Establishing Context of Digital Objects' Creation, Content and Usage

Rudolf Mayer
mayer@ifs.tuwien.ac.at

Andreas Rauber
rauber@ifs.tuwien.ac.at

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Vienna, Austria

ABSTRACT

The context of objects is essential for the interpretation of information entities, for establishing their authenticity as well as ensuring appropriate use. Thus, documenting the context of creation and use is an essential task in digital library and document management settings, for retrieval tasks as well as for digital preservation. Yet, context is notoriously difficult and labour-some to establish and document, and often missing or partially incomplete or incorrect when it has to be entered manually by the creator of the digital objects. This paper introduces an approach to (semi-)automatically determine the creation and usage context of digital objects. Various aspects of context in different dimensions are automatically detected, and different views at multiple levels of granularity allow the extraction of the most appropriate connections to other digital objects.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Context Analysis and Indexing, Information Search and Retrieval

General Terms

Metadata creation, Object context, Digital Libraries

1. INTRODUCTION

Digital objects – as all information – do not exist isolated, but are embedded in and also form themselves a larger context, an information space. This context of information objects may be the setting and intention within which they were created, the persons and activities involved, as well as the time frame and potentially correlating activities that are somehow linked to or influence a digital object. Context thus consists of the relationship that an information object shares with other information items (such as creation time, type, purpose, creators, users, and others), but also of the embedding of the the very content itself, that a piece of information is conveying, such as style, genre, facts, references to other documents, etc. Here, a piece of information may be both a digital object as such, but also a subset of it, e.g. a certain information item in it, or a content snippet, or on the other hand, also a larger group of objects which are already bound together by a common context, such as e.g. an email body and its attachments. Thus, context of digital information basically describes all relationships and commonalities of that piece of information with other information items, and dimensions along which these can be structured. All these types of context are essential for its interpretation, and form a core aspect of establishing its authenticity, constituting the record characteristics, an essential subset of the significant properties of objects in digital preservation settings. They are also an important feature for identifying relevant pieces of information for a given information need, and as such form a basis for retrieval tasks.

Even so, establishing and documenting the context of information is a notoriously difficult and time-consuming endeavour. Professional institutions such as cultural heritage institutions or archives in industry settings, press agencies, etc. take great care and invest substantial effort in correctly and extensively documenting the various types of context of individual pieces of information. Due to the tremendous effort required to establish and document this context, a range of approaches has been developed, aiming at automatically capturing it. This may range from documenting essential metadata, e.g. author and time information as part of the creation process of a document (as found, for instance, in standard office applications), to more complex workflow environments, where groups of related objects are bundled, cross-linked and the various stages in their creation and usage are documented automatically by a document management system.

However, in less controlled or integrated environments, the documentation of the context of information is in most cases utterly neglected. This applies to many small and medium enterprises, as well as small and home office (SOHO) environments. On the one hand, this constitutes a problem for the very people creating and using the content, who are finding it increasingly difficult to locate certain pieces of information, such as a specific photograph, a letter, or a certain e-mail discussion. On the other hand, it poses a significant challenge in professional settings such as archives and libraries, when they are receiving donations or bequests that are to be ingested into their collections. So far, this process relies on predominantly manual work, supported partially by collection profiling tools that analyse the various file types, and assisting in browsing directory structures, e-mail folders

InDP'09, June 19, 2009, Austin, TX, USA.

– or handling piles of CDs or other optical storage media.

Correctly establishing the context of a digital object may never be a fully automated process. Still, a range of techniques may be employed to lower its level of complexity. It can, e.g., be of great help to identify potential correlations between objects, events, persons, or activities across time to semi-automatically establish a basic context of pieces of information in a bottom-up manner. Context between digital objects can be present and established on several different and orthogonal dimensions, for example objects created in the same *time*, within the same *project*, with similar *content*, with a certain set a of *people*, and of a certain *type*. In this paper, we thus propose methods to automatically and semi-automatically, in an interactive process, recover and detect inter-object context on all these dimensions.

We build on our existing research prototype [16] which facilitates analysing information objects in different dimensions, relying on techniques from information retrieval, natural language processing, machine learning, and concepts found in on-line analytical processing. We give application scenarios of how this explicit context information can be used to enhance the metadata descriptions of digital objects ingested into an archive or serve as essential finding aids for personal information repositories.

The remainder of this paper is structured as follows. Section 2 gives an overview on related work. Section 3 presents possible object sources, while Section 4 then introduces the different types of context we consider for our (semi-) automatic extraction and detection process. Section 5 outlines some possible application scenarios for utilising the newly generated context, while we give a conclusion and outline directions for future work in Section 6.

2. RELATED WORK

The preservation and management of digital objects has become one of the major research objectives in the field of Digital Libraries. The transition from a view of mere documents to digital objects includes additional aspects such as relations between the objects as well as authenticity considerations. Along with metadata providing additional information, the relations between the objects and their context, such as the setting in or intention for which they were created, are important cues. As an example, considering a news article, knowledge about its setting, time, and persons involved in its creation, means of distribution, and discussion on it in different media, may be vital for a better or even correct understanding and fair assessment of its authenticity and reliability. Articles published on ephemeral blogs or private Web pages have a whole different impact than official, proofread and corrected, news bulletins. While such context assessment can be well accomplished by humans, integrating it as an automated processes into digital library systems has proven to be difficult and identified as a major task for the future [19]. Taking into account these properties of digital objects thus has become a major research challenge in the digital library community.

The importance of context for digital objects is, for example, detailed in [24], which focuses on the specificities of *digital literature*. The authors point out that, additionally to the technical environment a digital object was created in, such as the hard- and software used, also the sociological environment should be preserved. This is particularly important for volatile pieces of information, such as so-called networked writing and pending writing, which involve several authors working collaboratively, along with reader comments and interaction, both being actively involved during the creation step. Similar issues can be found in collaborative tools such as Wikis.

