H. Gellersen et al. (eds.), ECSCW 2005: Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work, 18-22 September 2005, Paris, France, 367–386.

Context Grabbing: Assigning Metadata in Large Document Collections

Joachim Hinrichs[1], Volkmar Pipek[2] and Volker Wulf[3]
[1]Institute for Information Management Bremen GmbH, Germany; [2]International Institute for Socio-Informatics, Bonn, Germany and University of Oulu, Finland; [3]University of Siegen and Fraunhofer FIT, Sankt Augustin, Germany
jhinrichs@ifib.de, volkmar.pipek@iisi.de, wulf@fb5.uni-siegen.de

Abstract Classification schemes are an important issue in the collective use of large document collections. We have investigated the classification of technical documentations in two engineering domains: a steel mill and a sewerage plant company. In both cases we found a coexistence of different classification schemes and problems resulting from distributed local archives. In supporting human actors to maintain different classifications schemes while working on a common archive, we developed the concept of context grabbing. It allows assigning context information efficiently in the form of metadata. Based on a document management system, a tool kit for context grabbing was developed. Its evaluation in a sewerage service company allows us to comment on important aspects of understanding the role of classifications in collaborative work.

Introduction

Knowledge management has become an important topic for the CSCW community within the last couple of years (Davenport and Prusak 1998; Probst et al. 1999, Ackerman et al. 2003). Since cooperative work is often based on existing documents, document archives and their organisation are an important research issue in the context of sharing knowledge. To maintain a shared document archive proves to be a complex task. Large numbers of documents and additional information need to be categorized, a task involving different actors and stakeholders. This problem is of particularly relevant in the manufacturing

and engineering sector. Maintaining an appropriate structure in vast collections of technical documents is a challenge for practitioners as well as scientists (Carstensen and Wulf 1998, Trigg et al. 1999, Lutters and Ackerman 2002). Accessing specific documents can become a labour-intense and error-prone activity (Hinrichs 2000).

The transition from paper-based archives towards electronic document collections holds the opportunity to capture additional information about a document's context by enriching its representation with meta-data. Context in this sense can be understood as a document's set of present or past relationships in the world. Examples of a document's context dimensions are: objects (e.g., machines, plants) of the 'real world' the document refers to, other documents the document is related to (e.g. same project), human actors who created or accessed the document, or work processes in which the document was relevant (including administrative processes like accounting). A document's context consists typically of an immense variety of different dimensions. When making use of context in digital archives, a small selection of relevant dimensions is typically represented in specific attributes (metadata). Each attribute is defined by a set of values that represent the variation within this dimension of context (capturing one personal, physical, organisational, etc. aspect of a 'situation', see Klemke 2002). The representation of context–based meta-data can be used to constitute classification schemes that support human actors to structure large collections of entities (Simone and Sarini 2001).

The benefit of maintaining context data in digital archives has to be weighed against the effort necessary to capture and maintain the attributes' values for each of the many documents. To deal with this problem, we will propose the concept of 'Context Grabbing' which allows capturing attribute values efficiently. By maintaining a richer representation of context, context grabbing supports human actors to build their specific classification schemes on shared collections of documents.

Additionally, assigning context data is not a straightforward task, Documents and document collections become boundary objects (Star 1989) of different organisational communities, with different sets of 'relevant' dimensions of context that represent and establish the perspective of the respective community. 'Maintenance' of documents, metadata (context) and classifications becomes a matter of multilateral interest, with every actor or stakeholder expecting to find a manifestation of his/her perspective in the archive data available. Changing interests, perspectives and – thus – contexts require adaptable context representations, and 'tailoring' the metadata becomes a crucial task for maintaining document collections. The need for appropriate management support becomes even stronger if large amounts of 'new' documents have to be included in a collection.

In this contribution, we present our idea of providing 'context grabbing' techniques to support classification work in large document collections. These ideas have been informed by earlier research we discuss in the 'State of the Art' section, and by two case studies in industrial settings we we present and comment. After that, we describe one 'context grabbing' prototype we implemented and evaluated. In the concluding sections we discuss the ideas in a broader context of archive management.

State of the Art

In many domains cooperative work is based on collections of stored documents. Current file systems are insufficient for the administration of large amounts of documents. They restrict the users by limited indexing functionality and insufficient support to organize documents in an intuitive way (Dourish 2000). Another problem is the loss of context information when documents are passed on through different departments. Without additional documentation, information about the original context gets lost (Freeman and Gelernter 1996; Rekimoto 1999). Technical functions to record context and at a later point in time to restore previous compilations of the document stocks are missing (Lutters and Ackerman 2002). Even Document Management Systems (DMS) especially designed for the purpose of document administration often prove to be too rigid and are not sufficiently adapted to cooperative work processes (Timmermans 2000). In summary, the technical support for the classification of documents is too inflexible with regard to evolving schemes.

In order to analyse the use and evolution of classification systems, Bowker and Star (2000) gave the static notion of classification systems (as being a segmentation of the world with a set of consistent classificatory principles that operate on a disjunct and complete set of categories) a pragmatic turn. They suggested to accept anything that is "consistently called a classification system and treated as such" under this term. However, the use of the term with regard to the implications for the design of Information Technology (LaMarca et al. 1999) softened the sharp edges of the strict definition even more to allow the inclusion of all activities of classification that are relevant for work.

