# Facilitating Design Learning through Faceted Classification of In-Service Information

Yee Mey Goh*, Matt Giess and Chris McMahon

*** Contact Author:**
Dr YM Goh
Innovative Design and Manufacturing Research Centre (I*d*MRC)
Department of Mechanical Engineering
University of Bath
Bath, UK.
Telephone: +44(0)1225386131
Fax: +44(0)1225386928
Email: y.m.goh@bath.ac.uk

**Abstract**

The maintenance and service records collected and maintained by engineering companies are a useful resource for the ongoing support of products. Such records are typically semi-structured and contain key information such as a description of the issue and the product affected. It is suggested that further value can be realised from the collection of these records for indicating recurrent and systemic issues which may not have been apparent previously. This paper presents a faceted classification approach to organise the information collection that might enhance retrieval and also facilitate learning from in-service experiences. The faceted classification may help to expedite responses to urgent in-service issues as well as to allow for patterns and trends in the records to be analysed, either automatically using suitable data mining algorithms or by manually browsing the classification tree. The paper describes the application of the approach to aerospace in-service records, where the potential for knowledge discovery is demonstrated.

**Keywords**
Information organisation, faceted classification, in-service records, knowledge discovery

## 1 Introduction

Information captured from in-service phases represents a significant intellectual resource to engineering companies. Knowledge gained from in-service events may indicate how well the design of an engineering artefact satisfies performance, operational and maintenance requirements. Such knowledge can also be used to improve the planning and execution of maintenance, repair and upgrade activities. Doultsinou et al [1], for instance, identified several areas of service knowledge related to tooling, spares, serviceability, maintenance and training issues. The research reported in this paper is conducted within the "Through-Life Knowledge and Information Management" (KIM) project [2]. The project aims to establish a holistic framework for the design and use of information and knowledge-support systems within the new and evolving landscape of the long and total product-service life cycle. One key aspect in this is associated with how information about the product in use can be captured and fed back to support activities within the product-service enterprise.

The importance of engineering knowledge obtained as a result of learning from practice and experience is widely recognised [3-5]. Maidique [6] studied success factors in new product development of 158 electronics products in the USA and reported on the importance of learning from past experience that includes dimensions of learning by doing and using. Learning by doing (internal learning) and using (external learning) were introduced by Rosenberg [3] based on study of the aircraft industry. Two outcomes may arise from the learning by using process: embodied knowledge that results in design modifications for improved performance, and disembodied knowledge that results in improved operation of the original or the modified product. Embodied knowledge is knowledge that is present inside persons or manifested within the characteristics of the artefacts which they design. According to Rosenberg, learning by using usually results in disembodied learning, e.g. operational experience that results in change in maintenance practices. As a consequence of this, however, new information may be generated that eventually results in design modification (embodied knowledge).

Product-Service Systems (PSS), where the designing companies are also contractually involved in the support and sometimes operation of the product, present greater opportunities in terms of learning than traditional product support approaches. In the PSS context, it may also be argued that learning from use is even more critical to the success of the ongoing maintenance, product support and end-of-life activities and to the generation of knowledge which will then feed into the next design cycle. For instance, learning about the design can be about the use conditions assumed at the design stage, such as the operating loads or temperature variation, to allow better understanding of actual operating conditions and how the design performs under various conditions. Also, learning about the design can be associated with the engineering knowledge applied at the development stage, e.g. simulation technologies and test data, which can be updated to improve predictive capabilities of the tools and methods. Such continuous cyclical learning process will increase the chances of *design right first time*, prevalent key objective in agile manufacturing [7]. More directly, the knowledge gained will be useful in the decision-making involving operational and maintenance activities.

In practice, knowledge management can be achieved via two fundamentally different approaches: personalisation or codification [8]. The personalisation approaches

emphasise human resources and communication, such as the development of Communities of Practice [9]. Codification approaches on the other hand, place emphasis on the collection and organisation of knowledge into useful resources. The experience accumulated from in-service can be held by the service personnel, which can be transferred and shared through formal communication and informal conversations between employees. Alternatively, information resources that are generated from the in-service activities can be reused to support decision-making throughout the lifecycle of the product. Personalisation and codification approaches are not mutually exclusive but emphasis may be placed on different techniques to suit the type of engineering work (for instance, degree of design originality and complexity [10]). In this paper, it is argued that a codification centric approach well matched to to the dynamic and geographically distributed nature of the teams involved in long-lived PSS.

Whether or not the PSS model is fully adopted, companies often have to deal with technical issues arising from their products in use. Such events may be triggered for warranty claims, service call-outs, or through general feedback or surveys. It is common to expect documentary records associated with these events to be generated. More recently, state-of-the-art monitoring technologies have been used to deduce from collected data indications of a particular condition or anomaly in the product. The systems are used in condition-based maintenance to help predict and plan for operational and maintenance activities more effectively [11]. At present, although much information is being captured, the mechanism for retrieving and reusing in-service feedback for decision-making is still often found to be ad hoc and inconsistent. Information that is captured is stored in different formats, held in distributed systems, owned and maintained by different parties etc. As a consequence, many companies rely greatly upon personal and social processes to learn about their products. In order to build on these ad hoc resources, this paper reports on a codification approach based on an information classification technique to enhance retrieval of in-service information and to support greater learning from use. An application case study in an aerospace company is reported to demonstrate the approach.

