INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED11
15 - 18 AUGUST 2011, TECHNICAL UNIVERSITY OF DENMARK

APPLYING CONTEXT TO ORGANIZE UNSTRUCTURED INFORMATION IN AEROSPACE INDUSTRY

Yifan Xie[1], Steve J Culley[2], Frithjof Weber[3]
(1) University of Bath, UK (2) University of Bath, UK (3) Airbus Operations GmbH

ABSTRACT

A large portion of information created within business activity is unstructured and as a consequence much useful information is buried inside. Within the aerospace industry, an added challenge is that long product lifecycles require such unstructured information to be accessible over a long period of time. In this paper, the authors have examined and identified promising techniques that can collectively contribute to the better organization of unstructured information. Two industrial case studies were conducted to examine the current practices of organizing this unstructured information in some engineering departments. As a result, a number of key challenges in organizing and dealing with the unstructured information *elements* within an engineering setting are identified. Subsequently, a set of requirements of a desired intelligent system are developed, particularly associated with the important guiding topic of context. These requirements are then used to guide the design of an example Context Based Search Platform which demonstrates promising potential for dealing with multi-dimensional and complex data sets.

Key words: information context, information semantics, context aware, unstructured information, information retrieval, knowledge discovery, industrial case study

1. INTRODUCTION

The adoption of information technology (IT) has completely changed the way information is generated, propagated and consumed in the working space. Terms such as “file” and “document” now relate more with digital entities in the cyber space than piles of paper. However, the rise of IT also ushered in the explosion in the volume of digital information [1, 2]. Information is generated with ever increasing ease and captured in various digital data repositories. This leads to the phenomenon of information overload: on one hand there is too much information stored in too many repositories, whilst on the other hand knowledge workers face increasing challenges to organize and retrieve useful information. From an information life cycle perspective, namely that of creation, storage, retrieval and maintenance [3, 4], increasingly focus has shifted from creation to retrieval and maintenance of information. In the last decade, the ‘vibe’ of IT has been set by well established companies such as Google who specialize in information retrieval, while one of the most promising benefits of Web 2.0 has been the potential to facilitate better understanding of information [5, 6]. There seem to be 2 important dimensions that effect engineers, namely that of unstructured information and also how context can be linked to these information sets to aid, both retrieval and understanding.

1.1 Unstructured information

A topic that increasingly attracts interest is how to organize unstructured information for better retrieval and reuse. Unstructured information generally refers to information that either does not have a data model or has one that is not easily usable by a computer program [7, 8]. Various researchers have indicated that as much as 80% of information generated in the workspace is captured in various forms of unstructured information such as audio, image and unstructured text [9, 10]. Despite advances in information management systems, much useful information is still buried within these large volumes of unstructured information. A large body of work has been conducted from various research and commercial fields for the better organization of unstructured information. Many of these fields, such as social computing, data mining, semantic technologies and context awareness, are inter-related and inter-complementary [12]. It is fair to say with the increasing complexity of today’s IT infrastructure, there is no “silver bullet” to address this challenge, however the use of context may be one of the elements that may help.

1.2 The use of context

Depending on existing infrastructure and community practices, contextual elements that are available to be captured or exploited can vary greatly, and so do the choice of specific technique(s). These issues relating to context will be discussed in relation to engineering and design activity in the aerospace industry.

For the aerospace industry, an added challenge for organizing unstructured information is that long product lifecycles require such information to be organized and accessible throughout long period of time. For example, the repair and operational history of a given aircraft typically spans 2-3 decades. Within such a long lifecycle, multiple generations of information systems would have been used and become obsolete, and access to information needs to encompass such diversity. Moreover, industrial protocol, engineering language, and business practice are changing all the time, yet future generations of aerospace engineers and designers still need access and to make sense of today's information.

1.3 The purpose and content of the paper

Four elements to this overall problem will be addressed by this paper. 1) There is relatively little work to study the challenges of organizing unstructured information in such an industrial setting, 2) Also it is not clear how techniques that capture and exploit contextual elements can bring actual benefit to businesses. 3) There is a lack of research work that systematically looked into the lifecycle of unstructured information with specific business scenarios. 4) The facilitation of coherent collaboration between different techniques and existing practice has not been fully considered.

This was achieved by the research activities described below that are then presented in the paper. The authors performed a survey in research fields, examined and identified promising techniques that can contribute to better organization of unstructured information. Two industrial case studies were then conducted to examine current practices of organizing unstructured information in engineering departments. As a result, key challenges in organizing unstructured information within an engineering setting are identified. Subsequently, a set of requirements of a desired intelligent system is developed. The authors then introduce a Context Based Search Platform, a system that is designed according to the identified requirements. This system is then evaluated in relation to the two case study areas.