Various studies show that the order and classification of data are often linked to specific work conditions (Bowker and Star 2000) and that the compilation of the documents reflects the know-how of the actors handling the processes (Hertzum and Pejtersen 2000). While the file structures used are comprehensible and self-explanatory to individual users, the comprehensibility of the classification schemes gets lost at the collaborative level. Severe problems occur when classification schemes for cooperative processes are to be developed (Dourish 2000, Wulf 1997). Different terms and terminologies, but also different modes of operation and understanding complicate the process of coordination

(Bannon and Bødker 1997; Carstensen and Wulf 1998; Trigg et al. 1999). Classification schemes that are introduced in a centralised way and that cover the whole organisation are often too rigid and restrict the users in a disproportionate way (Hinrichs 2000; Pipek et al. 2002). The standards for building classification schemes (IEC 61346 – structuring principles, classification objects and codes) and for structuring technical documentation (IEC 61355 – classification for plants, systems and equipment) still have to be tested in practice. Categories arising by themselves during a more decentralized process often hold better opportunities (Bowker and Star 2000; Dourish 2000; Simone and Sarini 2001) to access relevant information.

With regard to classification schemes for storing and retrieving documents Simone and Sarini (2001) discuss case studies from the CSCW literature. With respect to the degree of centrality, they distinguish between endogenous and exogenous classification schemes. Endogenous classification schemes are defined by a high degree of overlap in common practice between the producers and the consumers of a classification scheme. Exogenous schemes are given in case a "relevant distance" in practice between producers and consumers of classification schemes exists. Simone and Sarini (2001, p. 28) assume that exogenous and endogenous classification schemes coexist and should be both supported by technical means.

When supporting different endogenous and exogenous schemes, capturing information about a document's various contexts seems to be crucial. The Placeless Documents approach offers an infrastructure for highly flexible document administration (Dourish 2000). Applications can be implemented which offer emerging classification schemes by allowing adding new attributes flexibly (LaMarca et al. 1999). While this is a very interesting approach in case new categories for classification come up, the more mundane question remains how to grab the values of these attributes efficiently.

In the Lifestream approach, document administration is supported by temporal information which is automatically recorded. Unlike traditional file structures that are organized in a hierarchical way, time bars represent the chronology of a work process and thus symbolize aspects of the temporal context (Freeman and Gelernter 1996, also in the Time-Machine Computing approach, Rekimoto 1999). Awareness services are often implemented as procedures that record a specific aspect of a documents context automatically (e.g. Fuchs 1998). The display of awareness information may be understood as (short-term) classification. However, automatic procedures are not always suitable to capture those dimensions of a document's context that are relevant for classification.

In more general considerations on 'organisational memories', Ackerman and Halverson (1999) explained that the documents in collaborative contexts themselves represent boundary objects - in the sense of Star (1989) - for the different actors (tasks, organisational entities) that use them, and active processes

of decontextualisation (losing context) and recontextualisation (giving context) mark the crossing of these boundaries and counteract static notions of 'organisational memory'. If context is represented explicitly (as it is when using classification schemes in document collections), there is an immediate need for flexibility in representing the different contexts a document might pass through. As this process is highly dependent on unpredictable organisational changes that every company experiences, the problem to maintain changing context representations becomes a highly important task. The contribution of Ackerman and Halverson also demonstrates the importance of empirical work for understanding the pragmatics of archive maintenance, and for understanding the emergence of classification schemes. It also becomes clear, that there is a need for better technological support for these processes.

Case Studies

We have investigated the practice of document management in two different organisations running complex technical facilities. The first case study deals with the handling of drawings in maintenance engineering of a major German steel mill (Hinrichs 2000, Pipek et al. 2002, Pipek and Wulf 2003). The second case study investigates the document management practice of a company that runs the facilities for wastewater treatment of a major German city.

Running and maintaining complex technical facilities is highly cooperative work. It requires cooperation among different actors typically distributed across various organizational units. Running and maintaining complex technical facilities is highly constrained by the work carried out by other actors in the past. Here, technical drawings play a crucial role in representing states and history of technical facilities.

Organisations that run large-scale technical facilities have to handle vast amounts of drawings and other types of documentation. The two companies investigated employ rather different strategies with regard to the degree of centralization of the document archives. The focus of our analyses was on investigating the role of a document’s context for storage and retrieval.

A Central Archive in a Steel Mill

We have investigated the maintenance engineering processes of a major German steel mill in the Ruhr area. The mill employs about 3,500 employees and is structured into rather independent plant operating units, such as the coke chambers or the blast furnace. Various central units provide services to these plants and manage the mill. The maintenance engineering process involves different central and decentral organizational units as well as external service providers. A central construction department inside the mill coordinates the

planning, construction and documentation of the plants. Important parts of construction work have been outsourced to external engineering offices. In each of the different plants, a small group is responsible for the execution of the maintenance work, often supported by hired external construction companies.

Research Methods

The OrgTech project aimed at improving the maintenance engineering process by introducing groupware technologies over a period of three years (Hinrichs 2000; Pipek et al. 2002, Stevens and Wulf 2002). During the course of the project, the steel mill's central drawing archives turned out to be the crucial bottleneck of plant maintenance. Therefore, we investigated the practice of document storage and retrieval. The results are derived from a variety of different sources:

- Analysis of the work practice: 25 semi-structured interviews, workplace observations, further informal inquiries into special problem areas of work.
- Analysis of the documents, particularly the technical drawings and the descriptions of archiving facilities and processes.
- System evaluation: The existing archiving systems were examined (usability evaluation, with a focus on task adequacy).
- Project workshops: In a number of workshops organisational and technological interventions were discussed to improve the maintenance engineering process.