## 2 Literature Review

Information reuse occurs when information is assimilated and used in a new application and also, when subject to some processing, it yields useful new insights and knowledge. Data loggers and monitoring systems are commonly available on high-value equipment to continuously log its performance data through life [11]. The data may be transmitted to a computer or Personal Digital Assistant (PDA) via infra-red or radio-frequency and interrogated using specialised analysis software. Many case studies have shown that such operational data might be used to monitor equipment health, thus enabling preventative and predictive measures to be deployed. Jardine et al. [11] provided a comprehensive review on machinery diagnostics and prognostics using condition-based monitoring. These decision-support systems allow the detection, isolation, identification and prediction of fault condition(s) based on data processing and analysis. Statistical, artificial intelligence, model-based and rule-based approaches may be employed for pattern recognition in condition-based monitoring systems. Sensor data such as loading, number of operational cycles and frequency of use can allow inference to be made with regards to its application and operating environment. The Distributed Aircraft Maintenance Environment (DAME)

is an e-Science project [12] demonstrating the use of GRID infrastructure to implement online health monitoring and fault diagnostic systems for the maintenance of aircraft engines in distributed environments [13]. General Motors (GM) OnStar Vehicle Diagnostics is a telematics service, for instance, that uses a combination of Global Positioning System (GPS) satellite and mobile phone network, along with onboard diagnostics, to link the vehicle and driver to the OnStar centre.

Knowledge gained through collective experiences of similar products used in similar or different environment may be incorporated into new design solutions. For instance, GM is planning to use the customer insights gained from the OnStar System to better design its future vehicles to meet customer needs [14]. A closed-loop condition monitoring system with feedback to the Original Equipment Manufacturer (OEM) that might affect the design is shown in Figure 1. The figure also illustrates the nature of such systems that involve numerous stakeholders and functional entities, which may act as barriers to knowledge transfer and sharing in traditional organisation and business arrangements. Although of significant importance, issues associated with these challenges will not be discussed in the current paper.

**Figure 1 Condition monitoring and feedback to Original Equipment Manufacturer (OEM), adapted from [15]**

Maintenance and service activities are frequently recorded and such records become a valuable resource to the company. These records may be generated using paper or electronic forms or document templates, which may mandate a specific form of entry for fields such as engineer and customer whilst allowing more subjective information such as issue resolution to remain free-text.

Increasingly, electronic and online reporting systems such as e-maintenance systems and Computerised Maintenance Management System (CMMS) are used with associated workflow management and event logging capabilities [16, 17]. For example, when a customer reports a problem through a phone call to a service centre, this issue is logged automatically and becomes a live issue that is managed electronically. These systems allow companies to manage and systematically respond to issues thus potentially improving customer relationships. Some systems also allow automatic user information capture, correspondence, reminders and archiving. The information within these systems may be structured[1] (with predefined fields such as date, issue, customer, location etc.), but it is fundamentally different to the numerical data provided by sensors.

Where monitoring systems may invoke some algorithmic method to model expected outcomes in-service, textual reports (particularly free-text) require some processing to achieve such ends. Research in Natural Language Processing (NLP) seeks to use computers to automatically understand human languages. To date, significant progress has been made in this area, particularly within the subjects of bioinformatics where samples of human language are converted into more formal representations that are easier for computer programs to manipulate.

[1] 'Structured' data or information is that which is represented electronically in a way that makes it interpretable to a computer, such that computational processes can be used to manipulate it when it is encountered.

A number of such approaches rely upon suitable identification of the role each term plays in speech. For example, Automatic Term Recognition (ATR), the identification of key terms within a corpus, generally relies upon both linguistic and statistical measures [18]. The linguistic measures are generally based the identification of various types of term (nouns, adjectives, prepositions, verbs etc), referred to as Part-of-Speech (PoS) tagging. By identifying these constituent elements, linguistic filters may be deployed which utilise these tags to construct candidate term sets which are then statistically analysed to deduce their relevance as key terms. The specific composition of these filters depend upon the detail of the corpus, in terms of language deployed [19], hence such approaches are most readily deployed in domains with a degree of consistency in the means of expression.

Text mining supports pattern-searching in natural language text, and may be defined as the process of analysing text to extract information from it for particular purposes [20]. Text mining, as a broader phrase covering activities such as information extraction and categorisation, describes the activity of extracting information from text algorithmically and discovering relationships from the extracted information in a similar manner to data mining. Commercial text mining software is widely available, such as SPSS Text Mining for Clementine (www.spss.com/) and IBM Intelligent Miner (www.ibm.com).

When textual information is recorded in a structured manner (where semantically meaningful tags are used), certain methods in statistical analysis and data mining can be employed to reveal patterns or correlations in the information that may not have been previously apparent. In particular, if a taxonomy is used, the information may be treated more like discrete data and therefore can be similarly processed. For example, reliability engineering is an area that uses the collection of field data from a population of ‘identical’ components to predict system reliability. A number of such databases are collected by the defence, process and offshore industries [21, 22]. Technique such as “Slice and Dice” analysis (a form of user-based dynamic manipulation and visualisation of data) [23] enable analysts to look for root cause and trend data across specific criteria such as product type, life cycle episodes (e.g. particular manufacturing batch points), application, or individual user. This may give the analyst further understanding of the nature of the reported events that may be characteristic of the environment in which the artefact is operating. By focusing on understanding such correlation, knowledge to enhance design decisions aimed at mitigating root causes of in-service issues can be actively acquired.