2. LITERATURE REVIEW

Various reports have identified the need to better organize unstructured information as a major IT challenge for the next decade, and the key to overcome this challenge is to better capture and use the context of such information [12, 13]. The word "context" is somewhat ambiguous. The main driver to understand and organize information from a contextual perspective has been in research fields which have a "context-aware" focus or attempt to exploit semantic technologies. Context-aware approaches mainly address the topic of information usage to establish context, while semantic technologies focus on extracting the meaning of information for this purpose. On the other hand, fields such as data mining and social computing focus on understanding what is available within content of information entities.

An analysis of these separated strands of research work indicates that there seem to be 3 main contextual dimensions, namely: 1) *Context of Information Semantics*: What does a piece of information mean in certain domain? 2) *Context of Information Usage*: What is the context of information usage? 3) *Context of Information Content:* What can be extracted, analyzed and learnt from the content of a collection of information entities? These 3 dimensions are discussed within this section from 2.1 to 2.4

2.1 Context of information semantics

The term "semantics" often refers to *"meaning"*. In the context of information technology, semantics particularly refers to the meaning of different concepts as opposed to their syntactic forms [14]. The term "meaning" is related to information users' understanding of any given concept under a certain context. For example, the concept "Bull" in the context "Animal" refers to the male of a bovine animal, while in the context "Stock" refers to increasing stocks.

There are two core issues in capturing semantic context [15, 16]: The first issue is facilitation and construction of semantic models on given domain. This issue focuses on capturing domain concepts and relationships between concepts in a manual, semi-automatic or automatic manner. The second issue concerns using rules or logic to perform reasoning, so that information systems can automatically establish implicit connections to information pieces that meet users' information need, this is predominately achieved by understanding user inputs such as search queries. In terms of maturity of specific techniques, most notable are the set of XML based technologies advocated by W3C [17]. These are: 1) OWL for ontology semantic modelling; 2) RDF for resource description; and 3) Description Logic (DL) for semantic reasoning.

2.2 Context of information usage

This contextual aspect concerns the circumstance under which users retrieve and use information. The term "context" can be defined as any information that can be used to "characterize the situation of an entity" [18]. Systems that capture contextual information and adapt their behaviour are often referred as context aware systems [19]. As to what information is to be considered as elements that constitute context, Dey and Abowd identified identity, location, time and activity as basic contextual elements [18, 19]. Henricksen et al investigated the nature of such elements, for example whether they are dynamic or static [20]. Budzik [22] looked into how users' task-related context can be captured and exploited. Rhodes and Maes [23] explicitly took user's specific actions into account. Much of the literature in this area considers context in generic settings [20, 21, 22] and the most mature systems are those that exhibit location-aware capabilities [22]. From an industrial perspective, the VIVACE project [24] focuses on exploiting user context (e.g. roles, discipline, and task) for information retrieval.

Usage of information is often influenced by many factors such as a user's background, location of the usage, nature of information entities [11]. Increasingly the layered approach has been adopted to handle issues such as context capturing and context modelling separately [25]. For context capturing, a wide range of techniques are applied to extract contextual information from various sources, for example, physical sensors on mobile devise are used to capture user location; activity profiling techniques [4] are used to capture user interaction with computing devices; while static context such as user identification can be captured from profile information of traditional systems. Many existing works utilise similar set of technologies that support semantic technologies, namely XML based technologies such as OWL for ontology modelling, RDF for context entity description and DL for context reasoning [26].

2.3 Context of information content

Traditionally, techniques that describe and analyze information content mainly consist of automatic or semi-automatic approaches that originated from fields such as data mining. These techniques apply advanced algorithms such as clustering and also process specified patterns such as taxonomy based rules. Their key features are: 1) generation of descriptive information such as metadata and summarization; 2) dynamic grouping such as information clusters; 3) static grouping such as information categories; and 4) trend projection [27]. These in turn provide better support for information recognition, retrieval, navigation and decision making. Research in the aerospace sector [29] has indicated that clustering and classification techniques can help to improve the retrieval of engineering information.

More recently, social computing has provided a new generation of platforms [5, 28] for information users to describe information by leveraging collective knowledge. These platforms hold tremendous disruptive potential in today's business world [5]. By methods such as social tagging and commenting, users can provide the metadata that helps to distinguish useful information and also share their experience of usage.

2.4 Summary of Literature Review

Table 1 provides a summary of techniques that can handle information context from each of the contextual aspects described above. It has been noted that different techniques can be applied with inter-related efficiency. For example, social computing and semantic technique may be applied to achieve a tagging scheme that is adaptive to user contexts [30].