Empirical Findings

A central organizational unit, the archives group, is responsible for storing the documents that represent the technical state of the steel mill. The central drawing archive represents a history of 100 years. It contains more than 300,000 documents, such as technical drawings, technical descriptions, part lists, static information and calculations. A large part of these documents is filed in conventional paper form and saved on microfilm. In 1995, an electronic archiving system was introduced which contains more than 50,000 drawings, old documents scanned from microfilms or new ones stored in raster format. So far the central archive contains only few CAD files.

The classification scheme of the central archive is based on 'Basic Numbers' that break the mill down into plants and their components. However, this classification has been created for accounting purposes, and was not always meaningful for engineers. The 'Drawing Numbers', the other index, are used rather arbitrarily. The central archive gives sets of Drawing Numbers to internal and external engineers who assign them to drawings. They roughly classify drawings in the temporal order of their creation. These sets of numbers do not reflect the amount of drawings created within individual projects, a project may cover Drawing Numbers from different engineers and different number sets. It is the responsibility of the archives group to classify newly delivered drawings into

the scheme of Basic Numbers, to add certain keywords to the documents and to enter the Drawing Numbers. The consistency of the paper-based archive has suffered from several changes in the classification schemes over the 100 years of the steel mill's history. Within the electronic archive about one quarter of the documents are not appropriately categorised according to the correct Basic Number or stored without keywords. Finally, the electronic archive system does not offer search functions beyond Drawing Numbers, Basic Numbers and keywords.

The central classification scheme and its implementation within the archive system are obviously problematic for maintenance engineering purposes. Information relevant for local work is not considered. To overcome these problems some engineers developed different types of local classification schemes that enabled them to deal with the problems of the central archive. One important context information is provided by project-specific 'Drawing Lists'. Whenever a project is finished, the internal or external engineers create a document that lists all the drawings that have been created or modified during the course of this project. After handing over the drawings to the central archive, the engineers of the internal construction department preserve the Drawing Lists in paper form in their offices. When searching for drawings they cannot find easily in the electronic archive, the engineers refer to the Drawing List to locate drawings from the same project.

While maintaining its own classification schemes, the internal construction department still uses the central archive to store the technical documents. In some plants, local classification schemes lead to the existence of local drawing archives. Annotated copies of drawings are stored by the actors who are responsible for the execution of the maintenance work in the local plants. These local archives can contain up to 500 drawings. Even 'physical' information, such as a drawing's position in a pile or the level of dust covering it, indicates when these drawings have last been used.

The existence of local archives has also implications for the quality of information provided by the central archive. The workers in the maintenance department of the different plants annotate their locally stored drawings when changes in the state of the plant happen without prior construction activities. For instance, plants can be modified without prior planning (and without the creation of any documentary drawing) when accidents happen. This 'sloppyness' also occurs when at the end of a budget year, work is carried out to use up still available funds. Since these annotations are only carried out in the local drawings, the local archives are often more accurate than the central ones.

Local Archives in a Sewerage Work Company

The second field of study was done in a company that runs the sewerage system of a major German city. The allocation of a fixed yearly budget to be invested into

into the extension and maintenance of the sewerage facilities is part of the contract between the city and the service provider. The company has about 400 employees. The technical services of the company are divided into operating and construction departments. There are two operating departments: one deals with the sewer system of about 1000 km length, the other runs two sewerage disposal plants and various pumping facilities. A construction department plans the extension and maintenance of the different facilities. It is divided into two groups: one deals with the sewers themselves, the other with over-ground facilities. External construction companies support both efforts.

Research Methods

The research with the sewerage service company directly focused on problems with handling the technical documentation. In a socio-technical approach, we accompanied the introduction of a document management system (DMS) by means of a socio-technical approach. The results presented in this paper have been collected from a variety of different sources between 2001 and 2003:

- Analysis of the work practice: >30 semi-structured interviews, workplace observations, and further inquiries into special problem areas.
- Analysis of the technical documents and the archiving processes.
- Analysis of the of the organisational appropriation of norms and standards for documentation and classification structures.
- Feedback workshops with the project's 'Steering Committee': Based on the results of the steps above, requirements for the selection of a DMS were specified and discussed with the steering committee of the project that involved stakeholders from all organisational units.
- Introductory workshops: Opportunities to improve the document handling with a DMS were discussed with engineers from all and with members of the steering committee.

Empirical Findings

In the beginning, the sewerage service company did not run a central archive for technical documentation. We found a broad variety of different locations all over the company, where technical documentations were stored. In our analysis, we focused on the construction process and the two operating departments.

The construction department initiated the process of technical documentation of a project, and planning and documentation efforts were intensified after a project's approval by the management. Usually a project was carried out by one engineer, larger projects by small groups of engineers, lead by a manager.

In a project, the engineers in the construction department kept electronic and paper-based folders in parallel. Most of the technical documents, especially CAD drawings, were created on the engineers' computers and stored on a file server. Each engineer had his own folder on a file server that he could structure

according to his individual way of working and classifying. Those folders were only accessible for members of the same group in the construction department.

During the course of the project the engineers started to create a paper-based documentation, as well, resulting in up to 40 DIN A4 folders per project. When the responsibility of a project moved within the construction department or from the construction towards the operating department, only the paper-based version of the technical documentation was handed over. Often, the electronic version of most documents stayed only in the creator's folder on the file server. Electronic versions of drawings considered important were stored on a CD and attached to the physical folders. The operation department usually only got copies of the folders. Those were extended further as the work proceeded. The original documentation either stayed in the engineer's office, or was moved to the local archive of the construction department. 'Projects' are the main dimension for classifying technical documents. Within the project-related folders, the individual engineers were rather free to create the categories for structuring their documentation, and sometimes even individual schemes overlapped significantly.