Case-Based Reasoning (CBR) is an artificial intelligence (AI) learning approach that represents knowledge (previous examples) as cases using some pre-defined key parameters. When presented with a new, incomplete example, CBR can retrieve similar previous cases by comparing the index parameters and infer the missing or incomplete information from these complete examples. CBR has been applied successfully for information reuse in diverse areas such as help desk automation and health monitoring [24, 25], where the domain is well understood and knowledge is relatively static. Whilst it may be effective for retrieving the most similar cases and reusing the cases to attempt to solve the problem, it is less useful in facilitating learning through experience.

Ontologies have been applied for improving information organisation and retrieval in various domains. An ontology, which is a specification of a conceptualization, defines all the entities (objects or concepts) that are of interest in a domain and the

relationships that connect these entities together, usually in some formal and preferably machine-readable manner [26]. An ontological approach has also been employed to organise service information to support the ability to make inference between design and in-service outcomes [27]. In this approach, concepts related to design and service engineers' knowledge, e.g. engine deterioration mechanisms, engine models and parts, airport locations etc. were derived from document analysis and knowledge acquisition interviews. The resulting ontology contains concepts ranging from engine failure mechanisms, engine models and parts to airport locations. A prototype system allows users to access previously heterogeneous document resources, integrated by a shared ontology via a web portal. The system allows users to search maintenance records through a taxonomy and to query the statistics. Ontological systems rely on key concepts and their relationships being identifiable from the documents.

# 3 Information Organisation

In order to infer how a product performs in service, for purposes of improving the product's fundamental design or of providing some temporary treatment of suboptimal performance, it is important that a practitioner can contrast performance of a product across numerous operating environments for a large enough sample of products for consistent patterns to emerge. For example, a certain fault or issue may occur only under certain usage conditions or in conjunction with the use of certain other components, which will not be apparent when considering one fault or product in isolation.

Records from service generally contain information describing the specific conditions under which a fault occurred, however where this information is obtained from multiple external sources (as is often the case for in-service queries) the specific terminology and form of the information is subject to significant variance. In such a situation it is necessary to abstract this information into broader, aggregated and consistent viewpoints such that emergent patterns across numerous records may be identified. The abstracted viewpoint may also serve as a browsable organisational scheme, such that past cases may be identified and retrieved and their resolutions either directly or indirectly reused when addressing similar issues. This work seeks to evolve such an organisational scheme from service records, and to identify how emergent issues may ultimately be identified by population and interrogation of this scheme.

## 3.1 The Form of Abstraction – the Organisational Scheme

The intent of an organisational scheme is to provide a means of collocating or consistently describing records whose contents share certain commonalities. By categorising relevant items of information from within a record (for example, identifying the specific component of interest) it becomes possible to prescribe specific universal categories for application across all records. This allows the contents of each record to be treated identically, regardless of the form and terminology deployed. By arranging such categories within a broader classification scheme, indicating how the categorisations are related, a browsable structure may be generated to facilitate retrieval of past cases. Alongside such organisation, the assignation of consistent categorisations (and their arrangement into aggregated categories at higher levels of the classification scheme) also allows analysis to be conducted to deduce common patterns across the assigned categorisations.

Information classification involves the development and deployment of schemes for the systematic organisation of knowledge (represented as information), whose approaches were founded within the field of Library and Information Science (for example Foskett [28], Rowley and Farrow [29], Taylor [30]. Within this field, three distinct forms of classification may be seen [30], the enumerative, the synthetic and the faceted.

Enumerative classification involves the recursive partitioning of an information corpus into progressively smaller subsets, with the aim of constructing subsets whose content is identified as consistent enough for the purposes of the classification scheme. The criteria used to partition are referred to as principles of division, whose selection is generally steered by the purpose of the scheme. As many of the principles are commonly deployed, it is possible to derive standard sets of principles which promote consistency across the scheme – for example, a scheme would be confusing if it categorised people by gender, but at the next hierarchical stage classified men by age and women by nationality. By promoting a standard set of principles, to be deployed in a consistent manner, more comprehensible schemes may result. This is basis of synthetic classification, where auxiliary tables have been generated which pre-define the principles of division.

Faceted classification takes this idea of consistency to arguably its logical conclusion, identifying that some ‘dimensions’ or facets (not entirely equivalent to principles of division) are so inherent that they may be treated separately. For example, in a wine cellar each bottle of wine will be the product of one or more grapes, so by treating the dimension or facet ‘grape’ separately, and combing with other facets (such as ‘vintage’ and ‘region’ for example) to provide a complete categorisation, a robust, consistent but versatile classification may result. Users are not constrained by the order in which the principles are deployed as per the enumerative classification, instead they may interrogate only those facets of interest.

### 3.1.1 Compound Concepts in Classification

The selection of a suitable principle of division necessarily introduces some viewpoint-dependency into an enumerative scheme, as certain principles may be of greater or lesser utility to a given group of users. Synthetic schemes provide a common form of decomposition but do not address this issue of viewpoint dependency.

A faceted scheme, however, allows the various dimensions to be treated separately, such that a user may place greater or lesser emphasis upon a given categorisation within any given facet. This is of great importance when considering the multi-dimensional nature of many information resources, which may describe compound subjects. For example, treating resources that describe the ‘structural analysis of titanium strut for aircraft XX’ within an enumerative schedule requires that precedence be given to either the component, material, method of analysis or aircraft, whereas in a faceted schedule each may be treated separately. The ultimate aim of the work reported here is to illustrate to the user, either visually or through some automated pattern identification, the prevalence of certain concepts in combination. Treating each concept separately allows the user to assemble a compound concept which is of interest at the point of browsing, and by allowing users to rapidly change their perspective in terms of the specific combinations of concepts (the idea of a dynamic taxonomy) such that the population or volume of records or documents

returned can be assessed with each change in concept. Thus, the user may retrieve a useable volume of material by searching at various levels of granularity, whereby if too great a volume is returned the granularity of a given concept may be increased or further concepts may be added to the compound to increase discrimination of retrieval. Further to this, the user may make some assessment of which combinations are sparsely populated and which are heavily populated, which in itself provides insight into the domain.