Contextual Aspects	Techniques (Specific Technology)
Information Semantic	• Ontology modelling (OWL) • Resource description (RDF) • Context inferring (DL)
Information Usage	• User activity profiling • Physical sensing • Ontology modelling (OWL) • Contextual entity description (RDF)
Information Content	• Clustering • Classification • Pattern Recognition • Social Metadata Facilitation(Tagging, Bookmarking, Comments)

Table 1: Techniques to handle context of information

The next phase of the work was to understand the potential to exploit the use of context in engineering situations. This was achieved through 2 separate case studies that had quite different characteristics.

3. INDUSTRIAL CASE STUDIES

This section contains 2 case studies which have different characteristics, but both have the potential to use context to deal with their rapidly expanding information sets. It is first necessary to set the industrial context and highlight the elements of the operation that are relevant.

3.1 Industrial background

As a leading manufacturer of civilian and military aircraft, Airbus develops and produces a wide range of aircraft models. The typical development cycle of a new model, from the early feasibility studies to testing and ramp-up and use is about 10-15 years. On the other hand, the typical in-service lifespan of an aircraft is about 15-30 years. Altogether these amount to a total product lifecycle of potentially up to 45 years. During this time, knowledge about the design and service of any individual aircraft need to be maintained despite the change in industrial practice, manufacturing and production technology, information technology and information users' habits.

A related issue is the way that the organization is structured in that this creates natural boundaries for internal information and knowledge transfer. Today the organization is constructed along effectively four dimensions: 1) Product portfolio which groups workforce by aircraft programmes (i.e. A320, A380); 2) Organizational functions which groups workforce by functional units (i.e. Finance, Customer Service); 3) Product structure which groups workforce by major aircraft components (i.e. fuel system, landing gear); 4) Location which naturally divide workforce among different locations (i.e. Filton, Toulouse).

The two industrial cases presented in this paper are located in Filton, U.K. The first case looks at how an In-Service Support department, which supports the repair and maintenance of aircraft wing structure, uses past repair cases to create instructions for customer repair requests. This case study looks at the issue with unstructured information from the perspective of an individual functional department. The second case looks at how the Fuel System group, which is responsible for all activities related to aircraft fuel system from design to in-service support, uses lessons learnt and best practices to improve the overall performance of the group. This case study looks at the issue from the perspective of an organizational entity that is responsible for major components of aircraft.

Each case study will give a general account on the nature of information practices and the objective of information usage, followed by description of existing practices to organize unstructured information under existing IT infrastructure. Discussion is then presented to summarize the findings on both case studies, identifying ongoing challenges in organizing unstructured information.

3.2 Case Study 1: Past Cases Usage in Wing In-Service Support (ISS)

The Wing ISS department is responsible for repairs associated with the aircraft wing structure. They provide support on a 24-hour-a-day, 365-day-a-year basis to design and validate repair inquiries by customers. The repair requests principally can be classified into two types: major repairs which require significant support such as an on-site repair party; and simple repairs which require the provision to the airline of repair instructions for on-site implementation. Simple repairs represent the majority of the repair requests and are the main focus of the industrial practice studied in this paper.

ISS employs two types of service engineers: design repair engineers for designing repair methods, and stress engineers for validating the proposed repair. The two groups are managed respectively by senior design and senior stress engineers. When a repair request is received, the typical in-service support workflow includes the following steps:

- *Case Allocation*: Senior design and stress engineers assigning the task to respectively a design and a stress engineer.
- *Repair Design*: The design engineer proposes a repair method for the reported damage
- *Repair Validation*: The stress engineer validates the repair method via stress calculations.
- *Repair Approval:* Both senior design and stress engineers examining and eventually approving the repair method. The repair solution is then ready for feedback to the customer.

ISS relies heavily on using past cases as reference for new cases. Typically, at the beginning of the *Repair Design* stages, design and stress engineers will collaborate to identify similar past repair cases that can be used as reference for repair design and stress calculations. One reason is that a similar past case often contains useful information such as repair methods, stress thresholds, damage photos, drawing information which can greatly reduce the response time of a new case. Another reason is that using past cases has proved to be a good source of on-the-job learning for novice engineers.

ISS keeps records of over 30,000 past cases in various structured enterprise information systems that provide information such as repair cases workflow, document templates, and product structure drawings. For various operational and historical reasons, none of these systems provides satisfactory amounts of information for past case records, in other words, none of these systems gives what can be thought of as the “complete picture”. In order to retrieve appropriate past case information from such a big case corpus, ISS compensates the lack of an “all-in-one” information portal and efficient retrieval facility by creating and maintaining Past Case (PC) documents within the corporate shared drive:

- *Creation*: A PC document in PDF format is created when a repair case is finalized. Each document contains key information pieces, which were used during the handling of the case, taken from various related enterprise systems. Such information includes customer emails, damage photos, related drawings, repair method description and stress validation.
- *Storage*: Each PC document is sorted and stored into various locations within the corporate shared drive, following allocation guidelines specified by domain experts.
- *Retrieval*: A record for each PC document is created in a Microsoft Excel spreadsheet which contains key metadata of past cases in tabular format. Excel VBA Marco is then used to programmatically recreate hyperlink to the actual PC document by analyzing the metadata. It is worth noting that direct access to the past case documents is inefficient due to the shear amount of documents and lack of past case context when accessing via folder/file structure.
- *Maintenance:* An experienced engineer is assigned to oversee the creation and storage of PC documents and maintain the spreadsheet full time. Typical tasks include removing inappropriate information, reviewing and update past case metadata in the spreadsheet.