An important basis for classification was provided by the standardised German 'scale of charges and fees for architects and engineers' (HOAI), which is also part of the professional education. The scale of fees distinguishes nine consecutive phases/activities in construction work (e.g. 'Planning', 'Detailing', etc.). This scheme was also applied for purposes of external subcontracting and internal controlling. So, in some cases the project folders got structured in this way. Other engineers created an internal folder structure based on the time of a document's creation or based on the document type (drawings, drafts, statistical calculations, protocols). One engineer kept specific folders that contained documents and notes that the engineer did not want to share with his colleagues later on in the process.

The engineers were offered some freedom to implement their project-specific classification schemes, although the relevant standards for documentation (DIN 6779 resp. IEC 61346) were well known in the organisation. The pattern of decentralization led to a couple of severe problems. Documents were redundantly kept in different locations, which left it unclear whether a document version represented still the actual state. The documentation in a local archive became incomplete in the course of time, since folders were taken away when needed and not returned. Archiving and working processes were also suffering from media discontinuities, since there was no direct linkage between the electronic documents and their paper versions. Lacking access to the appropriate documentation led to severe problems. Incomplete or inaccessible documentation e.g. lead to costly exploratory 'digging by hand' to avoid damaging power lines.

Supporting Classification Work

The two case studies indicate that a broad variety of context dimensions were selected by the different actors to create classification schemes for technical

documentation. In both of the case studies, the historic context of a document's creation played a major role. In the case of the steel mill, the Drawing Lists were an important resource for finding those documents that were created in the same project. In case of the sewerage work company the historical context of creation was the main classification scheme for all technical documents. A second dimension in classification was provided by the structure of the facilities the drawings referred to. This was the main classification dimension in the steel mill. However, there were different versions of this scheme. The central archive was based on an economical interpretation that divided the plant up into cost centres while the plant operators' local archives were rather structured according to a technical interpretation of the plant's structure. In the sewerage service company, reference to the facilities was not used as a classification scheme, since the facilities did not have the complexity to make this necessary. Instead, the geographical position of the facility the drawing referred to was documented in each drawing as part of a descriptive set of information. Another dimension mapped historical aspects. The phase of a document's production in the engineering process was part of a documents' context in the case of the sewerage work company. An important dimension of classification with regard to local archives in the Steel Mill was the reference to the actor in charge. In both companies, local archives were kept in the actor's offices. When looking for certain documentation, one usually asked those engineers to provide help.

Interestingly, the different classification schemes do not always create fully distinct subdivision of the documents. For instance, geographical and technical interpretations of the structure of the plant do overlap in a considerable manner. A project-based classification overlaps greatly with one that is based on the 'engineer in charge'. Obviously there exist similarities between different context dimensions that could be exploited to maintain classification schemes efficiently.

Coexisting central and local classification schemes resulted from different tasks and work practices in the organisational subunits. The coexistence of different classification schemes led to the problem of a redundant storage of technical documents, which again led to inconsistent document bases. The transition from paper-based archives to electronic archives often results in the loss of a dominant (physical) order, but it also offers the opportunity to operate with several different classification schemes that can be extended with new attributes when needed. In decentralised architectures, synchronisation mechanisms can help maintaining a consistent database.

So far research on technical support for classification work has mainly focused on flexibility. Architectures should allow flexibly adding or modifying the represented dimensions of context (e.g. Trigg et al. 1999; Dourish 2000, Sarini and Simone 2001). However, it is not only a question of being able to define attributes flexibly. The more attributes of a document's context are modelled and the more dynamic they change, the more classification work results (cf. Trigg et

al. 1999). To make this classification work more efficient, we have developed the concept of context grabbing.

Context Grabbing

Under the label of ‘Context grabbing’ we collect a set of techniques to support categorisation work in large document collections. The goal is to provide a time-efficient way to maintain context metadata of documents. These techniques can complement DMS, but also file sharing applications. They need to be customisable with regard to the existing local work practices.

We distinguish two possibilities to capture context information: automatically or computer supported. Since documents are created, manipulated and stored on computers many aspects of a document’s context can be grabbed automatically. For instance the time of a document’s creation or last modifications can be extracted automatically. Additionally, information about the set of other documents a document was ever stored with in a folder can be grabbed automatically (e.g., to produce the ‘Drawing Lists’ in the Steel Mill).

Capturing context automatically does not work if the relationships of a document (with actors, documents, tasks, etc.) are not represented in the computer. For instance, it is difficult to decide automatically which part of a plant a drawing refers to. This information has to be provided by those human actors who possess the relevant knowledge. Computer support should make their classification work more efficient.

Computer support in grabbing context information can be based on similarities either between the value sets of different documents or between value sets of different context attributes. Exploiting these similarities allows both, assigning attribute values in an automatic or computer-supported manner. For instance, in the sewerage work company we found that a project-based classification strongly resembles the one based on the ‘engineer in charge’. So, in case the attribute ‘engineer in charge’ is newly created in a digital archive, its value can be assigned to individual documents by referring to the values of the attribute ‘project number’. Since the value sets of different attributes show similarities but are typically not identical, fully automated, e.g. rule-based, approaches to the problem are not feasible. The human actor needs to stay in control.

We can distinguish two cases of context grabbing. In a first case, values of a newly created attribute have to be assigned or the values of an existing attribute have to be updated. In this case different values of the same attribute have to be assigned to many documents. Secondly, there are cases in which a newly created document has to be classified with respect to all relevant context attributes. These cases require different kinds of tool support.