### 3.1.2 The Theory of Faceted Classification

Although not the inventor of facet analysis, Ranganathan is credited as the first to systematise and formalise the theory [28, 29, 31]. According to [30], facets are clearly defined, mutually exclusive, and collectively exhaustive aspects, properties or characteristics of a class or specific subject. Ranganthans's early work was revisited in later years by the Classification Research Group, who proposed certain amendments to Ranganathan's work. A key departure is within the work dealing with Fundamental Categories, a key part of Ranganathan's work which identified five fundamental dimensions as the basis for facets. The CRG suggested that any attempt to impose such fundamental categories mechanically should be resisted [32], as each scheme and domain must be tackled on its own merit.

The published literature describing the construction of faceted schemes tends to be predominantly evaluative, in that is provides a means of assessing a constructed scheme, as opposed to generative [33], although high-level procedural guidance is in evidence. For example, Facet Analysis, applied to the presentation and physical layout of *thesauri*, involves [29]:

- Identifying sets of terms representing concepts. This involves the identification of multiple 'simple' concepts from the compound subjects (s) that describe a document.
- The grouping of the terms representing the simple concepts into a number of mutually exclusive categories (called facets).
- Organising the facets into a limited number of fundamental categories – these fundamental categories can be viewed a being different types of classification schemes. Thus the process of organising facets is essentially analogous to a process of classification scheme construction.

This general approach was followed within this work and will be discussed in later sections.

## 3.2 Waypoint

Although faceted classification was originally intended to assist in the classification of tangible entities such as books, it is of great use when used to electronically treat information as the browsable structure and its content may be dynamically adapted to meet any user viewpoint. A faceted classification system, Waypoint, has been developed for this purpose [34]. This system allows users to browse through any combination of pre-defined facets and retrieve information that meets all selections made across all facets.

A potential weakness of such a system is in the ability to select empty sets, essentially combinations of categories that are infeasible in practice [34]. The Waypoint system populates a pre-defined faceted classification according to a series of logical rules (which are described in section 3.2.2). A user may browse any given facet of interest,

from which they may select a category of interest. All documents that are not relevant to this category are removed from the available set, such that if this process is repeated within a separate facet the remaining document set is the intersection of sets returned in each separate selection. The total number of documents relevant to each category at each point in the selection process is indicated, such that a user may have some indication of the discrimination of search but also such that the user may identify densely populated combinations of categories. It is this facility that is of interest in learning from past cases as described in this paper.

### 3.2.1 The Treatment of Documents within Waypoint

Waypoint acts as a 'stand-off' system which allows a user to browse distributed document corpora through a single interface and retrieve documentation in its native format. Where Document Management Systems generally host documents, and thus may mandate the manual assignation of appropriate categorisations to each document at the point of upload, the stand-off nature of Waypoint requires an automated method of document classification be deployed.

The automated classification is conducted via the use of constraints - syntactic rules that identify key terms within a document which indicate its relevance to a category within the classification scheme (Figure 2). Other approaches may be deployed in this automated classification, for example NLP (discussed in 2) and Machine Learning [35] which includes techniques such as CBR (also as discussed in section 2).

This provides a potential avenue for further work, however for the purposes of this research constraints provide a useful approach as they may be generated (namely, the specific syntax for each rule may be specified) at the same time as the construction of the faceted scheme itself. As the faceted scheme is constructed by arranging specific concepts identified within the corpus into a logical scheme, the terms from which the concepts were identified (the concepts may differ slightly from the terms themselves as they are collated and abstracted) may be immediately entered as constraints for that concept. A hypothetical example of such a constraint is given below.

IF document contains Term 'Landing Gear' AND Term 'Bush' OR Term 'product SN XXXXX' THEN document is relevant to concept 'Landing Gear Bush'

**Figure 2 Faceted classification associating documents to categories within each facet [36]**

These constraints may be constructed for each *concept* within the *concept map* (recorded as a combined XML-based map indicating the faceted scheme and each constraint), and is typically carried out concurrently with the construction of the faceted scheme – the scheme is itself a distillation of terms that appear in the documents (as concepts). This construction does not have to be conducted for all documents, a subset of documents may be interrogated for this purpose (called the training set). Other established techniques exist, for example it is possible to automatically parse the documents to extract all significant terms (for example, by comparing the prevalence of certain terms within a document as compared to the prevalence within the overall corpus, referred to as "term frequency – inverse

document frequency" [37, 38]) and to organise those key terms to form concepts. Methods for assisting concept identification will be discussed for the case study in later sections although a controlled vocabulary can greatly assist in this task.

### 3.2.2 Exploiting Document Structure

'Structured' data or information is that which is represented electronically in a way that makes it interpretable to a computer, such that a computer can understand what to do with it when it is encountered. Structured data can be found in such things as relational databases and in files which are tab or location delineated. Here the data are organised into a data structure according to the relationships and data type definitions prescribed by a data model. Unfortunately, much of the information generated in the course of the engineering process (in such things as reports, communications, procedures, catalogues, etc) is not only unstructured – making it difficult for electronic information systems to handle – but by its nature difficult to structure in its entirety. Such information often consists of incomplete or changeable content, constructed of text as well as visual elements such as tables, drawings and graphs.

