focus of the research has been on the users' experience with information use, seeking, and retrieval with the aim of understanding this in greater detail. Most recent research in information behavior focuses almost exclusively on descriptions of users' interactions with information at various

levels of detail and does not include analysis and recommendations for systems design. While this research provides insight into the users' information behavior in particular situations, the goal of the research has not typically been to produce design requirements that can guide the organization and representation of information and/or the design of controlled vocabularies. Furthermore, much user-centered research in information retrieval and information behavior has focused on the needs, use, and behaviors of individual users and has attempted to develop models of how and why people seek and use information (Pettigrew, Fidel, & Bruce 2001). This focus on individual users has greatly advanced our understanding of how people interact with information. However, information behavior research has left nearly all the organizational and environmental context outside or in the peripheries of models of information behavior. To gain an understanding of users' information interactions and usage, it is important to study then in context of real situations (Fidel et al. 2004).

To inform design of controlled vocabularies, studies of human information behavior need to investigate many more aspects than simply the methods users employ to seek information and their interactions with the system. Limiting investigations to users' interactions with particular systems generates a rather narrow view of the users and their activities. To gain a fuller understanding of their activities, the designer needs to consider not only information seeking activities, but also activities of which the information behavior is part. By limiting the investigations to users' information behavior and focusing on users' interacting with particular systems, one essentially does not take into account the majority of users' activities and those activities that define users' information needs. By expanding the units of analysis from humans' interactions with information systems to include all activities humans engage in that might lead to information needs, the number of influential factors grows significantly. Furthermore, many of these factors, while they might influence people's behavior, are uniquely personal and cannot be accounted for in the design of controlled vocabularies.

The challenge is to study people in context, in order to account for significant factors, while, at the same time, keep the complexity of the analyses manageable and informative for the design of controlled vocabularies.

3.2 Actors and domains

Most information behavior studies "have been primarily descriptive, if in different ways" (Fidel et al. 2004, 942) and have often focused on individual users in an effort to understand and describe how individuals interact with information systems in particular situations. Limiting investigations of humans to their interactions with systems, as users, keeps the focus on system—and humans' interactions with that system. However, to understand the full complexity of humans' activities, to understand the context that creates their information needs, and since "users don't think of themselves as *primarily* having anything to do with the computer at all ... [but] as professionals, working with others, and using computers to support those interactions" (Lamb & Kling 2003, 200), it is fruitful to think of them as *actors*. In this context, the term 'actors' simply means *humans that are involved in activities*. These activities can be directly related to information seeking or they can be other activities in which the actor is involved.

The goal of changing the focus from "users" to "actors" is to change the focus from the systems to the humans. When the focus is on "users," as defined as humans interacting with systems, the focus tends to be on the human-system interactions and the systems take a predominate position and humans are only viewed as interacting with systems. If the focus is changed to "actors," the focus will be on the human-activities interactions. A human-activities interaction focus will provide an insight to the full complexity of the tasks in which humans are engaged. Furthermore, a human-activities interaction focus centers the investigation of humans on those activities that need to be supported by systems and thereby does not take into account personal idiosyncrasies and other non-activity related matters.

When the focus of the human-centered approach is expanded to actors and activities in which they are engaged, the broader context in which these activities take place becomes equal important. It is recognized that individuals are part of a context that shapes their behavior, and this context needs to be included in our understanding of how individuals act and make decisions about information (Hjørland 1997; Nielsen 2001; Lamb, King, & Kling 2003; Lamb & Kling 2003; Fidel et al. 2004; Fidel & Pejtersen 2004). For the design of controlled vocabularies it is important is have a good understanding of the information needs and problems people might have as they start seeking information. However,

since "the context that gives rise to an information need is an inherent part of the requester's understanding of his/her information need" (Hertzum 2003, 175) it is necessary and important to gain an understanding of that context. In fact, just understanding current needs and problems might not help the designer of controlled vocabularies understand how to structure a controlled vocabulary and which terms to include. Furthermore, an individual's information needs "can only be understood within the framework of a systems theoretic model of the communication structures of knowledge domains" (Hjørland 1997, 174); to gain an understanding of future requests and information needs requires an understanding of the domains in which actors function. The domain provides a context for actors and creates the information needs for which they seek information.

The notion of domains has been used in the information science literature for some time now, most predominantly in the conception introduced by Hjørland and Albrechtsen in the mid-90s. Their use of the notion of domain is purposely broad and inclusive (Hjørland & Albrechtsen 1995, 400):

> The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study knowledge-domains as thought or discourse communities, … The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective.

The object of Hjørland and Albrechtsen's paper was to steer the attention of information science research away from studying individuals toward gaining understandings of context for the purpose of systems design. While their paper was successful at that, it lacks a more concrete notion of how to operationalize the notion of a domain. I have elsewhere defined the notion of domain as "an evolving and open concept that will develop as the concept is used and applied in research and practice. [T]he concept is [here] used to refer to *a group of people who share common goals*. A domain could, for instance, be an area of expertise, a body of literature, or a group of people working together in an organization" (Mai 2005, 605). While this discussion makes it clear that the notion of domain is not limited to or the same as scientific disciplines and it focuses on human-activity interactions, the definition is still vague and difficult to operationalize.