In the first case one can exploit the similarity between the value sets of different attributes. The tools for assigning a value to a particular context attribute

for a set of documents need mechanisms to specify the scope of validity of an assignment operation, maybe by exploiting the existing folder structure and the value sets of those attributes that are already defined. Since these specifications can become quite complex, users have to be supported in understanding them. In the second case, the user needs support to identify documents or sets of documents that already have been classified. These documents can then be taken as points of reference to copy the values of all of their context attributes. Here again the user should stay in control to check whether all the appropriate values get assigned.

The strategy to exploit similarities of classification dimensions for assigning meta-data not only requires appropriate editing functions. Since the similarities themselves are often hard to detect, additional support for detecting and visualising these similarities is also helpful. Relations between different documents that are represented by means of context attributes can be used by specific search tools to provide graphical representations. For instance, in case of the steel mill it would be very helpful for the engineers if all those documents could be displayed together that once had been stored in the same project folder.

A Tool Kit to Support Context Grabbing

We now describe our approach to support context grabbing to one of our fields of study. In the course of the project, the sewerage work company decided to introduce a document management system (DMS). Based on requirements developed in the initial phase of the project, windream©[1], a commercial DMS product, was chosen. Contrary to traditional DMS that run as separate document management applications, windream's document management functionality is integrated into the file management of the operating system. It adds functionality of a DMS such as version control, document life cycle management, differentiated access control, and a sophisticated search tool.

As an important prerequisite for our approach, windream supports the evolution of classification schemes by allowing to structure the meta-data as a basis for classification and to dynamically add new attributes to existing schemes. Beyond the typical features of operating systems, windream offers additional functions to grab values of a context attribute automatically (e.g. regarding data of a document's history). However, there is no appropriate support to assign values of context attributes manually to larger collections of documents. Administrators of windream can create 'index sheets', specific pop-up windows to enter values of a document's different attributes. Depending on the type of attribute, specific value sets and interface elements can be defined, as well. An example of such an 'index sheet' is presented in Figure 1.

---

[1] http://www.windream.com/

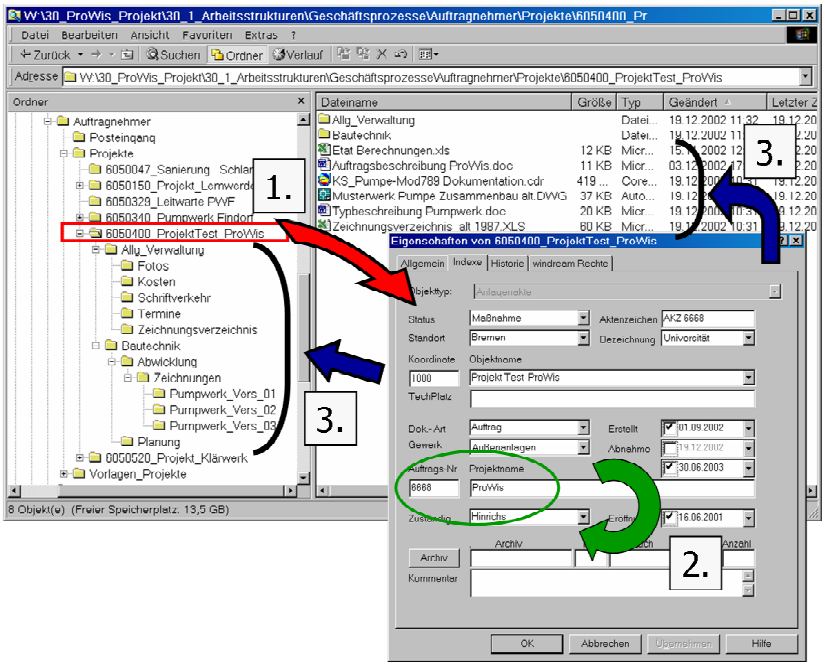

Figure 1. Context grabbing supported by the Windexer: (1) after selecting a folder, the index sheet appears, (2) values of specific attributes can be modified, and (3) assigned to selected documents

To make the manual assignment process in the sewerage work company more efficient, we implemented a tool kit based on the DMS. The tool kit consists of two applications, called Windexer and PreWindexer. The Windexer allows assigning a specific attribute value to a set of predefined documents. The PreWindexer supports the classification of newly created documents by assigning a whole set of predefined attribute values. The classification process via the Windexer operates on the basis of folders. When the Windexer is activated the folder's index sheet appears, (first step in Figure 1). The user can enter and modify attribute values (second step in Figure 1) that serve as the basis for assignment operations. For the assignment operation, the user can select which of the different attributes in the sheet should be assigned to what group of documents (by criteria like name of the creator, the date of creation, or the type of document; third step in Figure 1). Finally, a description of the operation in plain text is presented for user confirmation, and a list of altered documents is produced. For each assignment operation, the Windexer creates a unique identification number that is automatically assigned to all altered documents of the operation. This code number allows recreating the grabbing context by searching for the documents that have been "windexed" together.

The PreWindexer is a tool that helps assigning attribute values to newly created documents. The assignment of the attributes' values is again based on the

index sheets assigned to the folder structure of the DMS, which provide the meta-data that then is assigned to every document placed in that folder. Usually this happens automatically, but optionally the user can modify them for every operation.