Nevertheless, it is often possible to bring some order and machine-interpretability to unstructured information by making it semi-structured. This is done conventionally using such things as headings, paragraphs and sections (physical structure) and more recently by tagging or marking up interesting elements of the content explicitly using purpose-built formal languages such as HTML or XML [39]. In structured data, the data are organised according to some external, pre-specified data model. In unstructured information, it may be possible to identify commonly occurring information elements, and develop a scheme or model by which these information elements in a document can be partially structured (and indeed, new documents constructed). Particular useful are the semantic (that is, meaning-bearing) information-bearing elements of content. One document may have many semantic dimensions, each of which is represented by a different structure.

Revealing the semantic content of information allows information search based not on conventional pattern-matching techniques, but on meaning of the content. The process of making semantic structure explicit by marking up information content effectively means that documents can be decomposed into smaller and meaningful chunks. Waypoint specifically makes use of the explicit structure of semi-structured, marked-up documents when making an association between a document and a concept. This is done by specifying the part of structure to which the constraints apply. Figure 3 illustrates the following rule in associating a structured document to a *concept node* in the *concept map*.

Document is relevant to concept 'Damage' IF AND ONLY IF document contains Term 'Damage' within 'Description' field

**Figure 3 Association of a structured document to a node in the concept map**

## 4 Case Study

The aircraft manufacturer's In-Service Support (ISS) teams maintain databases of queries of a non-trivial nature from operators for their aircraft. Each incoming query into the central organisation is given a unique identification (ID) number before the

case is passed on to the team responsible for the relevant part of the aircraft. Within these teams, several distributed resources of database and records management systems may be used to support various activities (design, analysis) that are carried out by the ISS teams. These systems usually contain repetitive and additional information related to a particular In-Service Query (ISQ). The unique ISQ number and other means of cross-referencing may be used to trace instances across systems and files. Databases of the ISQ queries from operators that are assigned to the ISS team responsible for issues regarding the "Fuel Systems and Landing Gear" and the "Wing Structures" were available in this study.

The ISQ databases contain entries collated from years of in-service experiences associated with a number of aircraft variants. The databases contain standard fields such as the unique ISQ number, airline, aircraft type, ATA chapter (an industry standard, high-level decomposition of aircraft systems), manufacturer's serial number for the aircraft, dates, engineer responsible, urgency category, case status, description of the issue, keywords and conclusion. Information for keyword and conclusion is not always included but more likely to be included in recent records. Typically, the ISQ records are generated after the in-service issues have been resolved and most of the records are also linked to the original correspondence generated between various stakeholders such as the aircraft manufacturer, component manufacturer and the airline representative during the resolution of those events. Often, the linked documents are in the form of unstructured reports and e-mails scanned and stored in PDF format (for legal reasons). The ISS engineers mainly rely on field filters and a keyword search facility to retrieve past records that help them resolve new issues. The keyword search may be effective when the user is familiar with the vocabulary used in a particular domain. Although it may be more efficient to domain experts, such system is not particularly useful to new users, nor is it useful in highlighting classes of systemic in-service issues that may warrant further investigation for continuous improvement.

The objectives for classifying the ISQ records using the Waypoint faceted classification system were twofold. Firstly, the classification should enhance the retrieval of successful past resolutions when a new query is raised; secondly, treating each instance in a consistent manner in order to reveal issues that are frequent and enduring enough to suggest remedial work would be beneficial. Not only should this evidence allow the identification of design-induced issues, where consistent faults are seeen with given components or systems, but also provide some understanding of the operating environments in which they occur as these may be significantly different from those assumed in design.

In this manner, engineers can prioritise their efforts in identifying root-causes such that they can actively learn from the accumulated in-service experiences. The next sub-sections describe the construction of classification schemes in the case study and how Waypoint can be used for retrieval and patterns analysis. Methods considered for improving/automating some of the work are then discussed.

### 4.1 Constructing the Classification Schemes

The construction of a classification scheme is a far from trivial task, requiring significant intellectual effort. It is possible to construct a classification scheme in a number of ways, for example to decompose a domain according either to the requirements of users or an expert opinion, although such a top-down approach presupposes the practitioner's view of the world is both accurate and consistent.

The Library and Information Science approach to the construction of a faceted classification, as discussed in section 3.1, may be considered bottom-up in nature in that it focuses on extracting the content of a domain (in terms of constituent concepts) from the underlying documents. This depends upon adequate possession of documentation, and an accurate distillation of each document to provide a suitable scheme. In practice, the construction of classification schemes may use both top-down and bottom-up approaches, for example a top-down view of the scheme may steer the practitioner to identify certain forms of concept from each document at the expense of others.

The method of identifying key concepts, terms and relationships within documentation and structuring these into a classification scheme can also be influenced to some extent by *warrant* [40], used here in the sense of justification for the choices in the classification. Warrant may be considered simply as the authority by which the specific categorisations and their arrangement into a coherent scheme were arrived at, whether that be by the opinion of expert users or by some measure of the information corpus itself. Many types of warrant have been suggested, including user warrant, scientific warrant, educational warrant and cultural warrant. For example, domain experts could create classification schemes based upon the *scientific warrant* whereas the *user warrant* aims at supporting the end user by identifying the terms and structures that would be of greatest utility to a given set of users. *Literary warrant* describes the practice of constructing a classification scheme based upon the specific content of literature [41]. In terms of applicability, analytico-synthetic schemes generally rely upon literary warrant where the concepts which are contained within the document corpus are identified beforehand and the scheme arranged to fit these concepts. In this case study, the faceted schemes were constructed bottom-up using this approach, and were then compared, refined and verified by domain experts.