Rasmussen, Pejtersen, and Goodstein (1994) give a slightly different definition of the concept of domain. Instead of defining the notion in relation to discourse communities or human activities, they define a domain as "the system to be analyzed and thus it represents the landscape within which work takes place" and as such the domain is "independent of particular situations and tasks" (Rasmussen, Pejtersen, and Goodstein 1994, 28). The domain is then the context, the landscape, in which actors operate and this landscape is defined and analyzed independently of the activities that take place in the landscape. Simon's fable about an ant making a zigzag path on a beach to find its way home, explains this best (Simon 1969, 24):

> Viewed as a geometric figure, the ant's path is irregular, complex, hard to describe. But its complexity is really a complexity of the surface of the beach, not a complexity in the ant. On that same beach, another small creature, with a home at the same place as the ant, might well follow a very similar path.

Focusing solely on the ant, his decisions and behavior will reveal a host of factors that influence his path through the beach, but many of the factors are best understood and explained in relation to the constraints and obstacles he encountered on the beach along the way. To understand why the ant followed a particular path, one needs to understand both the ant and the beach, and "to ignore the latter is to make an enormous mistake because it is not possible to understand the ant's path without an understanding of the contributions made by the beach" (Vicente 1999, 150-151).

When the ant is substituted with a human actor, we find a similar pattern, namely that "the apparent complexity of his behavior over time is largely a reflection of the complexity of the environment in which he finds himself" (Simon 1969, 25). The focus, thereby, changes from looking at individuals to investigating the context in which they operate to understand their decisions and choices. When turning to design of controlled vocabularies, we investigate the context because the "path taken by a human actor through a work space can only be explained on the basis of the complexity of the work space together with the goals and resources of the actor" (Rasmussen, Pejtersen, and Goodstein 1994, 36). In other words, we cannot explain the actors' actions by investigating actors' interactions with particular systems,

and we cannot explain the actors' activities by looking at the actors alone, we need to include analysis of the context, the landscape, the domain in which they operate. The description and designation of the particular domain to be analyzed depends on the goal and purpose of the design; there is no set way to determine domains, the "identification depends on a pragmatic choice of boundary around the object of analysis that is relevant for the actual design problem. This choice depends on the circumstances" (Rasmussen, Pejtersen, and Goodstein 1994, 35).

I propose that both the notion of actor and the notion of domain are needed for a contextual, human-centered approach to the design of controlled vocabularies. The notion of actor is important to view humans more broadly than just interacting with systems and domain is useful and important to set the situation in which actors operate. A third notion that is equally important is the notion of constraints; this notion ties actors and domains together and lays the foundation for a formative design approach.

3.3 Behavior-shaping constraints

Constraints are factors external to individual actors but common to all actors within a particular domain. The challenge is to identify those constraints that shape actors' information behavior—and not just to identify actors' specific preferences, perceptions, and experiences.

The focus on shared constraints is necessary to avoid having systems design being guided by descriptions of actors' observed behavior. Designers need in-depth knowledge about the constraints that shape actors' information behavior to determine what "*can* take place, or what strategies *can* be used, independently of how observed actors interact with current systems" (Fidel & Pejtersen 2004). This moves the design approach from a normative approach, which legislates the situation by describing how things should be, to a descriptive approach, which portrays the situation by describing how things are, and then further to a formative approach that describes how "things could be by identifying novel possibilities" (Vicente 1999, 112). Descriptive approaches have demonstrated that actors "do not, cannot, and should not consistently follow the detailed prescriptions of normative approaches" (Vicente 1999, 94) and that greater insight can be achieved by focusing on the context-conditioned variability of situations. No matter how detailed descriptions of particular situations are, they do not provide for systems design, because designers need more than just information about current practice; design that is based only on current behavior and practice can offer very little new.

Furthermore, the goal of the new system should not be merely to support current practice, but to allow for future possibilities and practices as well. While the descriptive approach provides insight into the domain, it does not provide for systems design. The goal of the design is to improve current technologies and practices; to do so, designers need to approach the situation differently than simply building on descriptions of current practice. Current practice is achieved and simultaneously limited by the use of existing technologies and practices, design techniques that are restricted to descriptions of current practice, no matter how thorough, therefore can only improve current conditions. A descriptive design approach cannot suggest new and innovative technologies and practices (Vicente 1999).