The tool kit also contains a search tool (ContextSearch). It can be activated by selecting a folder or a document. After activating the search function in the context menu, a window to specify the inquiry pops up. The structure of the search tool window is similar to the one of the index sheet. To simplify entering the query, the attributes of the search window are initially filled with the value of the selected folder's or document's index sheet. These search values may be altered, but may provide an easy starting point for complex queries. The retrieved documents are displayed as a list of hits that can be saved and used as a reference for further search processes. The tool kit was implemented using Microsoft©'s DCOM-technology (Distributed Component Object Model) and the API of the windream software.

Context Grabbing in Practice

The DMS was introduced to the sewerage work company to overcome the problems caused by the coexistence of the various local archives. The tool kit for context grabbing played an important role in enabling the transition from the local archives towards a better integrated pattern of storage.

Introducing the DMS

The introduction of a DMS was the technological part of the management's agenda to improve the overall performance of the formerly state run company. On the organisational side, the construction department was split up and integrated into the two operating departments. The change in the formal organisation had an impact on the way the DMS was applied to centralise document management.

A pilot installation of windream was run for half a year on data from one completed project to experiment with the functionality, then a field trial was conducted with a small group of engineers. During that time the system was also presented to various actors from the two operating departments. During these presentations, requirements for the context specification using the index sheet were collected. Based on prototypical implementations of the index sheet, these requirements were discussed in the project's steering board. The integration of the local archives was prepared, and a centralised concept for document management was developed. A classification structure for the file repository was built. The folder structure resulting from prior archiving strategies built the basis that was complemented by the metadata of the index sheet that provided classifications according to work practice and technical standards for documentation (e.g. IEC

61346). The need for the suggested functions of Context grabbing became even more manifest with this experience. The training of about 60 actors during the introduction addressed DMS as well as toolkit functionality, and followed the new conventions on document management.

Classifying Documents

Our evaluation of the context grabbing tool kit covered about 50 workplaces that were observed for the period of about one year. Most of the experiences we were able to record came from field notes from informal communications during site visits and from the conversation in the steering committee. Additionally, 10 semi-structured interviews were conducted regarding the use of the DMS and our tools. The introduction of the DMS lead to far reaching changes in the handling of electronic documents: vast collections of individually structured documents suddenly got shared among different actors. To enable this transition the individual as well as the newly established organisation-wide classification schemes had to be entered into the system.

The generation of classification schemes in the DMS is restricted by its original functionality and its local configuration. Classification schemes relied on both, the folder structure (as the basis) and the index sheet of the DMS (as additional classification scheme). Each of the two departments worked in one folder. On the next structural level, three project phases were distinguished by corresponding folders: “planning”, “detailing”, and “operating” (based on HOAI), that again contained project folders with all documents belonging to that project. Folder movements followed the proceeding of a project. There was no general template about organising the project folders, but the engineers were asked to keep a flat structure. Project folders usually were created by project managers and then passed to the engineer carrying out the technical work.

While the structure of the individual project folder was still rather specific to the individuals in charge, the attributes represented in the index sheet allowed for additional classification schemes. Some of these attributes have an organization-wide meaning (e.g., the seven-digit project number also used in the ERP system). The value sets of other attributes are less well defined (e.g., the project name is an arbitrary character string chosen by the project manager). Interestingly, there is a considerable redundancy among certain attributes. The project number and the project name always characterise the same project, but both attributes were included in the index sheets since different actors are better able to interpret attribute values of the one or the other type. Some attributes of the index sheet represent super-/subclass relations. The reference number (“Aktenzeichen”) was a superclass of name, location, coordinate and object name, ‘object name’ was the superclass of ‘technical location’, craft, project and order number. These super-/subclasses served as a flexible classificatory orientation for users.

The values of attributes are assigned and modified by various actors at different points in time. When launching a project the project manager uses the PreWindexer to configure the project folder. Values of the initially known attributes are suggested whenever a new document is stored in the folder. Typically also attributes such as project name and number, cost center, facility, engineer in charge, and status are assigned at that point. During planning, the engineer adds attribute values such as geographical coordinates, object name, and object location. When the project status shifts from "planning" to "detailing" additional attributes may have to be assigned or modified (e.g. time of completion, engineer in charge). The engineers in charge can also add comments in plain text (e.g. information about a customer). Depending on the internal folder structure and the time of assignment, for these activities the Windexer or the PreWindexer were used. When the construction work was finished, the project documentation was archived electronically. At the same time, copies of certain documents were created and passed to various actors (plant operators, external construction firms). In these copies, the classification provided by the folder structures is not present anymore. Thus, the folder-based classification schemes were fully duplicated by means of attributes of the index sheet.

The Windexer proved also helpful for the classification of documents of about 120 to 150 running and approximately 300 completed projects. After being transferred from the file server to the DMS, these documents had to be also classified. One problem in the course of the introduction was the workers' refusal to accept a delay of about 30 minutes until the context assignment was effective in the DMS. The delay was caused by a problem with the file locking mechanism of the office software used. The tools were only used to their full capacity when an immediate storage (and presentation) of context in the DMS was guaranteed.

Reconsidering Classification Work

When observing and supporting work that relates to classification schemes, it is important to understand the way how classifications are objectified, used and altered (Simone and Sarini 2001). Our studies as well as the evaluation of the context grabbing tool kit suggest that it is important to embrace deviations in the use of classification systems instead of fighting them with standardisation efforts.

Star's (1989) notion of boundary objects helps us to further argue in that direction. Documents are not simply 'work results', they also became the anchor of different perspectives on work goals and work processes. In the times of paper-based documentation, their location, attached markers and comments, and other 'physical' attributes often documented the state of work processes as well as the meaning of current work tasks. That way, work practices have made 'documents' meaningful beyond 'documentation'. They became boundary objects of different communities that collaborate in an organization to get work done. The 'context'

every actor or group of actors subjectively associates with a document is a manifestation of the meaning the document has for their work, and it is as important as the documents' content[VW1].