### 4.1.1 Constructing the Classification Scheme within Waypoint

From the ISQ database with many thousands of instances, the records for a selected number of years were scripted into structured XML documents with each piece of information tagged using a scheme based on the fields in the ISQ records (shown in Figure 3 “Document”). An XML document can be viewed as the meta-information about an ISQ instance, where hyperlinks to the original documents such as reports, photographs and e-mails may be embedded for future retrieval. As mentioned previously, the structure improves the effectiveness in the Waypoint classification, such that rules can be coded under each tag/field. Some of these tags were inherited from the current ISQ database, particularly those referring to the airline, aircraft type and the ATA chapters which naturally become facets by which to organise the documents. These facets however, mainly relate contextually to the ISQ instance (except the ATA chapters which reflect the subsystem and functions to a limited extent). They do not sufficiently describe the technical detail of the issues. Some fields may contain numerical information such as the dates, the flight cycles and hours, which can be classified according to ranges specified in the Waypoint *concept map*, using the flexibility provided by the use of XQuery [42] as a technology for querying the underpinning Lucene index [43] upon which Waypoint is constructed. Nevertheless, information was found in the brief descriptions (typically expressed in terse, perfunctory free-text) which proved useful in deriving additional facets by which to classify the documents. The descriptions provide content-based information to the issue such as the failure mode, assembly, operational phase, flight type and

topological location on the aircraft[2] (RH/LH, rib #, leading/trailing edge). From a relatively small subset of the ISQ descriptions (100 documents), a number of concepts were manually distilled, aggregated using domain knowledge, and organised into different facets summarised in Table 1. It is argued that these additional facets are of great importance in learning the patterns across issues as they describe the specific detail of each issue, which when taken in conjunction with the more contextual facets provide a complete description of each issue.

**Table 1: Concepts distilled manually from a subset of the "description" of Fuel Systems and Landing Gear ISQ records**

Having derived the classification schemes, the next step requires the constraints to be coded for each node in the *concept map*. It is this step which determines how documents should be associated with the classification scheme. The following syntax is used to associate documents with the term "fuel leak" in its **description** field to the classification node (concept) labelled as Fuel Leak (refer to Figure 3 concept map and document syntax). The use of XML as a representation provides human-interpretability, allowing users to readily comprehend the structure and content of the classification. It also becomes possible to provide schemata which will allow the concept map to be treated in alternative environments, for example in order to edit the scheme, which is one of the inherent strengths of XML approaches.

```
<concept id="operation_3140">
    <desc>Fuel Leak</desc>
    <term type="0">
      <search>description:"fuel leak"</search>
    </term>
</concept>
```

The Waypoint system includes such functionality by incorporating the Open-Source Lucene indexing and search engine library [43]. This library has a number of different document analysers which cater for different forms of treatment including stemming. In the example above, the double quotes ensure the phrase 'fuel leak' is present in the document description field whereas without such quotes will associate document with either one of the terms to the node. In such a case, documents containing any combination of the stemmed words of 'fuel' and 'leak' will be returned (e.g. fuel leaks, fuelling leak, fuel leaking) under this node. Indeed, free-text description allows for spelling variations, synonyms and abbreviations that simple stemming procedures do not resolve, and which occur outside of the sample set of documents used in the construction of the classification scheme and constraints are thus not catered for. In addition, common concepts such as a component or a failure mode are not apparent if they are being inconsistently referenced in the description. Text analysis and mining approaches are being considered to improve and automate the identification of concepts and their constraints and are discussed in the following section. By performing the bottom-up approach a number of additional facets were identified from the underlying document corpus that are consistently used by the ISS engineers.

[2] This information is more pertinent in the Wing Structures ISQ.

In the future, this information may be requested during reporting to enable an effective classification of new instances avoiding records with missing vital information.

### 4.1.2 Automating the Concept Identification

Methods generally classified under text mining can help towards automating the identification of concepts from the descriptions and the generation of rules/constraints for those concepts [37, 38]. Text mining is useful as the classification is constructed using literary warrant, i.e. evolved from the underlying content of corpora. The identification of concepts for classifying the ISQ records was originally derived manually by going through a sub-set of records. As the domain is quite constrained, many concepts are repeated. Although methods proposed in NLP such as PoS tagging appear to be relevant in the first instance, they are found to be less useful in dealing with technical reporting which does not conform to standard grammatical rules, although progress has been made in this regard, for example through the approach for handling sublanguages described in [44].

As shown in Table 1, the concepts need to be grouped into classes to provide a logical classification structure (either enumerative or faceted). For instance, aircraft types can be grouped according to their families of variants which share similar characteristics e.g. single/twin aisle, medium/long range, twin/four engine. By doing so, behaviour displayed by the families of aircraft could become distinguishable. As such, the means of aggregating the concepts into classes and the granularity of the classification have significant influence on ability to make inference from the patterns and trends in the records. Although methods like WordNet and Formal Concept Analysis (FCA) have been used with some success in some applications [45, 46], domain knowledge is still important to provide meaningful abstraction at present.

In this case study, statistical methods were used to analyse the frequency of collocated words to extract most frequent word clusters. High frequency word clusters are a good indication of common phrases referring to some meaningful concepts. The concepts extracted from the ISQ records reflect the domain vocabulary used by the in-service engineers. Table 2 shows the results of word clusters with a minimum frequency of 10. The clusters shown are generated from indices in concordance with the keyword “tank”. The number of words in a cluster (known as a window) and the minimum frequency were arbitrarily set depending on the characteristics of the corpus. The significance of the word clusters as concepts can be improved by breaking the sentence at punctuations and common stopwords. Also as can be seen from the table, concepts automatically extracted can be related to the main assembly, sub-assembly and component of tank, and thus, help towards the construction of the classification schemes. This is similar to the concept pairs heuristics proposed by Yang [47].