The task of mediating between available information objects in Digital Libraries, and certain information needs, and how the context of these information objects can facilitate this task, is discussed in [18]. A taxonomic description of context for this purpose, which organises context in *Information context*, *Community context* and *User context*, is presented. While we capture some aspects of *Community context*, mostly by the analysing who used and modified a certain object, our primary focus is on *Information context*, which 'includes information about the digital information object that is directly related to the individual artefact, and to its surrounding information structures like an information collection it is part of' [18].

A system organising digital photographs by their time and location context is presented in [17]. The organisation is along several hierarchies, and mimics the way people think about their collection. An investigation into organisation of digital photographs by hobby photographers and professionals is presented in [8], where the authors make the observation that *meta information* is an important factor when retrieving specific images from a larger collection. Specifically, associated information on events, locations and activities are considered important contextual aspects.

A study on contextual data of videos on Youtube, e.g. discussions on those videos in blogs, and the importance of this contextual data for the understanding of the relevance of a certain video is presented in [4].

A range of tools and services that perform content characterisation specifically for digital preservation has been developed. The National Library of New Zealand Metadata Extraction Tool[1] extracts preservation metadata for various input file formats, while Harvard University Library's tool JHove[2] enables the identification and characterisation of digital objects. Collection profiling services build upon characterisation tools and registries, such as DROID and PRONOM [3], to create profiles of repository collections.

Besides the predominantly metadata-oriented information extracted by the tools discussed above, we further employ text mining techniques to extract content-specific information. Specifically, natural language processing tools such as GATE [3] [5] allow the detection of concepts such as names, places, dates, or part-of-speech structures such as pronouns, adjectives, etc.

Many aspects of social and community data prove useful for digital library tasks. User annotations, e.g., can be a vital aspect of a museums on-line catalogue. A hybrid approach for merging authoritative metadata catalogues based on library standards with community annotations and tags is proposed in [9]. A test study was performed to show the applicability to image data. Yet, in many settings, the representation of these relations between various objects are very fragile and may not even exist in a single place, as for example with hyperlinks and embedded objects.

[1] `http://meta-extractor.sourceforge.net/`
[2] `http://hul.harvard.edu/jhove`
[3] `http://gate.ac.uk/`

After extracting a multitude of characteristics from digital objects, these need to be analysed in order to identify patterns, first individually within each of them, then by combining them to identify correlations. To this end, machine learning techniques are employed to analyse either the cluster structure or to assign category labels to objects. Specifically, Wards clustering [11] as a representative of the agglomerative clustering algorithms is used to identify patterns, and provide hierarchical organisation. Other approaches that have been used to detect and visualise structures in large document collections include the self-organizing map (SOM) [15, 22].

While the detection of topical aspects has received a primal focus in the domain, functional characteristics of documents are as well important for their correct interpretation. This can be addressed by performing genre analysis of the digital objects, extracting textual, structural and layout features [2] as well as specific keywords, that can be used to train a classifier sorting documents into specific genre categories such as minutes, memos, papers, websites, notes, listings, etc. [13, 12]. Yet another line of research addresses the type of use or sensitivity of objects, such as discriminating between potentially private/conversational or public/official documents [21].

The principle of analysing digital objects introduced in this paper is inspired by data warehouses, which store vast amounts of (typically numerical) business transaction data and dimensional data, which is used as reference information to bring context to the business data. Analysing data in a data-warehouse is performed using the online analytical processing (OLAP) approach, with the OLAP cube preparing the data for fast multi-dimensional analysis as central concept. A detailed explanation of these concepts is given e.g. in [14]. Data warehouse and OLAP principles have also been applied to less-structured digital objects, creating so-called information warehouses. Examples for these are warehouse for analysing web pages [1, 20].

When recovering objects from (e-mail) archives, means for structuring the resulting object collection and navigating it are essential. A range of different visualisations in the context of desktop search are evaluated in [7]. Its user evaluation showed that Tree View and Cloud View are the most useful visualisations for presenting search results. The Tree View takes into account user defined folder structures, which further motivates the research presented in this paper.

Social network techniques have been studied in a variety of contexts ranging from specifically designed on-line communities to automatically constructed networks from e-mail data. Various publications have been based on the Enron corpus[4]. This corpus was also studied in [6] showing the applicability of social network analysis to real-world scenarios as well as problems therewith.

3. OBJECT SOURCES

In this section, we give a short and illustrative, but certainly not complete, overview on potential sources of digital objects, their characteristics and what kind of information can be extracted from them.

3.1 E-mail mailboxes

E-mail repositories in general contain a rich set of metadata. Single e-mails contain information about the sending date, sender, and recipients, and other information such as the program used to compose the e-mail. Attachments in e-mails normally come along with an object type, and are embedded in the context of the e-mail they are attached to, and potential other attachments from the same e-mail. Further, e-mails might be filed in (hierarchical) folder structure, which indicates relationships between e-mails.

3.2 Wikis, Versioning systems, BSCW...

Wikis and similar online system generally contain a rich set of authoring and versioning information, i.e. for each version, the date, person and changes to a previous version are stored. Files attached to a page may share similar characteristics as e-mail attachments.

Versioning systems also have rich data about creation, modification and authoring history of their holdings. While they are often used for storing source code of software, such systems may also be employed by writers and thus contain literature.

Other collaborative systems, such as BSCW [5] may additionally contain information about users accessing the system, and reading/opening specific documents.

Further, this type of repositories may comprehend less collaborative systems, such as a user's blog.

3.3 File-systems

Object stored in a file-system might provide less details as the above mentioned repositories, but still a lot of information can be obtained. In most cases, either the file or the folder containing the file will bear a meaningful name, containing keywords, acronyms, and the likes. Also, folders are very often organised in hierarchies. Further, file-systems generally store modification times, and sometimes creation and access times for each file. Generally, for every file, may it be part of a file-system or attached to an e-mail, specific tools allow to obtain a lot of information from metadata embedded in the file itself, such as titles or authors. File-systems may cover hard discs, as well as external media such as CDs or DVDs or online storage.