The constraints that shape behavior of actors in particular situations are the parts of the context that limit and enable actors to perform their work. It is important to recognize this duality of constraints; constraints limit and enable actions at the same time. For instance, a scholarly domain's history, schools of thought, and paradigms *limit* as well as *enable* actors to act in the domain; the constraints thereby shape possible information needs. A domain's history, for instance, enables actors to formulate questions and inquiries about particular phenomena by providing a narrative of the evolution of the knowledge about the particular phenomena. Simultaneously, the domain's history limits the kinds of questions and inquiries actors can pose about the phenomena by providing current, consensual understanding of phenomena. In other words, without a domain's history we would not know how we came to the current understanding of particular phenomena, so it enables us to pose questions and inquiries. At the same time, the domain's history provides a context for how questions and inquiries about the particular phenomena can be framed today.

Understanding the behavior-shaping constraints gives designers insight into the context of actors' work and provides an understanding that facilitates systems design. The outcome is not a prescription of what actors should do (a normative approach) or a detailed description of what they do (a descriptive approach), but an analysis of the constraints that shape the domain and context.

While the information behavior of individual actors varies enormously "there is something that remains relatively constant, and thus can be analyzed" (Vicente 1999, 151); the actors act within a given set of constraints that remain relatively stable from person to person. Designers of controlled vocabularies do not need to know how a particular person would employ a specific search strategy "or what exact circumstances would motivate the person to this strategy selection" (Fidel & Pejtersen 2004) to make decisions about the design; the designer only needs to understand the "possible strategies for people in a particular context" (Fidel & Pejtersen 2004) in order to design controlled vocabularies that support such strategies. In other words, designers need a map that gives a picture of the constraints that limit and enable information behavior of actors; they need a map of the beach, the domain. One framework that offers such a holistic approach is Cognitive Work Analysis.

4. Cognitive Work Analysis

Cognitive Work Analysis (CWA) provides a framework for analysis of actors' activities, domains, and preferences. The outcome of the analysis gives the designer an understanding that facilitates the creation of design recommendations and designers can use the recommendations to make decisions about systems design. CWA provides a holistic approach for studying human-information interaction in which it is possible to account for several different dimensions of activities and examine those dimensions in-depth and in context.

The CWA framework has been presented and discussed in general terms (Vicente 1999; Rasmussen, Pejtersen & Goodstein 1994) and also by the information science community in particular (Fidel & Pejtersen 2004; Pejtersen & Fidel 1998). A few information science studies have applied CWA; it guided the development of a retrieval and classification system for fiction (Rasmussen, Pejtersen, & Goodstein 1994; Pejtersen 1989) and informed the analysis of data collected in a study of Web searching by high school students (Fidel et al. 1999). More recently, a project to support multi-institutional collaboration in indexing and retrieval among three national film archives used CWA (Albrechtsen et al. 2002; Hertzum et al. 2002), as did a study of collaborative information retrieval among engineers (Fidel et al. 2004; Fidel et al. 2000).

The CWA framework views human-information interaction in the context of goal-driven activities. The activities are "steered by some goals, whether explicit or implicit, personal or organizational, stable or situational" (Fidel et al. 2004, 942). CWA analyzes actors' work activities, their organizational relationships, the constraints of the work domain and environment, and the actors' personal preferences and priorities. In other words, CWA focuses simultaneously on individual actors, actors' tasks, and the contexts in which actors operate. A graphic representation of the framework is given in figure 2.

Each circle in figure 2 represents a dimension that impacts on human activities and decisions. Each dimension represents a number of attributes, factors, or variables that can be analyzed. Each dimension presents a range of constraints for the dimensions that it holds. In other words, the work domain presents a set of constraints on the activities that can take place in a particular work domain. The environment (1), work domain (2), organization (3), and activities (4) together constrain the actors' resources and values.

CWA is attractive as a framework for studying human information behavior to support design decisions about controlled vocabularies because it is flexible and rigorous at the same time. It provides a structure for studying complex phenomena without limiting and directing how the phenomena can be understood. With CWA, one can examine the information behavior of actors in a domain and create a map of the domain that can be used to design domain-centered controlled vocabularies. This map is created by analyzing each CWA dimension and determining the factors, attributes, or variables that are important for the design of each dimension.

4.1. Dimensions for analysis of human-information interaction

While previous research into user-centered controlled vocabularies typically has focused on one or a few factors, the CWA framework focuses on multiple dimensions simultaneously. Rather than enumerate all or some of the factors that influence actors or activities, CWA gives a small set of dimensions each of which contains constraints of various sort. This is illustrated in figure 2 and will be discussed below. The factors that might influence actors are vast, whereas the number of potential constraints is somewhat limited. The first step in the development of a controlled vocabulary—as discussed in section 2 of this paper—involves the analysis literature, needs, actors, tasks, domains, activities, etc. It is for this