**Table 2 Two to five word clusters in concordance with “Tank”, minimum frequency = 10**

From the information contained in the Fuel Systems and Landing Gear database, the full classification shows sparse records in some of the facets, inevitably for those derived from the description as summarised in Table 3. The free-text description allows flexibility for engineers to report issues according to facets of information which are deemed relevant by the engineers. The flight type is the least frequently

quoted information in the records with only 1.03% occurrence rate. The operational phase and the failure mode have 30.14 % and 25.79 % respectively of the whole document set successfully classified. These statistics reflect the inconsistency in the facets of information that are considered important to each ISQ instance as well as the efficiency of the constraints used in the classification. It should be noted that the statistics only apply to a fraction of records in the database. The company in-house system has subsequently been upgraded with more structured information entry following the proposal from this work. Therefore, the statistics are expected to improve if more recent records are being classified.

**Table 3 Statistics of the Classification**

## 4.2 Retrieval and Patterns Identification

In face of ever reducing time to respond to unplanned in-service events, the ISS teams rely greatly on past experience to expedite their responses to the aircraft operators. Previously, they depended greatly on knowledge of the work context and organisation structures, along with a keyword search on the text descriptions and keywords to retrieve and reuse similar past cases. An example scenario may be searching for ISQ based on the engineer responsible or the dates from memory of a similar event. Although the approach may be efficient to familiar users, it is problematic to users outside of the ISS teams (such as designers). By classifying the ISQ records according to additional facets such as assembly/component, failure mode, operational phase etc. retrieval by users who are unfamiliar with the organisation context can be facilitated. For instance, the databases become more useful for designers to interrogate for component-related failure modes, operational-induced issues etc. This added advantage was recognised by the ISS engineers.

In the Waypoint implementation, the ISQ instances (XML documents) were classified according to the faceted schemes that were constructed as previously described. Through the interface, a number of concepts in the faceted schemes can be selected to form a compound query. For instance, if the user is interested in corrosion on aircraft type X, they will select the concept “X” under the facet Aircraft Type and the concept “Corrosion” under the facet Failure Mode. The returned documents are pruned from the system leaving only relevant records that satisfy the conjunction of the selected concepts. At this point, a much smaller but highly relevant set of documents can be retrieved as a list by clicking on the ‘Results’ button on the top right hand corner as illustrated in Figure 4 (a). A list of results that are relevant only to both concepts will be displayed as shown in Figure 4 (b). The display can be customised as well as hyperlinked to other sources of documents that are relevant to those ISQ instances. Due to confidentiality, the actual numbers and some fields are not shown in the figures. The interface shown in this figure uses Dynamic HTML (DHTML) to allow the output to be easily formatted or inserted into other applications or web pages. If the returned document set is still too large, the user can introduce further constraints (for example by selecting the concept “Main Landing Gear” under Assembly) to reduce the set further and increase precision. This way, the faceted classification schemes help to quickly identify and retrieve experiences and resolution to past issues based on a combination of the concepts selected.

**Figure 4 (a) A Waypoint faceted classification interface (numbers do not reflect actual ISQ records) (b) List of documents relevant to selected concepts**

As the users interactively browse through the classification tree and introduce constraints (by selecting more concepts) in Waypoint, the document set is pruned and the remaining relevant document set is updated dynamically. The document count is reflected in the numbers next to each node (at all levels of the tree), where only those documents classified under that node and also relevant to the current selected concepts are displayed (Figure 5). For example, if the user wants to find out about the failure modes that most frequently affect the main landing gear of aircraft X, with these two nodes selected, he/she can navigate to the Failure Mode facet. The ISQ documents will be distributed across the four classes (Damage [a], System [b], Event [c] and General Query [d] as shown in Table 1 with the number of relevant documents displayed next to each node in brackets. This allows the user to infer dominant class(es) of failure mode reported about the main landing gear of aircraft X, for example Damage [a]. By expanding the Damage class, all the active concept nodes (still relevant) are displayed with numbers next to each one of them (Figure 5 (inset)), showing which aspects of Damage are pertinent. At any stage, the concepts with no document count against them will be hidden or shown depending on the user's preference. Again, browsing the classification the users can see which categories have higher occurrences (e.g. mechanical [x], corrosion [y]). The number of documents at a parent node [a] may be equal to or greater than the sum of documents for all the children nodes ([x] + [y] + [z]) as some instances may be classified under more than one concept node. For example, an ISQ instance that reports both mechanical and heat issues will be classified and counted against both the nodes but only counted once for the Damage class. As previously mentioned, it can be anticipated that the construction of the hierarchical structure of the classification schemes will affect inferences that might be drawn because the number of instances in a class is determined by the lower level nodes that are associated with it.