4. (SEMI-) AUTOMATIC CONTEXT GENERATION

In this section, we discuss a number of dimensions along which inter-object context can be established. Context exists in several different forms, ranging from a very low-level technical context in which the object was created, via its immediate context of use (people involved, the project or activity it is related to, etc.), to a wider sociological, legal or cultural context. All levels of context are of importance for the authentic interpretation and usage of a digital object. However, we focus predominantly on the narrower focus of context that can be determined (semi-) automatically.

We thus consider the detection and documentation of context of digital objects as a semi-automatic process along several different and partially orthogonal dimensions, each of which structures objects according to different aspects. We currently use the following dimensions in our first prototype:

- the *time* of object creation and modification

[4] `http://www.cs.cmu.edu/~enron/`

[5] Basic Support for Cooperative Work, `http://www.bscw.de/english`

- the object *type*
- the *people* involved
- the *content* across different sub-categories, such as
  - the topic
  - the genre
  - acronyms, for example in project names

It has to be noted that the number of potential dimensions that digital objects can be organised by may be larger, such as references to places and other more detailed semantic concepts, existing (hierarchical) structures at the document source site, e.g. directory structures, e-mail folder structures and conversation threads, as well as specific object characteristics, such as embedded metadata (for example EXIF headers for digital photographs), for which specific extractors are needed. Especially the content of objects allows content establishment along several characteristics, such as the structure, format or layout used, or the existence of specific chunks, such as usage of the same logo indicating a specific purpose or origin of the object, and many more. Thus, the space spanned by these characteristics can become very high-dimensional.

The dimensions can be used separately and independently of each other, or be combined. For example, we can establish the temporal context of objects generated in a periodically repeating process, such as a yearly process of defining the work for the upcoming project year, or work for a periodically repeating conference series. If the information objects are analysed along the project dimension only, it is not immediately obvious that the objects might be stemming from different periods, and might thus not be that highly correlated within each time period. On the other hand, grouping the objects only by the time dimension allows e.g. for easy discovery of the actual focus of work in a specific period, while those objects might come from several otherwise totally independent activities. Thus, different dimensions need to be combined to establish a more precise object context.

As mentioned above, the concept of using various different dimensions as orthogonal views on the data is inspired by the concept of data warehouses and the data analysis method used therein, on-line analytical processing (OLAP) [14]. A central concept is the *OLAP cube*, which prepares the data for fast multi-dimensional queries and analysis. The analyst can pivot the data in various ways, e.g. see all the sales for a specific city for a certain product, and do this at various different levels of aggregation, allowing easily to get more detail on demand ('drill down') or a more abstract, summarised view ('roll up' or 'drill up').

Typical examples for dimensions used in data warehouses are time, geographical information, product numbers/information, customer information, etc. Often, they can be be arranged in hierarchies, to provide contextual information at different levels of detail and aggregation, respectively. In a traditional data warehouse, the time dimension could be organised as certain periods of a day, the weekdays, weeks, months and years, while geographical dimensions might be aggregated from boroughs to cities, counties, states and countries. Organising a data warehouse with a central fact table and hierarchical dimensional tables is called a 'Star' or 'Snowflake schema'. A graphical representation of the latter is given in Figure 4, depicting some of the dimensions we'll present later in this paper.

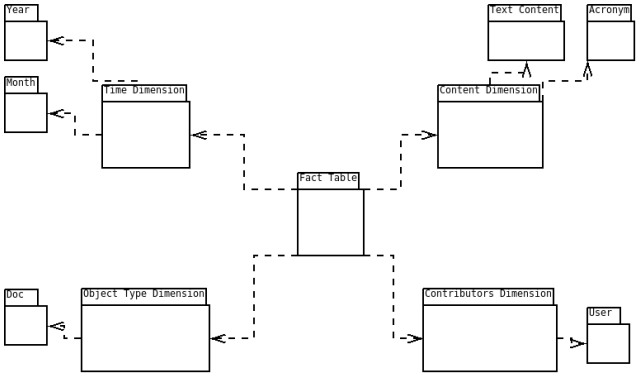

Figure 1: 'Snowflake' schema of a data warehouse for context information. The central table states the fact about the digital objects (e.g. metadata). The dimension tables such as content or type signify the various categories to organise the objects by

In the following we describe the dimensions identified for application to the process of establishing context for digital objects in more detail. The degree to which the context establishment can be automated varies among the various dimensions, ranging from fully automated along pre-defined categories, such as MIME for the file type, or the time dimension, to semi-automated e.g. for projects, where an initial hierarchy of projects can be proposed. Full potential is achieved by prompting for human interaction to improve the quality.

4.1 Time Dimension

The time-dimension can be defined very much in analogy to conventional data warehouses, and could be structured as follows: 'hour of the day', 'day of the week', 'week', 'month', 'quarter', and 'year', with 'week' forming a separate aggregation branch in the hierarchy. The desired level of hierarchies for this dimension can be very easily and quickly customised to the precise needs of the application and user.

Extracting the time dimension from digital objects is quite straight-forward. For e-mail objects, one can simply use the (sending) 'Date' field of the e-mail, for files in the user's directories, we can analogously use creation and modification dates. Other online systems such as Wiki or BSCW normally also store at least an upload time. Moreover, many file formats have embedded meta-data for creation or modification dates.

4.2 Object Type Dimension

Digital objects can be brought in context by their type. One categorisation for file types is the 'Internet media type', sometimes also called 'MIME-type' or 'Content-type'. This categorisation is a de-facto standard for object exchange over the Internet; it is well suited, as it provides an, even though limited, hierarchical order on digital objects, in the so-called (primary) 'type' and 'subtype'. Primary types are among others 'text', 'audio', image', 'video' and 'application', while subtypes can detail the content further, e.g. as 'text/rtf', 'image/gif' or 'application/pdf'.