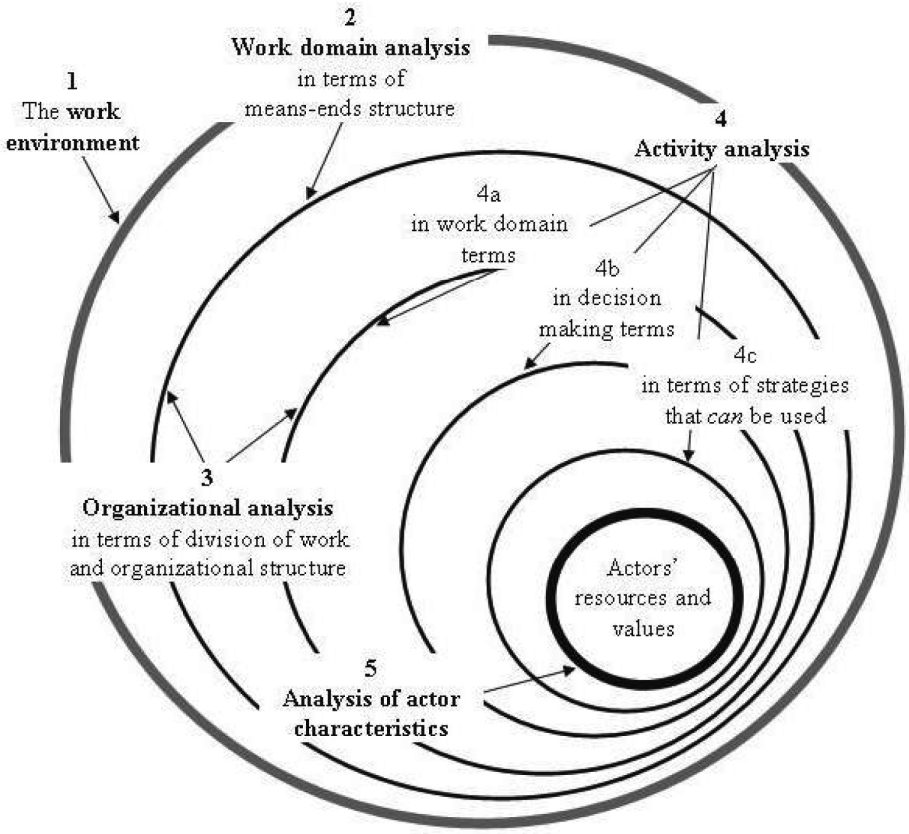

Figure 2. *Dimensions of Cognitive Work Analysis*

purpose the CWA framework is useful. The remaining four technical steps in the development process will be informed by the outcome of the analysis in step 1, but is as such outside the scope of this paper.

The issues to be addressed at each dimension vary from domain to domain, depending on the type of domain, goals of the controlled vocabulary, and activities that the system should support. I will define and discuss each dimension below and demonstrate how it contributes to the analysis carried out in step 1 in the development process of controlled vocabularies. I will keep the discussion of the dimensions at general level to show the strength of the CWA framework and demonstrate how it brings together many different factors that usually have been addressed in isolation. The purpose of the following is not to present a method for construction of controlled vocabularies, but merely to present a conceptual framework for conducting the analysis of the first step in the development process.

Dimension 1. The environment consists of the elements outside the actors' domain that affect their domain. The analysis of the environment reveals the context within which the actors operate and provides an understanding of the constraints under which the actors develop their information needs.

Actors in a given scholarly domain, for instance, are constrained by the domain's discourse, history, schools of thought, paradigms, research fronts, activities, etc.; these constraints limit and enable the types of information needs actors can have in the particular scholarly domain. An example of a constraint is Darwin's Theory of Evolution. Because of its importance, Darwin's Theory of Evolution limits and enables certain questions in biology, genetics, theology, and other areas of study by framing the areas in a particularly theoretical way. Any scholar in those areas has to respond to Darwin's theory in some sense. The theory is a constraint because it limits and enables the types of inquiry that are possible in those areas. Likewise, a commercial R&D division that is engaged in the development of web search engines is constrained by the long tradition of research in information retrieval. The creation of a controlled vocabulary for the intranet of the R&D division is influenced by the tradition of research in information retrieval. This history circumscribes and allows

certain kinds of questions about information retrieval techniques to be asked and addressed.

While it might be difficult to illuminate all the constraints that exist in a particular environment, it is important to be aware of their potential influence on the types of information needs that develop in the domain, even if the actors are not consciously aware of them.

Dimension 2. Work domain analysis examines the work that is done in the work domain; the work domain can be thought of as the landscape in which actors operate. Actors in particular work domains are constrained by a number of factors within the work domain and the purpose of the work domain analysis is to tease out the complexity of the work domain. Both the constraints within the work domain and those in the environment limit and enable the actors' information needs, the difference is that the constraints in the environment are outside the control of the actors in the work domain, whereas those within the work domain are controlled by actors within the work domain.