3.3 Case Study 2: Lesson Learnt and Best Practices Usage in Fuel System (FS)

The FS group is responsible for activities related to fuel systems across the products portfolio. These include high-level strategic decision making, design of fuel systems, research and development of next generation technological capability, monitoring production of existing designs, and providing in-service support to all aircrafts in operation.

The FS group handles a wide variety of engineering tasks from strategic to operational issues, from high level conceptual design to detailed engineering practices, and also contains engineers from wide ranges of disciplines. Therefore the top priority for FS is to ensure experience accumulated during various activities is captured and shared among the group as well as communities across the organization. The current practice to share such experience is via periodic review of lessons learnt and best practices at the beginning of each key project phases, with the following process:

- *Initial review*: Existing lessons learnt and best practices are reviewed by domain experts and the ones that have potential high impact for the current project phase are identified.

- *Allocation*: High impact lessons learnt and best practices are allocated to engineers who can benefit by learning and reusing the captured information.
- *Review*: Engineers review the lessons learnt and best practices that are allocated to them and take the captured information into account during their tasks.
- *Iteration*: The above process is iterated throughout each key phases of the project.

Lessons Learnt (LL) documents are created to capture lessons learnt from handling problematic issues in previous activities. They allow engineers to document the context of such issues, describe specific solutions applied for the concerned issue, reflect on experience generated from handling such issues and record any update to related best practices due to this experience. Best Practices (BP) documents are created to summarize best practices for performing various technical tasks. Each BP typically contains visual information (i.e. photos, drawings, CAD models) about the related component, and general guidelines in the designing, annotating, testing and handling such component. At the time of this case study, there were a total of about 900 LL and 70 BP.

Each LL is created and organized by the following process at each stage of the information lifecycle:

- *Creation*: Each LL documents is created by the owner of a knowledge stream in Microsoft Word format. Each knowledge stream represents a key knowledge category of the fuel system domain. Each knowledge stream owner is an expert in the respective knowledge category.
- *Storage*: Each LL document is sorted and stored within a corporate shared drive following allocation guidelines agreed by domain experts.
- *Retrieval*: The LL document is available for retrieval by 1) directly accessing the stored location, or 2) via a summary Excel spreadsheet that contains metadata for each LL in tabular format and a hyperlink to the stored location.
- *Maintenance:* The knowledge stream owners are responsible for the maintenance of each LL document. They also collaboratively maintain the lesson learnt spreadsheet.

The practices of handling BPs are similar to that of LLs, with the following exceptions:

- *Storage*: The BP documents are stored in a separated location to that of LL documents.
- *Retrieval*: The BP documents are available for retrieval via a departmental web portal which maps the shared drive folder/file structure and provides a list of documents of selected folder.

3.4 Summary of Case Studies

The main objective of using past case information is to speed up day-to-day engineering repair tasks. The PC documents also provide the added bonus that novice engineers can learn quickly by following previous examples. Each PC document essentially aims to provide a “complete picture” of the handling of a past case, and is created via well thought through process of taking information from a wide range of data sources and carefully assembling different information pieces together. Lessons learnt and best practices are mainly used for periodic review for overall performance improvement. Their usages are more focused on knowledge acquisition than actual day-to-day task handling. As well as targeting engineers within the FS group, they are also shared with other communities within the organization.

In the ISS department, throughout the information life cycle, the creation, storage, retrieval and maintenance of past case information is closely supervised by a domain expert dedicated to the process. This is opposed to the collaborate manner adopted by the FS knowledge stream owners. The reason is that the knowledge domains related to fuel system cover a much wider range than that of in-service. Another factor is that the volume of past case information requires a tighter supervision scheme for information management. Due to the shear amount of past case information, the folder/file structures used to allocate PC documents are significantly more complex than that of FS for LL and BS documents. For ISS engineers direct access via corporate file/folders system is not viable – engineers would have to manually navigate large number of folders and subfolders to get to any given PC document.

Both groups have adopted similar approaches to organize their respective unstructured information:

- *Usage of a structural layer to support retrieval*: Both groups dedicate significant *manual* effort to set up a retrieval facility in the form of Excel. This essentially introduces a layer of structural metadata information to support the retrieval activities.