Context has to be re-established every time the document is used, and 're-contextualisation' is an important activity in using organizational memories (Ackerman and Halverson 1999). Organisation-wide classification schemes are one way of maintaining (part of) a documents' context, but our experiences show the importance of local practices of context maintenance (e.g., copies, annotations), sometimes even their priority (e.g., higher accuracy of local archives in the steel mill). It is important to consider how we deal with these dynamics when designing the transition from physical to electronic archives.

The dangers for 'traditional' approaches that actors choose to maintain their contexts are manifold: Copying, arranging, annotating, modifying and sorting documents work differently with electronic archives. The seducing power to impose (finally!) a single classification scheme on all documents often tempts managers on all organizational levels. When classification schemes are centrally developed and imposed, power relations play an important, often dysfunctional role since they hinder the maturing of schemes (cf. Star and Bowker 2000). Before the implementation of the DMS, the engineers of the construction department of the sewerage work company were able to predefine the structure of the project documentation because they were the first to built up a local archive that was later copied. In a number of cases, their schemes influenced the way the succeeding actors in the operating departments went on in organizing a project's documentation. With a centralised approach this diffusion of schemes is not possible anymore. We see that on an individual as well as on a collaborative level the transition to electronic repositories holds challenges for context maintenance. But the danger does not always come from 'above': In a case study on the development of classification schemes in a German public administration, we saw that typists who had more experiences in classifying were able to impose their scheme for some time on their clients (cf. Wulf 1997).

To strengthen the argument, the effects and value of emerging classification schemes have to be the focus of additional practice-oriented research. From an action research perspective, we also need to better understand how to facilitate the negotiation processes that are necessary when local classification schemes merge.

Technological Support for Classification Work

Technologically, it is not enough to provide flexibility in classification schemes, e.g. by allowing the definition of new context attributes and value sets. The flexibility has to be complemented by appropriate tools to manage it even for large document collections. Automated approaches can only operate on the traces of context that are machine-readable (timestamps, etc.). To fully integrate appropriate context maintenance in document management systems, human actors

have to be supported in modelling their context descriptions and maintain their individual perspectives, as it was the goal of our concept of context grabbing. This requirement can obviously lead to large numbers of non-disjunctive, even redundant context dimensions. Our experiences indicate that this is by no means a problem. When context visualisation is appropriately integrated into the user interface and supported by search tools, disadvantages due to a lack of transparency can be avoided. However, redundancy among attributes can have positive effects. The differences in the naming of redundant attributes can support local interpretation and sense making processes. Again, the challenge is not fighting congruency and redundancy, but dealing with it. Our concept successfully exploited congruencies between context dimensions for assigning context metadata to documents. Our case studies even indicate that the acceptance of a central electronic archive can be greatly increased when tools for managing local or individual context dimensions are provided.

We regard it as most important to further exploit those similarities, e.g. in asking how value sets of attributes produce subdivisions of document sets. Providing an editing tool that allows using these similarities in assigning context descriptions is just a first step. In our case studies, the congruency of attributes was easily recognisable for users familiar with the organisational aspects of the documents. But there may be similarities between attributes that are harder to detect. Here, automated support for detecting these congruencies is possible and would further improve the usefulness of the concepts presented here.

Classification Cultures

Simone and Sarini (2001) already focused on the importance of classification schemes for intra- and intergroup collaboration. One of the dimensions they described as important is the ‘distance’ between definition and use of classification schemes. They distinguish exogenous (external to common practice) and endogenous (derived from common practice) classification schemes to capture this distance. In the sewerage service company, the HOAI and documentation standards (IEC 61346) supported inter-group cooperation in classification work. Those were exogenous classification schemes, but very much ‘in practice’. Similarly, the education of engineers in the steel mill provided a valuable background for classifications according to technological properties of the facilities. In our eyes, the dimension of ‘distance between definition and use’ in fact refers to a cultural distance between those defining a classification and those using it. The argument that frequent collaboration produces a shared culture of understanding which then again allows ‘endogenous’ classification schemes to occur just describes an effect of cultural dynamics at workplaces. A ‘cultural’ understanding of this ‘distance’ is not only a redefinition of terms, but it also suggests different research efforts to further deepen the understanding of the relation between collaborative work and classification schemes. In the light of this

argumentation, a future analysis of the long-term effects of the context grabbing concept and tool kit is likely to suggest not only improvements for technological support, but also new theories on the emergence of classification schemes.

Conclusion

Especially when it comes to knowledge-intensive environments, classification work in order to allow a later retrieval of valuable information, is an important part of knowledge work. We were able to describe the experience from two field studies in industrial settings. Classification in practice happens on various individual and organisational levels, along different local and emerging classification schemes. Document Management Systems (DMS) aim to organize large document collections, but they usually treat documents as once and forever classified according to an acknowledged classification scheme. To allow a more flexible use of classification schemes in practice we suggested 'Context Grabbing' techniques to build and maintain classifications according to the context metadata of documents. A prototype for (semi-)automatically assigning context metadata attributes to large groups of documents has been evaluated in one of the fields. The results of the evaluation stressed the need to support the emergence of classifications, and to support the maintenance of large document collections also in order to maintain them as boundary objects of collaborating organisational communities.