The work domain shapes actors at many different levels: at one end of the spectrum are the goals of the domain, which direct the actors' activities and at the other end of the spectrum are the resources that are available, which determine what is actually possible. Furthermore, actors are constrained by the priorities, functions, and processes that take place in the work domain. The various levels of constraints are interdependent and related; each provides the means for purposes (ends) at other levels. For instance, the processes that take place in the domain serve as means for the functions. The interdependence and internal relationships of the various levels of constraints in the work domain is teased out by an analysis of this means-ends structure.

Such a means-ends analysis teases out the structure and complexity of the work domains with the aim of understanding the constraints that affect the actors' information needs. The actors' information needs are generated in the work domain and reflect the constraints in the work domain, because the work domain provides the framework in which the actors operate. For instance, actors in information retrieval research are constrained by the goals, priorities, functions, processes, and resources of their particular work domain; while researchers in the domain share some of the constraints in the environment, their particular work domain presents other constraints that are unique to the work domain. Actors in a commercial R&D web search research division work under constraints that are significantly different from actors in a university setting. While actors in these two work domains may work on the same problem, the fact that they operate in different work domains—with different goals, priorities, functions, processes, and resources—will cause them to approach the problem differently. This difference affects their information needs and how they search for information, which should determine how the information is to be indexed.

The particular work domain will set certain constraints for the actors which will influence their information needs. The work domain analysis gives the designer insight into the domain and provides further understanding for the situation in which the actors' information needs develop.

Dimension 3. Organizational analysis examines how work is divided among actors in the work domain and examines the nature of the work domain. While analyses of the environment and landscape of the work domain give some insights into actors' work constraints, the organization of the work provides an insight into how the work is distributed among the actors. The organization of the work provides additional constraints to the actors' activities and potential information needs.

Workplaces are analyzed in terms of their organizational structures, management styles, organizational culture, nature of the organization, and allocation of roles. The organizational analysis gives the designer an understanding of how the domain is structured both explicitly and implicitly. While actors in research workplaces might have a high degree of autonomy in the work and their information needs therefore might develop relatively independently of the organizational structure, actors in more structured organizations, like an insurance company, develop their information needs in accordance with their particular tasks. Actors in such organizations are often assigned specific tasks and they develop information needs in reaction to these assigned tasks. An understanding of the organization is therefore needed to gain an insight into how work is delegated, assigned, or otherwise acquired.

The organizational analysis determines the constraints imposed by the structure, culture, and values of the organization. The designer uses this knowledge to make decisions about how the information can be organized and presented to the actor. The analysis may show, for instance, that actors in a

workplace do not communicate and collaborate as the workplace has been described in organization charts and it may therefore be wise to organize explicit systems for information flow according to how actors actually collaborate and work.

Dimension 4. Activity analysis examines what actors do to achieve their tasks. The CWA framework divides analysis of activities into three separate analyses: a) activity analysis in work domain terms, b) activity analysis in decision-making terms, and c) activity analysis in terms of strategies that *can* be used.

Dimension 4a. Activity analysis in work domain terms considers specific tasks that actors perform from the same perspective as that of the work domain (see 4.1.2 above). It illuminates the goals, constraints, priorities, functions, processes, and resources of specific actors' activities and tasks and establishes the means-end relations among these activities and tasks. The analysis provides a detailed view of the individual actors' work and framework in which they develop the information needs and information search behaviors.

Actors' activities are constrained by not only the environment, work domain, and organizational structure, but also by their activities. The activity analysis in work domain terms teases out the nature of the actors' tasks to understand how, where, and when they need information. While actors' activity based information needs may vary, the constraints under which the information needs develop might exhibit some stability. The purpose of this analysis is to determine these constraints. Actors in an insurance company, for instance, might want information about claims, police reports, photos of damaged material, etc. for their work and they often prioritize precise, compliant, updated information about the issues; however, these needs develop in response to specific activities the actors perform. We could ask whether the documents are needed to address a specific issue in a class action suit or they are needed in response to the retention schedule. Designers of controlled vocabularies need to understand actors' work activities to understand the difference between these two types of information needs and to make decisions about the organization and representation of the material. Likewise, faculty members at universities search for information in relation to their scholarly activities; they need information for their classes, their research, and their service activities and often priorities finding accurate, correct, and complete information on a given subject. We could ask whether a scholar is interested in a document in preparation for a class presentation or to confirm specific ideas when reviewing a colleague's manuscript. These activities constraints the types of information people are interested in and the type of information system they will use. Without an understanding of these activities and constraints, designers would not know how to design useful controlled vocabularies.

The activity analysis in work domain terms provides designers with a detailed understanding of the actors' work tasks and information-seeking tasks and the tasks' contexts in terms of goals, constraints, and priorities.

Dimension 4b. Activity analysis in decision-making terms examines the decisions actors make while performing their work activities. While much information is needed to support and help people making decisions, this analysis provides an understanding of the actors' work in terms of decision making and focuses especially on the information they need to make decisions and which sources provide useful information.