- *Usage of allocation guideline for information organization:* Both groups follow a strict allocation guideline to decide the location of each document in the folder/file structure.

3.5 Key challenges

It is clear from analysing both case studies, that their practices in handling unstructured information serve basic organizing and retrieval needs for their respective set of information. However, discussions with ISS and FS engineers have highlighted the following ongoing challenges as below:

Challenges for ISS:

C1: *Inappropriate retrieval facility*: There is a general admission that Excel is not a tool suitable for concurrent usage and comprehensive retrieval. Excel built-in search tools only allow for ad-hoc filtering, offer very little support for advanced search and search results can only be presented with ranking on pre-selected metadata.

C2: *Incomplete information retrieval*: The search routine does not take account of what is available within PC documents such as photos and key document references. After initial searching via the spreadsheet, engineers need to open each potential PC document to perform detailed study on the usability of the case. Such documents often amount to large number of pages, and currently there is no facility to assist detailed examination. However, traditional search tools such as search portals often lack the capability for engineers to describe their information according to their specific need, and are often not as readily available as tools like Excel.

C3: *Inconsistent past case usage*: The existing retrieval practice often leads to inconsistent past case usage for similar cases. There is no capability to recognize the synonyms and acronyms that refer to the same concept so that similar search terms often yield different past cases;

C4: *Loss of historical context*: It is also considered that past case search in this manner leads to "Loss of historical context". For example, although PC documents present a complete picture of how a repair case was handled, the information captured within the documents may be out of date and is no longer feasible for current engineering practices.

Challenges for FS:

C5: *Inappropriate retrieval facility*: FS engineers also expressed concerns that Excel is not an appropriate tool for information retrieval. It has been indicated that, as the number of lessons learnt increase, retrieval has become increasingly difficult. The filtering mechanism of Excel often requires users to have well defined ideas about what they are looking for before performing search. This is counter-effective to lessons learnt searching which is often a knowledge discovery activity. It was suggested that better navigation between related groups of LL documents are much desired for such learning purpose.

C6: *Lack of connection between related information*: Lessons learnt and best practices are often inter-connected, for example, reviewing of lessons learnt often leads to creation of new best practices or updating on existing ones. During usage of such information, it is often beneficial to take this inter-relationship into account. However the storing and retrieval of these two types of information are separated, and the recognition of inter-relationships between lessons learnt and best practices currently reply on the memory of knowledge-stream owners.

C7: *Lack of accessibility control*: Lessons learnt and best practices are knowledge intensive engineering information. To ensure authority of such information, rigorous controls are often required during the creation workflow and also during the reuse cycle. Different accessibility to the same document is often required for different engineers (e.g. knowledge stream owner vs. normal engineer) at different stages (e.g. proposal vs. approval vs. reuse). The current IT facilities lack intelligent control over such variety in accessibility.

With the information challenges identified for both the ISS and FS departments, the next phase of the work was to investigate how a potential intelligent system can help to address some of these challenges, and evaluate its effectiveness. This is discussed in section 4.

4. DISCUSSION AND ONGOING RESEARCH

Despite the different nature of the information and varying requirements of usage, engineers from both ISS and FS expressed that *more intelligent* information retrieval is a primary concern for both processes. It is thus necessary to establish what the requirements are and then to translate these requirements into an overall context enabled system structure or search platform. This platform has been created and an initial evaluation of its elements is presented here.

4.1 Contextual Requirements

Considering the challenges articulated in Section 3.5 it is possible to draw the following conclusions. Challenge C1 indicates that a better information retrieval facility would be appreciated by engineering communities, however compared to traditional search tools, such a retrieval facility also need to offer greater flexibility for users to describe or articulate their information request. C2 reflects the need for better ways to capture information with high "engineering value" such as key reference material and visual information. C3 reflects the need to take account of engineering domain semantics for more intelligent retrieval. C4 reflects the need to take historical context into account in a situation where the value of specific information may change across time. Similar to that of C1, C5 reflects the requirement for a better retrieval facility, this challenge also stresses the need to provide better navigation ability across related information categories (similar to C6). C7 indicates the need for an intelligent system to take user's identification and actual tasks into account during information usage.

Therefore based on the two case studies, for an intelligent system that apply context to facilitate better organization and retrieval of unstructured information, it is possible to establish the following initial set of requirements, which are then used to drive the design of the Context Based Search Platform (CSP) in the next section.

- **R1 (C1, C5):** Flexibility for the user to describe information.
- **R2 (C2):** Capture "high value" information such as key reference and visual information.
- **R3 (C3):** Take domain specific semantics into account.
- **R4 (C4):** Understand change of information value across time.
- **R5 (C5, C6):** Efficient navigation between related sets of information.
- **R6 (C7):** Take user identification and task into account during information usage.