Several repositories already provide object type information – online collaborative tools and e-mails generally provide MIME type information for attached files. Moreover, several tools exist to correctly identify the types from the object if this information is missing, e.g. when considering files in a directory.

However, the hierarchical structure defined by MIME is not very elaborate – on the one hand, it provides only two levels of hierarchy, and on the other hand, the primary type 'application' holds a huge spectrum of very different file types. However, a range of other characterisation tools that provide more detailed information of the characteristics of a digital object exist. Thus, instead of MIME, the Digital Record Object Identification tool 'DROID' might be used in combination with the registry PRONOM [3].

4.3 Contributors Dimension

The persons involved in creating, modifying or using digital objects indicate a relation between those objects, for example if the same group of people are working together on multitude objects. These groups constitute organisational units such as companies or departments, or orthogonal project teams. To populate this dimension with data, we need to first identify the persons involved in the object creation process and usage. For several file formats, parts of this information is stored as explicit metadata in the object itself; for other objects, this information can be derived from its storage repositories. With e-mail mailboxes, sender and receivers are known for each message exchanged, and subsequently also for other digital objects attached to the messages (this information can be further utilised, as often the creator of a document is the initial sender of an attachment). Wiki pages have a complete change history, so do revision control systems such as CVS or Subversion.

4.3.1 Identity detection

The identities extracted from the digital objects follow several patterns. They can be just the names, names combined with an e-mail address, just an e-mail address, or specific user names, which are in some way related to the e-mails or names, e.g. being parts of that, or abbreviated version, but may also be arbitrarily chosen. For the subsequent processing steps, we can thus enhance the quality of the context by first eliminating duplicates from the collection of digital objects. A duplicate in our case is a user contributing to the authorship of digital objects under more than one personal name or identity, e.g. the name info set in e-mail clients. This can in the simple case be a different spelling of the name, such as mixing up the order of first name and surname in two different word processing applications, or more complex issues such as having a different user name for the Wiki and the Source Code Repository. In the latter case, the person was maybe identified by a specific user name, but person identification can still be possible via the e-mail address associated with the user name. Obviously, such duplicates have an severe impact on the outcome of any data mining and context detection tasks. If the personal information on a person's home computer is set up slightly different from the one at the office workstation, a certain fraction of (e-mail) contributions (i.e. everything submitted from her or his home computer) will be under this alternate name. It is therefore important to resolve these issues before starting authorship analysis.

In our approach, we first identify persons using multiple personal names for the same e-mail addresses. We then substitute each personal name with the one most often used in combination with this address. This alone filters out many authors with duplicate addresses and finds an author's main address or personality. Then, we identify persons operating with multiple addresses and unify these based on substitution rules. Although rather simple, these rules reduce the amount of duplicates to a satisfactory level and expose relations in a better way. As an orthogonal step, users can define custom rules if they know two unresolved names are of the same identity, as detailed below. Then, we can match these identities to those extracted from file meta data or online repositories, using both the names and e-mail addresses associated. Some of the resolution steps can be done automatically, e.g. swapping the order of name and surname or handling initials, while others can be done semi-automatically, with the system proposing likely matches and the user interactively confirming them.

4.3.2 Author communication

For several object types and from several repositories, we can extract a richer set of information on the persons involved, beyond knowing just the creator. For many file formats, we can extract information on the people involved in the complete creation or modification process. E-mails exhibit rich information on people involved in the same conversations in their sender and recipients fields, while Wikis also keep a history of the people editing a specific page. Likewise do version control systems. Some online collaboration systems such as BSCW further keep information on the persons viewing specific attachments.

We can thus utilise this information to construct a social network graph of people related to each other by using the same objects, to allow the user to analyse the relationships of contributors in his repositories of digital objects. The connections in the graph are defined by people being among the creators and users of the same object, e.g. the senders or recipients of the same e-mail exchanged, while the strength of the connection is the count of these co-occurrences.

The graph can then be visualised and searched interactively, as illustrated in Figure 2. We provide highlighting mouse-over effects, which mark the selected person and highlight all the connections. This helps detecting sub-graphs of people closely related. Further, the user can control the level of detail with a variable threshold value of minimal co-occurrence, to only display major links in the graph. The persons (nodes) visible in the graph can further be controlled by providing criteria that have to be matched, such as part of the names or e-mail addresses. This view provides a good starting point for analysing the social connectivity among the users of the digital objects. Further, the visualisation can be used to interactively improve the quality of the identity resolution, as it allows to 'merge' two names that have not correctly been resolved to the same person. This modification is persisted by generating a manual rule for future identity detection.

Instead of a flat dimension of people working together, we want to introduce a hierarchy and aggregation, in the form of people working together at different levels. Therefore, similar to the graph above, we first construct a matrix of all the persons involved in the manipulation of our digital objects, and count how often they are among the persons

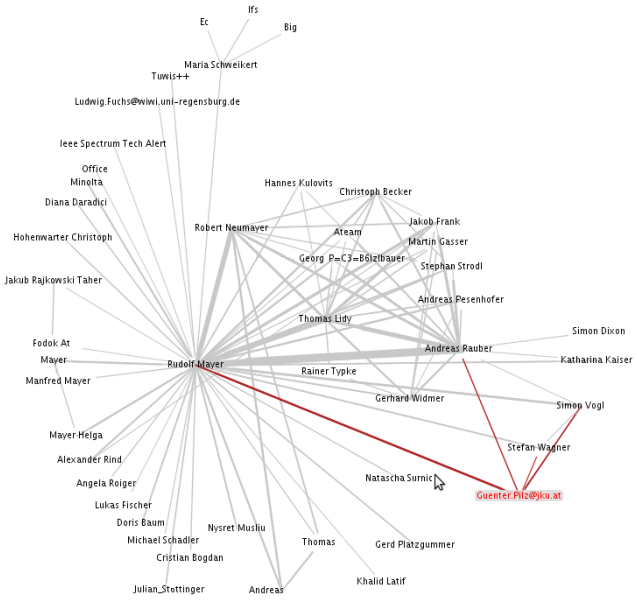

Figure 2: Graph view of the social network

involved in the same objects, as an indication on how much they are working together. We can then apply an agglomerative, hierarchical clustering algorithm. In the beginning of such a process, every item forms its own cluster, and in each subsequent step, the two most similar clusters are merged, until finally only one cluster remains. We employ Ward's linkage [11] (also known as minimum variance clustering) as one with the highest performance within the linkage clustering families. In this algorithm, the distance of each pair of clusters is defined by the increase in the 'error sum of squares' (ESS) if the two clusters are to be combined.