The purpose of this analysis is to clarify what information people need to make decisions, what information is actually available, and what information is desirable but not available. Researchers in a commercial R&D division might need information about a specific functionality in a search engine and they might be able to find this information in their personal files, intranet, public digital libraries, etc. Their search for this information is constrained by the decisions they have to make; depending on whether they are exploring issues related to functionality or are searching for design requirements, they will need different types of information. This difference in types of information needs will influence the design of the controlled vocabularies for this work place.

An analysis of actors' activities in decision-making terms provides designers with insight into the type of decisions actors make and their potential information requests.

Dimension 4c. Activity analysis in terms of strategies that *can* be used examines search strategies that can be adopted by actors to find information relevant for specific decision-making activities. While the work domain activities shape the decisions that actors make, the decisions they make shape the search strategies they can use.

The search strategies employed by actors in current systems can be good indicators of preferences in their search situations and might be valuable to understand as background for the formulation of search strategies in future systems. However, actors' current search behavior might not be relevant in future information systems and analysis of actors' activities in terms of strategies should focus on possibilities for searching and not be limited to descriptions of current practice. The analysis of strategies should therefore ask questions about possible strategies that actors can take, independently of whether actors actually use those strategies today. To identify possible strategies, the analysis would examine which strategies an actor could use to find specific information in an effective way; for instance could the actor search by using index terms, browse the system, or go directly to sources that are known to him/her.

The goal of the activity analysis in terms of strategies is to identify constraints that shape actors' possible and effective search strategies. Designers can use this knowledge to make decisions about which search strategies the system should offer in terms of controlled vocabularies, natural language searching, browsing, term coordination, etc.

Dimension 5. Analysis of actors' resources and values examines actors' experience, expertise, training, preferences, and values, and aims to identify characteristics for each group of actors in the domain. The actors' resources and values are constrained by the outer dimensions; the environment, the work-domain, the organizational structure, and the activities, and, as such, the resources and values of interest for this analysis are those that shed further light on the constraints facing actors.

The purpose of this analysis is to gain insight into the actors' cognitive resources and values, such as their knowledge of the subject matter dealt with in the domain, their preferences for information sources and format of information, values in terms of objectivity vs. subjectivity in representation of information. For instance, while designers of systems for actors in a scholarly domain might expect a certain level of subject knowledge, the information sources used in scholarly domains might vary among different actor groups. An analysis might find that senior researchers in the domain prefer short conference papers while students prefer review articles and monographs; such a finding should have an impact on the design of the controlled vocabulary. One would expect the controlled vocabulary to be able to make a distinction between the types of material and be able to distinguish between a topic covered in a review article and the same topic covered in a conference paper. Likewise, such an analysis might reveal that researchers in a commercial R&D division prefer more recent information in digital formats that contains lots of graphics representations.

The reasons for these preferences among the actors in these two domains can be attributed to the actors' resource and value constraints and while the specific actors might have very different preference and values, the constraints are stable across actors within specific actor groups. Designers can use knowledge about these constraints to understand how the information could be presented.

4.2. Summary

By moving the focus from descriptions of what actors do, to analysis of constraints under which actors operate, studies of human-information interaction can become useful for design. It is more useful because design of controlled vocabularies cannot be based on knowledge about individuals' behavior, preferences, and idiosyncrasies; design of controlled vocabularies is better served with analyses of the constraints under which actors operate. These constraints remain relatively stable over time and among different actors and therefore serve as better guides for the design of controlled vocabularies.

Cognitive Work Analysis (CWA) provides a useful framework for analyzing human-information interaction with the purpose of designing domain-centered controlled vocabularies. While factors that can affect human-information interaction are almost unlimited, the CWA framework offers a number of dimensions along which one can identify various constraints that influence the development of actors' information needs.

Each dimension contributes to the designer's understanding of the domain, the work and activities in the domain, and the actors' resources and values, and the analyses ensure that the designer brings the relevant attributes, factors, and variables to the design work. While analyses of each dimension do not directly result in design recommendations, they do rule out many design alternatives and such analyses offer a basis for which the designer can create a system for the particular domain. To complete the design, the designer needs expertise in the advantages and disadvantages of different types of indexing languages, the construction and evaluation of indexing

languages, and approaches to and methods of subject indexing.