4.2 A Context Based Search Platform (CSP):

Based on the identified above requirements it is possible to create a system to understand how different techniques can collectively contribute to the organization of unstructured information. Thus an experimental information retrieval platform - Context Based Search Platform (CSP) has been created based on the enterprise search software Vivisimo [31]. The CSP performs full-text information search within folder/file structures, and also utilizes a number of techniques that were identified during the literature review to capture and exploit information context. The approaches have to deal with legacy data and new data that, in the case of the service example, is expanding at the rate of approximately 3000 cases per year.

It is not the purpose of this paper to describe in detail how the Context Based Search Platform (CSP) works, but to describe in broad terms how a context based dimension can be implemented to address the requirements highlighted from the case studies.

CSP specifically tries to address requirements R1, R2, R3 and R5. For R1, CSP allows users to create *social metadata* such as public and private tags and comments on search results. Meanwhile, *metadata mining* is implemented to extract metadata that has been recorded in the PC and LL spreadsheets. For R2, *Pattern recognition* routines are implemented to extract key information such as references to industrial standard documents and production drawings. This extracted information then serves as metadata to enhance information retrieval. For R3, semantic models of domain specific context models are created, these are then used to facilitate *semantic query expansion* by suggesting concepts that are related to input search terms. Four types of semantic relationships are captured: equivalent terms (synonym, acronym and spelling variation), broader terms (parent, super-class and upper level concepts), narrower terms (child, sub-class and lower level concepts) and related terms (relationships between concepts of different types). For R5, *Dynamic clustering* is applied to provide dynamic grouping for the most relevant search results of each search query. Figure 1 illustrates key components of the architecture of the CSP and its key functionalities.

Currently, information for 5000 past cases and 900 lessons learnt have been incorporated into CSP, resulting in two demonstrators, one for each case study. The implementation for two demonstrators will enable better evaluation of the varying techniques that are being proposed. For the ISS demonstrator, all the illustrated functionalities have been implemented. For the FS demonstrator, search results clustering, social metadata and metadata mining have been implemented.

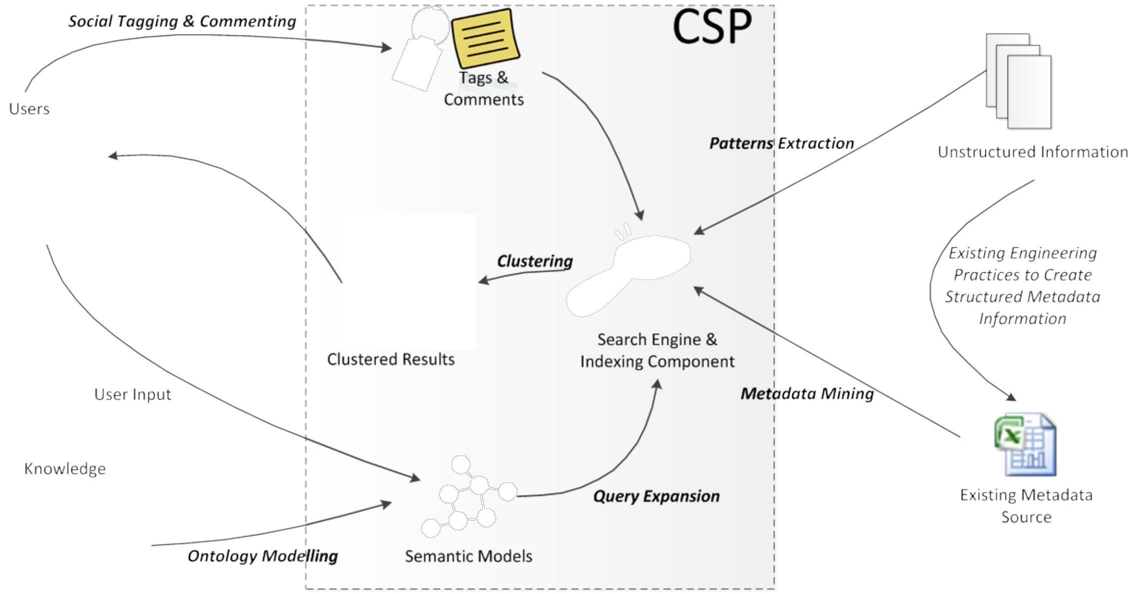

Figure 1: Architecture of Context Search Platform

4.3 Initial Evaluation of Benefits and Future Research Topics

Initial user evaluation has been performed for each demonstrator by inviting engineers to perform typical information retrieval tasks. Semi-structured interviews were then conducted to ask the engineers to evaluate the potential benefit of each technique and further research direction. The answers are summarised below:

- *Dynamic clustering* helped to provide interesting grouping of results particularly for large information volume such as ISS past cases. Engineers found this to be useful for discovering subtle relationships such as product structure and likely damage type. However it was also pointed out that when information was required to be classified in specified manners, such clustering approach often produced clusters of results inappropriate for search purpose. Most engineers interviewed didn’t have clear understanding of the underlying mechanism of such advanced data mining technique and found it complicate to configure clustering behaviour. It was suggested that static grouping technique such as classification with specified taxonomy may be more effective for such situation.
- *Social Metadata* allowed users to collaboratively create metadata on shared content, and therefore allow knowledge information such as past cases or lessons learnt to be shared. Engineers mentioned that this was useful for indicating information with high re-use value which would be otherwise difficult to identify via automatic methods. For instance, ISS engineers highlighted the possibility to indicate severity of corrosion cases which was interdependently influenced by factors such as damage level, location and repair history. However it was suggested that, at least for certain information, only selected experts should be allowed to create public social metadata. How to establish such privilege is a matter that will require further investigation.
- *Metadata Mining* extracted metadata information from existing structured metadata source. This approach allowed users to search for information via the search engine and utilised metadata created in Excel spreadsheet tools. Feedbacks from the engineers for this approach were mostly positive. On one hand, flexibility is retained for users to organise and describe day-to-day information as they see fit. On the other hand, the purpose-build search engine allowed much better utilisation of existing metadata compared to tools such as Excel. This was therefore perceived by the engineers to be a promising approach for organizing unstructured information. A pre-condition for this is that engineering departments need to dedicate significant resource to setup such structured metadata source. Further investigation is needed to investigate how common is this practice within the organization.

- *Semantic Query Expansion* driven by *semantic models* helped users to develop their search strategy in a structured manner. This was achieved by expanding the initial search criteria with semantic related concepts. Engineers expressed that this helps to provide more consistent results since users were prompted with query suggestions. In another hand, search queries were automatically expanded to include equivalent terms and therefore on average more results were yielded per search. However, the process to create useful ontology models was a labour intensive process requiring heavy collaboration between researchers and domain experts. Such process was perceived to be too expensive to be implemented in departmental levels. Also the process to validate and deploy such ontology models is a lengthy process under the current organizational IT framework.
- *Pattern Extraction* helped to extract useful information such as key references from unstructured documents. Engineers expressed that the extracted information helps to establish a "quick overview" on past case information. For instance, when references to other previous cases or standard procedures are extracted from a PC document, the engineer can quickly establish links to other related documents without examining the PC document in details. However they also pointed out that the extracted information alone does not provide enough information for effective past case selection, and is better provided along side with existing metadata source utilized by the metadata mining approach. In addition, the current process is an ad-hoc exercise tailored to the needs of the case study groups. It remains to be seen how such routine can be adopted to a bigger scale within the organization. The challenge is to how to have a generic approach integrated with an enterprise platform, but yet allow engineers to specify what is to be extracted.

This is the initial set of a subjective evaluation of the systems. It is further planned to run a series of scenario based benchmark assessments. These will be reported in subsequent publications. However it was clear from the evaluation that the addition of some form of context was of benefit and was appreciated and, interestingly, understood by the engineers who were part of this evaluation.

5. CONCLUSIONS

It has been discussed and shown from the literature that a large portion of information created in business activity is unstructured with much useful information buried inside. Within the aerospace industry, an added challenge is that the long product lifecycle requires such unstructured information to be accessible over a long period of time. The access to information and knowledge within these data sets is vital to both day-to-day engineering tasks and the improvement of engineering practices. It is this combination of unstructured information and longevity that is driving the requirement to incorporate some contextual elements into the data sets.

In this paper, the authors have examined current research that aims to enhance the organization of unstructured information from three contextual aspects: *information semantics*, *information usage* and *information content*¸ and identified promising techniques that contribute to each aspect. Two industrial case studies were then conducted to examine current practices of organizing unstructured information in respective engineering departments with particular focus on how such information is created, stored, reused and maintained.

As a result, key challenges in organizing and dealing with the unstructured information *elements* within an engineering setting have been identified. These are developed into a set of requirements of a desired intelligent context enabled system. An example Context Based Search Platform (CSP) has been described. An initial evaluation of its attributes has been given. The combination of a variety of techniques into a single platform, with the capability to support context based techniques seems to have considerable potential for dealing with the multi-dimensional and complex data sets that all engineers have to deal with.

REFERENCES

[1] Korth H.F. and Silberschatz A. Database Research Faces the Information Explosion. *COMMUNICATIONS- ACM , Vol 40, pages 139-142, 1997*

[2] Sweeney L. Information Explosion. Confidentiality, Disclosure, and Data Access: Theory and Practical. *Applications for Statistical Agencies*, 2001.