The result of the Ward's algorithm is a hierarchy of clusters the user can browse through. Increasing the number of displayed clusters means splitting existing clusters into two new ones, while reducing the number of clusters is achieved by merging two clusters into one. We can utilise the clustering tree to generate a hierarchical dimension, if we define a number of n levels, with m_n different clusters on that level. We can thus detect sub-groups among bigger groups of people collaborating.

Choosing a hierarchical clustering algorithm is motivated by the fact that in a non-hierarchical clustering, producing a clustering with a different (higher or smaller) number of clusters might change the layout and contents of the clusters to a large extent, which would thus not allow to create a hierarchical dimension.

4.4 Content Dimension

Even though the content of an information object does not constitute context per-se, other information objects that share some similarities in their content with a specific object do form a context of that object. Content relation can be detected on a plethora of different aspects – the usage of similar keywords, or similar style, similar chunks, etc.

The content dimension of digital objects refers primarily to the plain text of a document itself, or text documents which describe the object more closely. In the case of text documents the content is obvious – for non-textual attachments, e.g. the Wiki page or the e-mail body holding the attachment may serve as a surrogate. This method has the advantage of establishing context for all types of attachments, whether they can be textually indexed themselves, or not, as e.g. image or video data. The attachments as special purpose files can also be brought in context with the data found in the rest of the e-mail, Wiki page, or any other container they were originally attached to.

Different representations of textual content can be considered for analysis. These range from standard bag-of-words based full term indexing [23], to more advanced indexing techniques developed to identify specific patterns and types of content. Other aspects that can be extracted using natural language processing tools such as GATE, are names of persons, places, or dates. These indexing techniques result commonly in a vectorial representation of documents that can further be analysed using machine learning techniques.

We want to specifically point out the importance of 'within-project' similarity, as a special case of content similarity and vital aspect of an objects content. Digital objects created for and within the same project share a strong contextual relation, and automatically detecting them is thus a desirable goal. Projects may be characterised by dealing with a certain topic; however, that assumption might not hold for larger projects with several independent tasks. Further, it is difficult to label such topical relations with meaningful, unambiguous terms. Therefore, another approach to identify projects is simply to detect project or task names; very frequently, these names are in the form of acronyms, thus project identification may be reduced to acronym detection. However, acronym detection is a difficult task on its own. Acronyms are often characterised by being written in upper-case, or at least mostly upper-case, letters. Even though acronyms are often designed to be pronounceable, many of these are not dictionary words. However, there is still quite a number of acronyms that are actually (frequent) dictionary words, such as e.g. the 'MUSCLE' or 'PLANETS' EU projects. Consequently, more sophisticated approaches to detect acronyms in a text document, such as comparing occurrence frequencies of this term with the terms in a reference corpus or checking with a dictionary, have proven to exclude these kind of project acronyms. In our system, we thus employ a computationally simpler approach, by simply detecting all upper-case and mostly-upper-case terms in the subject lines of e-mails. Subsequently, applying a frequency-thresholding on the terms yields the most important acronyms – terms that occur too seldom can be excluded as being not relevant enough (or potentially wrongly capitalised words). The user can interactively adjust the thresholds for the cut-off, to fine-tune the results for his collection. Additionally, from the list of acronyms proposed by the system as result, the user can manually black-list terms as non-acronyms, and provide undetected terms in a white-list. With the knowledge of the users on the projects they are involved in, they can quickly scan for wrong terms, and thus have a clean list of acronyms within a few minutes.

Once acronyms are detected, we can re-scan the digital objects for occurrences of these terms, this time not regarding any capitalisation, as often, acronyms are not properly capitalised, especially in more informal objects such as the e-mail body. Then, we construct a feature vector representation of each acronym, where the e-mails it occurs in are the features. Applying clustering we create a hierarchical

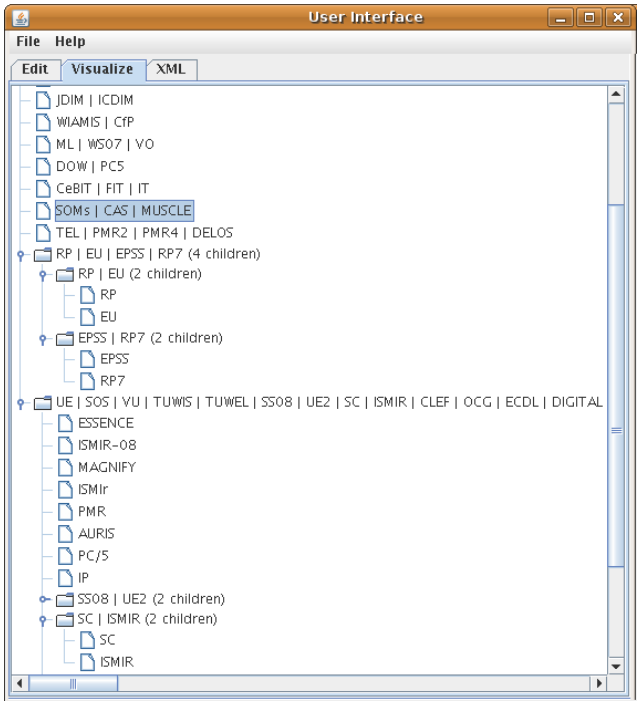

Figure 3: Hierarchical view of projects, identified by their acronyms

dimension of related projects.