5. Discussion

While others have suggested approaches to improve the design of controlled vocabularies, the CWA approach outlined in this paper is unique because it brings together many of the elements discussed by others. Hjørland (2002; 2004; Hjørland & Albrechtsen 1995) proposes a domain-analytic approach to Information Science and he outlines a number of methodological approaches that can be taken to study domains. The core of Hjørland's domain-analytic approach deals with establishing the constraints in what in the CWA framework is called the environment (dimension 1). Nielsen (2001) outlines a framework for studying professional domains and demonstrates that her mixed method approach provides sufficient understanding to design indexing languages. The majority of Nielsen's study focuses on the work-domain dimension in the CWA framework (dimension 2). Foster and Gibbons (2005) applied anthropological participant observation methods to study faculty members' work for the purpose of improving the design of an institutional repository. The core of their study is an analysis of the activities in work-domain terms (dimension 4a). Soergel's (1985) work on request-oriented indexing and Derr's (1982; 1984a; 1984b) and Saracevic's (1980; 1983) typology of information requests are mainly based on an analysis of actors' activities in decision-making terms (dimension 4b). Marchionini (1995) examined analytic and browse search strategies in detail to find ways to design systems that support both strategies; this approach closely matches CWA's activity analysis in terms of strategies (dimension 4c). Lastly, much of the work in the "cognitive viewpoint" movement (e.g. Ingwersen 1992) looked at the cognitive resources and constraints of individuals and much of their work matches CWA's analysis of actor characteristics (dimension 5).

While work along one or a few of the dimensions in the CWA framework bring further insight into the complexity of human-information behavior and strengthen the effort to design better information systems, a multi-dimensional framework is needed to capture the full complexity of the phenomena and move towards design of systems based on studies of actors and domains.

6. Conclusions

Designing information systems that facilitate the matching of actors' information needs with relevant documents is a challenging task, both in complexity and in importance for the success of such systems. While much information behavior research has focused on descriptions of how individuals seek and use information, and indexing research has increased our understanding of the technical aspects of the representation of documents, there have been few attempts to bring these two areas of study together.

To move towards a domain-centered approach to design of controlled vocabularies, knowledge and expertise from indexing needs to be infused with knowledge and expertise from information behavior. However, due to its complexity and contextual dependencies, the design approach taken needs to be based on a formative approach. A formative design approach with a focus on studying and understanding the information behavior constraints that actors face in particular domains, is the best foundation for moving towards domain-centered controlled vocabularies. An understanding of these constraints provides the designer with the right type of information to analyze the domain and recommend design features. Cognitive Work Analysis offers a possible framework to analyze information behavior that can lead to design recommendations. The advantage of using this framework is that it outlines relevant dimensions for analysis and provides tools for analysis and modeling.

References

Aitchison, Jean, Alan Gilchrist & David Bawden. 2000. *Thesaurus construction and use: a practical manual*, 4th ed. Chicago; London: Fitzroy Dearborn.

Albrechtsen, Hanne, Pejtersen, Annelise Mark, & Cleal, Bryan. 2002. Empirical work analysis of collaborative film indexing. In Bruce, Harry, Fidel, Raya, Ingwersen, Peter and Vakkari, Pertti eds., *Emerging frameworks and methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science*. Greenwood Village, CO: Libraries Unlimited, pp. 85-108.

ANSI/NISO. 2005. *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. Z39.19-2005. American National Standards Committee, National Information

Standards Organization. Bethesda, Maryland: NISO Press.

Belkin, N. J., Oddy, R., and Brooks, H. 1982. ASK for information retrieval. *Journal of documentation* 38: 61–71.

Blair, D.C. 1990. *Language and representation in information retrieval*. Amsterdam: Elsevier Science.

Buckland, Michael K. 1979. On types of search and the allocation of library resources. *Journal of the American Society for Information Science* 30: 143-47.

Bryant, Rebecca. 2000. *Discovery and decision: Exploring the metaphysics and epistemology of scientific classification.* Cranbury, NJ: Associated University Presses.

Derr, Richard L. 1982. A classification of questions in information retrieval by conceptual presupposition. In A.E. Petrarca, C. I. Taylor, & R.S. Kohn, eds., *Information interactions: Proceedings of the American Society for Information Science Annual Meeting* 19: 69-71.

Derr, Richard L. 1984a. Information seeking expressions of users. *Journal of the American Society for Information Science* 35: 124-28.

Derr, Richard L. 1984b. Questions: Definitions, structure, and classification. *RQ* 24: 186-90.

Fidel, R., Bruce, H., Pejtersen, A. M., Dumais, S., Grudin, J., & Poltrock, S. 2000. Collaborative information retrieval (CIR). *The new review of information behaviour research: Studies of information seeking in context* 1: 235-47.

Fidel, R., Davies, R. K., Douglass, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., Miyagishima, B. K., & Toney, C. D. 1999. A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science* 50: 24-37.

Fidel, R., & Pejtersen, A.M. 2004. From information behaviour research to the design of information systems: The Cognitive Work Analysis framework. *Information research* 10:1, paper 210. [Available at http://informationr.net/ir/10-1/paper210.html].

Fidel, R., Pejtersen, A.M., Cleal, B., & Bruce, H. 2004. A multidimensional approach to the study of human-information interaction: A case study of collaborative information retrieval. *Journal of the American Society for Information Science and Technology* 55: 939-53.

Foster, N.F., & Gibbons, S. 2005. Understanding faculty to improve content recruitment for institutional repositories. *D-Lib magazine* 11:1.

Hertzum, Morten. 2003. Requests for information from a film archive: A case study of multimedia retrieval. *Journal of documentation* 59: 168-86.