As reliable and unambiguous as classification schemes have to be to be operable, there is no point in pretending a timeless validity in collaborative contexts. Praxis reinterprets and changes the schemes frequently. Classification schemes can be understood as coordination languages for search and retrieval of information. Approaches to support 'classification work' should take into account what makes language useful[VW2]: Enough stability to guarantee mutual understanding, and enough ambiguity to allow for emerging changes.

References

Ackerman, M.S., Halverson, C. (1999): Organizational Memory: Processes, Boundary Objects, and Trajectories. In: IEEE Hawaii International Conference of System Sciences (HICSS'99).

Ackerman, M.; Pipek. V.; Wulf. V. (2003) (eds): Beyond Knowledge Management: Sharing Expertise; MIT-Press, Cambridge.

Bannon, I.; Bødker, S. (1997): Constructing common information spaces; In: Hughes, J.; Rodden, T.; Prinz, W.; Schmidt, K. (eds): Proceedings of ECSCW 97, Kluwer, Dordrecht, pp. 81-96

Bowker, G. C.; Star, S.L. (2000): Sorting things out: Classification and its consequences, MIT Press, Cambridge, 2000

Carstensen, P.; Wulf, V. (1998): Common Information Spaces in Engineering Design: An Analysis of the Structure and Use of a Project File; In: Proceedings of Concurrent Engineering (CE 98), Tokio, 1998, pp. 127–135

Davenport, T.-H.; Prusak, L. (1998): *Working Knowledge: How Organizations Manage What They Know.* Harvard Business School Press, Boston, MA, USA

Dourish, P. (2000): Technical and social features of categorization schemes; In: Schmidt, K.; Simone, C.; Star, S.L.: Workshop Classification Schemes; CSCW 2000; Philadelphia, 2000; available at: http://www.isr.uci.edu/~jpd/publications.shtml

Freeman, E.; Gelernter, D. (1996): Livestreams: A storage model for personal data; ACM SIGMOD Bulletin, 1996

Fuchs, L. (1999): AREA: A Cross Application Notification Service for Groupware, in: S. Bødker, M. Kyng & K. Schmidt (eds): Proceedings of ECSCW '99, Kluwer, Dordrecht , pp. 61 - 80

Hertzum, M.; Pejtersen, A. (2000): The information-seeking practices of engineers: Searching for documents as well as for people; Information Proc. and Management 36; 2000, pp. 761-778

Hinrichs, J. (2000): Telecooperation in Engineering Offices - The problem of archiving; In: Dieng, R.; Giboin, A.; De Michelis, G.; Karsenty, L.: Designing Cooperative Systems; COOP 2000, IOS-Press, Sophia Antipolis (F), 2000, pp. 259-275

Klemke, R. (2002): Modelling Context in Information Brokering Processes; Dissertation with the Rhenanian-Westfalian Technical University of Aachen, http://sylvester.bth.rwth-aachen.de/dissertationen/2002/120/02_120.pdf, 2002

La Marca, A.; Edwards, W. K.; Dourish, P.; Lamping, J.; Smith, I.; Thornton, J. (1999): Taking the Work out of Workflow: Mechanisms for Document-Centred Collaboration, in: S. Bødker, M. Kyng & K. Schmidt (eds): Proc. of ECSCW 99, Kluwer, Dordrecht, pp. 1-20

Lutters, W.; Ackerman, M. (2002): Achieving Safety: A Field Study of Boundary Objects in Aircraft Technical Support; Proc. CSCW 2002, ACM, New Orleans, 2002, pp. 266 - 275

Pipek, V.; Hinrichs, J.; Wulf, V. (2002): Sharing Expertise: Challenges for Technical Support; In: Ackerman, M.; Pipek. V.; Wulf. V. (eds): Beyond Knowledge Management: Sharing Expertise; MIT-Press, Cambridge, 2003, pp. 111 - 136

Pipek, V., Wulf, V. (2003): Pruning the Answer Garden: Knowledge Sharing in Maintenance Engineering. in European Conference on CSCW, (Helsinki, Finland, 2003), Kluwer, 1-20.

Probst, G.; Raub, S.; Romhardt, K. (1999): Wissen Managen: wie Unternehmen ihre wertvollste Ressource optimal nutzen; 3. ed., Gabler, Wiesbaden, 1999

Rekimoto, J. (1999): Time-Machine Computing: A Time-centric Approach for the Information Environment, in Proceedings of UIST 99, ACM-Press, New York, 1999, pp. 45-54

Simone, C.; Sarini, M. (2001): Adaptability of Classification Schemes in Cooperation: What does it mean? In: Prinz, W.; Jarke, M.; Rogers, Y.; Schmidt, K.; Wulf, V. (eds): Proceedings of ECSCW 2001, Kluwer, Dordrecht, 2001, pp. 19-38

Star, S. L.: The Structure of Ill-Structured Solutions, in: Glasser, L.; Huhns, M. (eds): Distributed Artificial Intelligence – Volume II, Morgan Kaufmann, 1989, pp. 37-54

Timmermans, H. (2000): Was wird von Dokumenten-Management-Systemen zukünftig erwartet? In: EDM-Report, Nr. 1, Dressler Verlag, Heidelberg, 2000, pp. 64-71

Trigg, R. H.; Blomberg, J.; Suchman, L. (1999): Moving document collections online: The evolution of a shared repository. In: S. Bødker, M. Kyng & K. Schmidt (eds), Proceedings of ECSCW 99, Kluwer, Dordrecht, 1999, pp. 331-350

Wulf, V. (1997): Storing and retrieving documents in a shared workspace: experiences from the political administration; In: Proc. INTERACT 97; Chapman & Hall, UK, 1997, pp. 469-476