One example of an acronym clustering result can be seen in Figure 3. We can see a lot of fitting acronym pairs, for example the highlighted grouping of the EU project 'MUSCLE' with a sub-project labelled 'CAS', and one of the technology applied therein, 'SOMs' (Self-Organising Maps). Also the other matches found by the algorithm group related acronym, for example a conference and its ensuing special issue of a journal ('ICDIM' and 'JDIM'), or the course 'Machine Learning' (ML), taking place in the winter semester (WS), and being of the specific course-type lecture (VO). These can be manually grouped to form semantic classes such as teaching, conferences or projects.

4.5 Other dimensions

Another possible dimension, not yet implemented in our system, contains information about which objects are used generally together. This can on the one hand be detected via direct dependencies of such files, such as an image being referred from an HTML site. On the other hand, systems tracking file-system calls to identify which files are accesses at the same time, such as the system presented in [25], give indications on files that are generally used together. Numerous other dimensions need to be added the system prototype to fully exploit its potential, including object-specific metadata such as EXIF headers, or electronic signatures, advanced features such as face and logo recognition for images, intention, style, modification (differences) in versions of objects, to name just a few.

4.6 Combining the Dimensions

The previous sections illustrated the usability of views on singular dimensions, and their ability to establish and visualise the context of digital information objects. However, combining the isolated dimensions opens plenty new exploration and analysis possibilities. In data warehouses, data in the OLAP cubes is often visualised by means of a pivot table, which is a summarisation tool that can automatically sort and aggregate data from a table, and display thus condensed information in a smaller, second table. Filters can be applied, roughly an equivalent to the 'drill down' concept in OLAP cubes. With such a tool at hand, users can in an interactive process quickly change the abstraction level of the data displayed. This can be an important aid to discover more complex contextual relations between the digital information objects.

An example of such a pivot-table is depicted in Figure 4, created from digital objects found in the e-mail mailbox of one of the authors. In its initial state, the pivot-table holds detailed information about each digital object in a simple tabular fashion. For this specific object source, the user can choose to show either e-mails or their attachments. Sorting by columns such as the subject, sender, or date, as well as several filtering tools, allowing to restrict the digital objects displayed to those fulfilling certain properties, are provided as basic operations. Filters can e.g. be applied to the sender or group of senders, acronyms or groups thereof, keywords and if the focus is on showing attachments, the content type.

Further, the user can switch the display to aggregating the objects along the x and/or y-axis of the view. For example, one can choose to group the objects by the associated acronym on the x-axis. This view then still gives the full tabular information for the objects, which is intended to give detailed information about the objects, but still allowing fast selection of relevant information by spotting the acronyms interesting to the user. This view is thus well suited e.g. for searching in a collection.

If the user additionally selects to group the objects also by a second dimension, the view changes slightly to the classical pivot-table, where each cell of the table now holds those objects that have common properties among both dimensions. This view is the one illustrated in Figure 4. An important aspect on this view is the wealth of interactivity and adjustment possibilities – for example, if we have selected a grouping by time, and chose months as the granularity level, we still can combine several months to form a new logical group. For example, in one of the authors collection, one can find a wealth of e-mails that got assigned to the acronym 'ECDL', the European Digital Libraries conference series. Filtering them by this acronym, and then selecting the time as second dimension, the user quickly realises that the digital objects stem from different 'instances' of this conference, namely once from organising the ECDL 2005, and subsequently for submitting papers to conferences in the other years. Also, an object type dimension can help in this case, as the types of files attached to the acronyms differ greatly, being mainly text documents (LaTeX and PDF) for the authoring activity, and spreadsheets and other office documents for the organisational work. To distinguish easily between the different stages of scientific aspects of a conference instance, we can combine units along the time dimension to form a logical grouping to spot activities related to the paper writing process, work for a final version, preparing presentation slides, and writing reviews.

Another example would be selecting all objects related with the project 'MUSCLE', a European Union funded re-

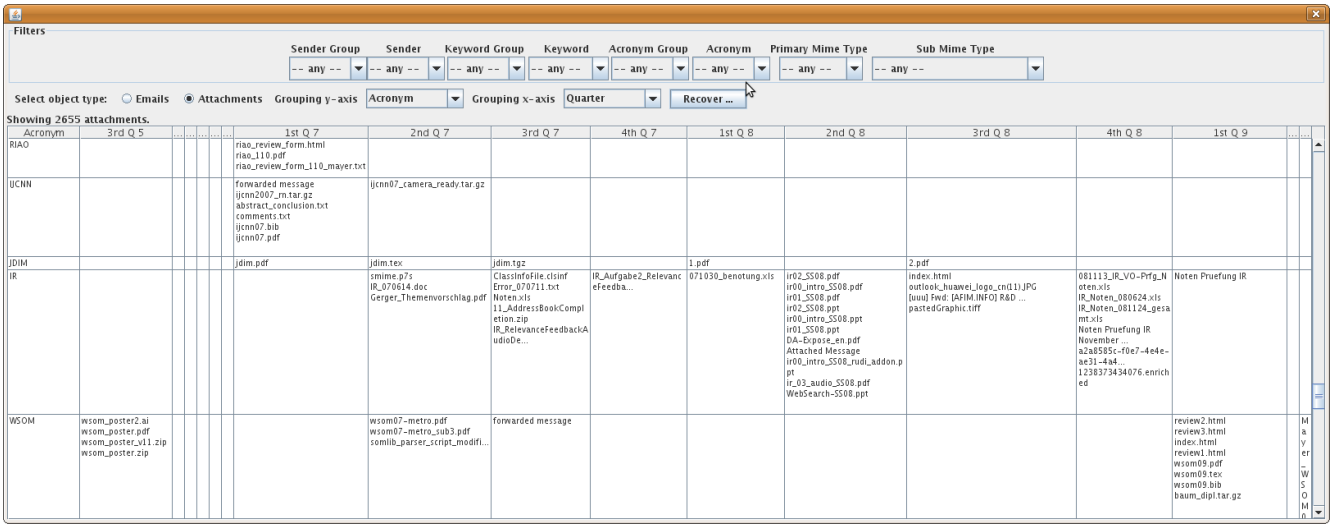

Figure 4: Pivot-table View on E-mail Attachments

search project with several dozens of partners. Thus, selecting the contributors dimension, objects get correctly separated by the different groups of people working on the specific work-packages in the project.