Hertzum, Morten, Pejtersen, Annelise Mark, Cleal, Bryan, and Albrechtsen, Hanne. 2002. An analysis of collaboration in three film archives: A case for collaboratories. In Bruce, Harry, Fidel, Raya, Ingwersen, Peter and Vakkari, Pertti eds., *Emerging frameworks and methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science*. Greenwood Village, CO: Libraries Unlimited, pp. 69-83.

Hjørland, Birger. 1992. The concept of "subject" in information science. *Journal of documentation* 48: 172-200.

Hjørland, Birger. 1998. The classification of psychology: A case study in the classification of a knowledge field. *Knowledge organization* 25: 162-201.

Hjørland, Birger. 1997. *Information seeking and subject representation: An activity-theoretical approach to information science*. Westport, CT: Greenwood Press.

Hjørland, Birger. 2002. Domain analysis in Information Science. Eleven approaches - traditional as well as Innovative. *Journal of documentation* 58: 422-62.

Hjørland, Birger. 2004. Domain Analysis in Information Science. In *Encyclopedia of library and information science*. New York, NY: Marcel Dekker.

Hjørland, Birger and Albrechtsen, Hanne. 1995. Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science* 46: 400-25.

Hjørland, Birger. & Pedersen, Karsten Nissen. 2005. A substantive theory of classification for information retrieval. *Journal of documentation* 61: 582-97.

Ingwersen, Peter. 1992. *Information retrieval interaction*. London: Taylor Graham.

Kuhlthau, Carol C. 1993. *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Ablex Pub. Corp.

Lamb, Roberta and Kling, Rob. 2003 Reconceptualizing users as social users in information systems research (1). *MIS quarterly* 27: 197-235.

Lamb, R.oberta, King, John L., and Kling, Rob. 2003. Informational environments: Organizational contexts of online information use. *Journal of the American Society for Information Science and Technology* 54: 97-114.

Lancaster, F.W. 1986. *Vocabulary control for information retrieval*, 2nd edition. Arlington, VA: Information Resources Press.

Mai, J-E. 2001. Semiotics and indexing: An analysis of the subject indexing process. *Journal of documentation* 57: 591-622.

Mai, J-E. 2005. Analysis in indexing: Document and domain centered approaches. *Information processing and management* 41: 599-611.

Marchionini, Gary. 1995. Information seeking in electronic environments. New York, NY: Cambridge University Press.

Nielsen, Marianne Lykke. 2001. A framework for work task based thesaurus design. *Journal of documentation* 57: 774-97.

Pejtersen, Annelise Mark. 1989. *The BookHouse: Modelling users' needs and search strategies as a basis for systems design*. Roskilde, Denmark: Risø National Laboratory (Risø Report M-2794)

Pejersen, Annelise Mark and Albrechtsen, Hanne. 2000. Ecological work based classification schemes. *Dynamism and stability in knowledge organization.* Proceedings of the Sixth International ISKO Conference. Advances in Knowledge Organization, 7. Würzburg, Germany: Ergon Verlag, pp. 97-110.

Pejtersen, Annelise Mark and Fidel, Raya. 1998. A framework for work centered evaluation and design: A case study of IR on the Web. Working paper for the MIRA workshop, Grenoble, March 1998. [Available at www.dcs.gla.ac.uk/mira/workshops/grenoble/fp.pdf]

Pettigrew, Karen, Fidel, Raya, and Bruce, Harry. 2001. Conceptual frameworks in information behavior. *Annual review of information science and technology* 35: 43-78.

Rasmussen, Jens, Pejtersen, Annelise Mark, and Goodstein, L.P. 1994. *Cognitive systems engineering*. New York: Wiley.

Rosenfeld, Louis & Peter Morville. 2007. *Information architecture for the World Wide Web*. 3rd ed. Farnham: O'Reilly.

Saracevic, T. 1980. A research project on classification of questions in information retrieval. *Proceedings of the American Society for Information Science* 17: 146-48.

Saracevic, T. 1983. On a method for studying the structure and nature of requests in information retrieval, *Proceedings of the American Society for Information Science* 20: 22-25.

Simon, Herbert A. 1969. *The sciences of the artificial*. Boston, MA: Massachusetts Institute of Technology.

Soergel, Dagobert. 1974. *Indexing languages and thesauri: Construction and maintenance*. Los Angles: Melville Publ. Co.

Soergel, Dagobert. 1985. *Organizing information: Principles of data base and retrieval systems*. Orlando, FL: Academic Press.

Taylor, Robert S. 1968. Question-negotiation and information seeking in libraries. *College and research libraries* 29: 178-94.

Vicente, Kim J. 1999. *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, Patrick. 1968. *Two kinds of power. An essay on bibliographical control.* Berkeley, CA: University of California Press.

Wilson, T.D. 1999. Models in information behaviour research, *Journal of documentation* 55: 249-70.