[3] LLic A. Burbridge T. Soppera A. Michahelles F. *A Treat Model Analysis of EPC-based Information Sharing Networks. Building Radio frequency Identification for the Global Environment,* 2007.

[4] Campbell D. *Approaches for the digital profiling of activities and their application. PhD thesis, chapter 6, page 12,* 2006

[5] Parameswaran M. and Whinston A.B. Social computing: an overview. *Communications of the Association for Information Systems, Vol 19,pages 762-780*, 2007

[6] Millen D. Feinberg J. Kerr B. Social bookmarking in the enterprise. *ACM Queue, Vol 3, Issue 9, pages 28-35, 2005*

[7] Ferrucci D. and Lally A. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, Vol 10, Issue 3-4, 2004*

[8] Unstructured data. *Wikipedia. http://en.wikipedia.org/wiki/Unstructured_data. Taken at Jan 14, 2011*

[9] Shilakes C. and Tylman J. *Enterprise information portal, Merrill Lynch, 1998*

[10] Grimes S. *Unstructured data and the 80 percent rule. Clarabridge Bridgepoints, 2008*

[11] Bhogal J. Macfarlane A. Smith P. *A review of ontology based query expansion, Information Processing and Management, Vol 43, issue 4, pages 866-886, 2007*

[12] Z_punkt and Accenture. *Information 2015-Reforming the paradigm, http://www.accenture.com/Global/Technology/Information_Mgmt/Information_Mgmt_Services/R_and_I/Information-2015-Reforming.htm, 2010*

[13] Clark W. *Context change everything: What CIOs must know, Gartner, 2010*

[14] Semantics. (n.d.). *Collins English Dictionary - Complete & Unabridged 10th Edition*. Retrieved January 14, 2011 from Dictionary.com, http://dictionary.reference.com/browse/semantics

[15] Mangold C. A survey and classification of semantic search approaches. *Int.J.Metadata, Semantics and Ontology, Vol 2, No 1, 2007*

[16] Ding L. Kolari P. Ding Z. Avancha S. Finin T. Joshi A. *Using ontologies in the semantic web: a survey. Integrated Series in Information Systems, 2007, Vol 14, Part 1, pages 79-113*

[17] Value-It Project, Final demand driven mapping report, 2010, Retrieved January 05, 2011 from http://portal.value-it.isoco.net/c/document_library/get_file?folderId=13060&name=DLFE-1015.pdf

[18] Dey A.K. Abowd G.D. Towards a better understanding of context and context-awareness in *CHI'2000 Workshop on the What, Who, Where, When and How of Context-Awareness*, 2000

[19] Dey A.K. Understanding and using context, *Personal and Ubiquitous Computing, Vol.5, No1, 2001*

[20] Henricksen K. Indulska J. Modelling and using imperfect context information, *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Coommunications Workshops (PERCOMW'04),* 2004

[21] Wang X.H. Zhang D.Q. Gu T. Pung H.K. Ontology based context modeling and reasoning using OWL, *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW'04),* 2004

[22] Budzik J. and Hammond K. User interactions with everyday applications as context for just-in-time information access, in *Proceeding of the 5th International Conference on Intelligent User Interface,* 2000

[23] Rhodes B.J. Maes P. Just-in-time information retrieval agents, *IBM System Journal, Vol 39, Issue 3.4, pages 685-704, 2000*

[24] Redon R. Larsson A. Leblond R. Longueville B. VIVACE context based search platform, in *Proceedings of CONTEXT'07, Roskilde, DK*, 2007

[25] Hong J. Suh E. Kim S. Context-aware systems: A literature review and classification, *Expert Systems with Application, Vol 36, pages 8509-8522, 2009*

[26] Baldauf M. *A survey on context-aware systems, Int.J.Ad Hoc and Ubiquitous Computing, Vol. 2, No. 4, 2007*

[27] Han J. and Kamber M. *Data Mining Concepts and Techniques,* Morgan Kaufmann Publisher, 2006.

[28] McAfee A. *Enterprise 2.0: New Collaborative Tools for Your Organizations Toughest Challenges.* Boston: Mcgraw-Hill Professional, 2009

[29] Goh Y. M. Giess M. Stewart D. McMahon C. Application of faceted classification on in-service records, *Academic Journal of Manaufacturing Engineering, Vol.6, No.1, 2008*

[30] Kim H.L. Passant A. Breslin J. G. Scerri S. Decker S. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces, The IEEE Internal Conference on Semantic Computing, Santa Clara, CA, 4-7 Aug, 2008

[31] Vivisimo, *http://vivisimo.com/, Taken at Jan 18, 2011*

Contact: Yifan Xie
University of Bath
Department of Mechanical Engineering
Bath, BA2 7AY
UK
Tel: Int +44 11793 64092
Email: Y.Xie@bath.ac.uk