5. USING CONTEXT: APPLICATION SCENARIOS

The context of digital objects – or of information in general – is essential to allow us to correctly interpret and use information. Automatically establishing context thus assists in virtually any task where specific digital objects are concerned and where the context is not obvious to the user anyway. This obviously affects very large archival holdings, where multiple users are working with specific pieces of information. It also affects small and medium enterprises and home office as well as private users, all of whom accumulate significant amounts of information, up to a degree where they lose oversight, and are unable to find all relevant pieces of information related to a certain task, or forget which of the several different versions of an object is the one most up-to-date or who it was sent to. The focus of the current prototype, still, is to first establish different levels of context, and to document them by enhancing according XML descriptions of objects for subsequent ingest into a repository. Still, to assist in evaluating and driving the development of the current prototype system we have considered a small set of different tasks to cover a range of application scenarios.

5.1 Object ingest in digital libraries

A scenario in a more professional setting relates to ingesting larger quantities of otherwise rather unstructured, and potentially unknown, information, such as personal data collections by famous persons bequested to archival or memory institutions. Ingest is a specific stage in the OAIS [10] reference model, where new entries (Submission Information Packages, SIP) are to be stored in the archive. One important stage in the ingest process is to generate descriptive information about this data to create the Archival Information Package (AIP) out of the SIP.

Identifying which objects are enclosed on a number of data carriers in a box and how they relate to each other, what they are dealing with, and when they were created, are important steps for correctly generating *Context Information* and *Provenance Information*, which are parts of the Preservation Description Information (PDI) descriptive information contained in the Archival Information Package.

In the OAIS model, 'Context Information documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects existing elsewhere' [10]. 'Provenance Information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This gives future users some assurance as to the likely reliability of the Content Information. Provenance can be viewed as a special type of context information' [10]. Manually establishing this descriptive information constitutes a rather tedious task, that can be supported by advanced tools that help to identify and suggest potential contexts.

Along the same lines, such a solution may assist small institutions and home users to move their ad-hoc 'curation' of their digital objects to an improved level by helping users to collect and maintain context information as part of a small office or home archive [26].

5.2 Disaster recovery

A surprisingly frequent scenario in less professionally managed computing environments is the loss of significant amounts of data due to hard disc crashes or loss/theft of computers or laptops. In many cases all information on a users home directory or personal hard-disk are lost, with no or only heavily outdated back-ups being available to recover data from. However, thanks to current work practices, evidence has shown that a rather large fraction of objects has been sent or received via external communication channels such as downloads from project websites or, specifically, via e-mail. These objects are usually hosted on external systems such as web or email servers, that are frequently not af-

fected by local disasters. Manually recovering the massive amounts of information from these repositories tends to be an almost unmanageable endeavour. By assisting users in automatically extracting and grouping objects, a goal of disaster recovery assistance would be to automatically re-create valuable content from such repositories on a user's hard-disk in a structure that can be used feasibly. This requires correct grouping of objects into a re-created directory structure that reflects a conceptual model comprehensible by a user, as well as an understanding of versioning aspects and others.

5.3 Semi-automatic user support

We see one of the application scenarios in supporting users in everyday tasks. Whenever interaction with digital objects is required, context data can help in finding, opening, storing, or accessing similar objects. The most simple application is suggesting folders as a context menu for the 'save-as' use case. A system relying on context information gathered in a way as in our prototype may be able to provide accurate suggestions for storage locations on different levels. An attachment from an e-mail received from another member in a certain research project may be saved along with other relevant documents for this project; via analysing multiple views, we can provide such suggestions. A primary task, however, definitely will be to support search. As most search algorithms currently only analyse the textual content of an object, retrieval is limited to objects containing the terms specified in a query. This is of limited use for identifying predominantly non-textual objects such as multimedia data (images, video, audio), or event textual documents with only limited quantities of text that sometimes may not match the query terms (such as, e.g., slides for presentations).

6. CONCLUSION AND FUTURE WORK

Documenting the context of digital objects is essential if we want to ensure proper usage and interpretation. This challenge becomes increasingly prominent as digital objects and pieces of information become more distributed and isolated. Many programs used for creating digital content, such as cameras, word processors, etc. already try to capture a set of context information automatically. On top of these, complex workflow support programs document the process and types of usage of digital objects in a given environment to further enhance the amount of context information that can be accumulated. Still, we find many situations where essential aspects of context have to be established and documented manually, often ex-post when documents are to be ingested into a central repository. This situation is even worse in less professional environments such as small office / home office settings.

In this paper we have presented a framework that assists users in establishing and documenting the context of digital objects by analysing their relationship across a number of different dimensions. More specifically, a range of tools and approaches from the information retrieval and machine learning domain are used to extract correlations between objects in time, according to the people involved either as creators or recipients, or according to their content in different degrees of abstraction ranging from acronyms to complete full-text indexing. Once these characteristics have been extracted, the interface facilitates automatic as well as semi-automatic analysis of groupings according to different dimensions. This helps in identifying which digital objects were of interest to which groups of people at which points in time, which allows the user to establish a certain context of use via a sequence of objects. Borrowing from the principles of on-line analytical processing in data warehouses, multiple views at different levels of detail help in establishing different types of context for each object, be it an e-mail attachment, a file in a home directory or on a CD, or even stored remotely on Wikis or being part of a blog.

The information gained by establishing and documenting this context may serve as additional metadata when ingesting an object into an archive, to improve the quality of interpretation of an object. It also serves predominantly as a finding aid, supporting to a certain degree semantic search. Based on the relationships and their qualifications documented, objects can be retrieved because they are obviously related to a certain aspect, even when the initial keywords in a full-text query do not match the content of a specific object. This way, photographs that relate to a project meeting can be identified together with the meeting's minutes due to correlations in the temporal and recipient domain if distributed via e-mail or shared web storage.

The current prototype implementation has revealed the potential of this approach. In order to fully exploit it, additional tools need to be integrated into the system that extract more specific metadata for various object types. This includes specific metadata headers, such as e.g. EXIF headers for digital photographs, or metadata embedded in office documents or PDF. Similarly, additional source-specific meta-information will be extracted, especially as additional sources such as web blogs and Wikis are integrated. Further improvements are required in terms of system usability, allowing more flexible and intuitive interaction when analysing correlations across various dimensions.

While initial studies focused primarily on establishing context at various levels of detail, we will move on to qualifying the type of context and adding interpretation (or at least suggestions for interpretation to be manually confirmed) in an automatic manner by analysing the semantic concepts between relationships, according to the dimension they occur in. These could be temporal sequences, and thus versions, of a document that otherwise has the same file name but increases in size, or discussions on a certain document if the filename or title appears in the content of another document, to give just a few examples. Apart from the focus on correctly determining the context of objects, test are planed to evaluate the suitability of these concepts of context for different tasks, ranging from search and retrieval challenges such as identifying all documents of a certain characteristic that are related to a certain project. Another core challenge is the evaluation in how far the context established is sufficient to recreate a usable structure on a file system from an otherwise rather flat repository structure, such as encountered when recovering lost data on a storage device from external repositories such as Wikis and e-mail servers in disaster recovery settings.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

7. REFERENCES

[1] S. S. Bhowmick, W. K. Ng, and S. Madria. Web schemas in whoweda. In *Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP (DOLAP '00)*, pages 17–24, McLean, Virginia, USA, November 10 2000. ACM.

[2] D. Biber. A taxonomy of english texts. *Linguistics*, 27, 1989.

[3] T. Brody, L. Carr, J. M. Hey, A. Brown, and S. Hitchcock. PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation*, 2(2):3–19, November 2007.

[4] R. G. Capra, C. A. Lee, G. Marchionini, T. Russell, C. Shah, and F. Stutzman. Selection and context scoping for digital video collections: an investigation of youtube and blogs. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital Libraries (JCDL 2008)*, pages 211–220, New York, NY, USA, 2008. ACM.

[5] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[6] J. Diesner, T. L. Frantz, and K. M. Carley. Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory*, 11(3):201–228, October 2005.

[7] S. Foo and D. Hendry. Desktop search engine visualisation and evaluation. In *Proceedings of the 10th International Conference on Asian Digital Libraries (ICADL'07)*, pages 372–382, Hanoi, Vietnam, December 10-13 2007. Springer.

[8] P. Holleis, M. Kranz, M. Gall, and A. Schmidt. Adding context information to digital photos. In *ICDCSW '05: Proceedings of the Fifth International Workshop on Smart Appliances and Wearable Computing*, pages 536–542, Washington, DC, USA, 2005. IEEE Computer Society.

[9] J. Hunter, I. Khan, and A. Gerber. Harvana: harvesting community tags to enrich collection metadata. In *Proceedings of the 8th ACM IEEE Joint Jonference on Digital Libraries (JCDL'08)*, pages 147–156, Pittsburgh, PA, USA, June 16-20 2008. ACM.

[10] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[11] J. H. W. Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.

[12] J. Karlgren. Stylistic experiments in information retrieval. In *Natural Language Information Retrieval*. 1999.

[13] Y. Kim and S. Ross. The naming of cats: Automated genre classification. *International Journal for Digital Curation*, 2:24, 2007.

[14] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*. Wiley, 2002.

[15] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, May 2000.

[16] R. Mayer, R. Neumayer, and A. Rauber. Interacting with (semi-) automatically extracted context of digital objects. In *Proceedings of the Workshop on Context, Information And Ontologies (CIAO 2009)*, Chania, Greece, 2009.

[17] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital Libraries (JCDL 2004)*, pages 53–62, New York, NY, USA, 2004. ACM.

[18] E. Neuhold, C. Niederee, A. Stewart, I. Frommholz, and B. Mehta. The role of context for information mediation in digital libraries. In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, and E.-P. Lim, editors, *Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004)*, volume 3334 of *Lecture Notes in Computer Science (LNCS)*, pages 133–143, Heidelberg, 2004. Springer.

[19] G. S. Pedersen, K. F. Christiansen, and M. Razum. The use of digital object repository systems in digital libraries. *D-Lib Magazine*, 14(11/12), November/December 2008.

[20] A. Rauber, A. Aschenbrenner, and O. Witvoet. Austrian on-line archive processing: Analyzing archives of the world wide web. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002)*, pages 16–31, Rome, Italy, September 16-18 2002. Springer.

[21] A. Rauber, M. Kaiser, and B. Wachter. Ethical issues in web archive creation and usage – towards a research agenda. In *Proceedings International Workshop on Web Archiving and Digital Preservation (IWAW'08)*, September 2008.

[22] A. Rauber and D. Merkl. Text mining in the somlib digital library system: The representation of topics and genres. *Applied Intelligence*, 18(3):271–293, May-June 2003.

[23] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[24] S. Schrimpf. Long-term preservation of electronic literature. In *Proceedings of the 5th International Conference on Preservation of Digital Objects (iPRES 2008)*, pages 29–30, London, UK, September 29-30 2008. British Library.

[25] C. A. N. Soules and G. R. Ganger. Connections: using context to enhance file search. *SIGOPS Oper. Syst. Rev.*, 39(5):119–132, 2005.

[26] S. Strodl, F. Motlik, K. Stadler, and A. Rauber. Personal & SOHO archiving. In *Proceedings of the 8th ACM IEEE Joint Jonference on Digital Libraries (JCDL'08)*, pages 115–123, Pittsburgh PA, USA, June 16-20 2008. ACM.