**Figure 5 Number of documents in each node with concepts selected and distribution of failure modes**

By browsing through the faceted schemes in Waypoint, one might be able to discover classes of in-service issues that are systemic such as design-induced (if there is apparent correlation between a component/topological location and a failure mode), operation-induced (if there is apparent correlation between an operational phase/event and a failure mode) and use-induced (if there is apparent correlation between a flight type/route/aircraft operator and a failure mode) issues. As indicated by [48, 49], the main advantage of faceted classification is that compound concepts can be generated dynamically, aiding discovery and synthesis of relationships between those concepts. Potential correlations can be indicated by higher than average number of instances associated with two or more concept nodes. Although a correlation does not necessarily indicate an underlying problem, by indicating a frequent occurrence, it acts as a prompt for engineers to determine root causes, which may lead to understanding of the operating conditions practiced by different airlines (e.g. maintenance procedure, type of operation – short or long haul). Furthermore, it becomes possible to discover less apparent and previously unsuspected correlations

such as a relationship between corrosion and the month in which issues are reported to suggest weather-induced issues. This capability will help towards highlighting recurrent patterns and correlations between different facets across a large set of ISQ instances. Such effort could have been achieved in the past through generated graphs and reports, through manual tracking and compilation of data or through expertise of the experienced engineers but it can be achieved interactively in the Waypoint environment. This mechanism can be used to provide evidence to the design team when formally requesting engineering change or in prioritising continuous development efforts as well as for analysing reliability of components and operational interruptions. Some of these insights may also lead to reporting of lessons learned, which will then benefit future aircraft development programs. A unified knowledge management solution for learning from experience can potentially be realised given that the company has also implemented a software tool for capturing and sharing lessons learned.

# 5 Knowledge Discovery from the Organised Records

The construction of faceted classification schemes involves significant intellectual efforts in defining the levels of granularity and the dimensions of information relating to the records. In doing so, faceted classification provides a mechanism for viewing and analysing the data similar to On-Line Analytical Processing (OLAP) [50]. For instance, aggregations can be built by changing the granularity on specific dimensions (facets) and aggregating data along these dimensions. As described in section 4.2, Waypoint allows the user to interrogate for patterns and trends through its dynamic interface. Where such an approach relies upon human inference, it is possible to augment this approach by using Data Mining (DM), whose Machine Learning algorithms assist in uncovering patterns by algorithmically interrogating data.

DM is essentially a progression of OLAP, allowing queries to be presented at a much more abstract level than those possible using OLAP [51]. DM may be defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [52]. Smyth [53] adds the further caveat that DM is a '…search for previously unsuspected structure and patterns in data', alluding to the freedom that DM has from requirements for prior assumptions or knowledge, such as is present in more traditional statistical analysis.

Figure 6 shows the CRISP-DM methodology, developed by a consortium including companies such as Daimler-Chrysler and SPSS to provide a standard cross-disciplinary approach to Data Mining [54]. The algorithmic part of DM, the modelling stage, is where Machine Learning algorithms are deployed to carefully prepared data to uncover significant relationships and patterns within the dataset. The CRISP-DM methodology highlights the tasks that must be performed before and after modelling, and that there is significant iteration between them – as new understanding is generated by interrogating the data, so different parts of the business (or data describing that business) may be interrogated.

**Figure 6 CRISP-DM process model**

The approach in organising records according to faceted classification schemes described in earlier sections can be considered within those of business and data understanding and (to a great extent) the data preparation stage. The data preparation

stage by itself can account for significant effort [55] and may comprise 80% or more of the time of the complete DM activity [56] and should thus not be seen as trivial. The preparation of semi-structured records, the abstraction of faceted classification schemes and the formulation of rules to associate records with the concept nodes allow for data sets to be interrogated at different levels of granularity. In this case, the business understanding should be reflected in the facets and the abstraction of the schemes. As indicated previously, the structure of the classification tree will necessarily influence inferences that can be made and similarly, the relationships that will be identified through DM. It is due to this reason that the role of domain experts is critical throughout the DM process because they possess the knowledge of the meaning and context of the data, as well as the understanding of both how the data might reasonably be analysed and how to interpret and evaluate any results produced.

The later stages involve the modelling, evaluation and deployment. The evaluation of DM models involves both ensuring the accuracy of the generated Machine Learning models and attempting to distil useful understanding from the results. Traditionally, little attention was paid to ensuring that DM models generated within engineering were accurate [57], which is important if faith is to be placed in their results. This should be seen as an intrinsic part of the Data Mining approach. In terms of understanding the results, it is important to consider the types of modelling that may be utilised. Witten and Frank [58] identify 4 distinct types, which may be listed in ascending order of computational inference as association, clustering, classification and numerical prediction. The first two are known as unsupervised methods, as there is no guiding output parameter which the model is attempting to describe as a function of the various input parameters. Instead, the methods seek to indicate common patterns such as consistently occurring groupings within the data. The latter two are, conversely, supervised methods, seeking to approximate the function which maps an input vector onto an output, and which use historical exemplars (with known outputs for each input vector) on which to base the approximation. Grossman et al [59] note the distinction between these two approaches, identifying two types of model, one which seeks to provide some description of the domain of interest and one which seeks to provide some future prediction. As this research is primarily interested in exploration and explanation of the domain, the first type of model (unsupervised) is preferred.

As the authors are not domain experts within the field of in-service support, it would be difficult to suggest or justify that the pattern identified from DM is either significant or interesting to the practitioner. Therefore, a decision was taken to extract from domain experts a number of past issues which the DM would seek to identify. These past issues are defined as situations where remedial action has been implemented within design teams to address consistently occurring problems in service. During interactions with engineers, one particular issue was frequently discussed. This issue dealt with crack initiation within a non-structural member, and although not detrimental to aircraft safety remained an expensive component to replace, thus remedial work had been undertaken. The ambition of the DM analysis was thus to discover whether such analysis would reveal or otherwise highlight this particular issue within the compiled data.

The Clementine DM software tool [60] was used to conduct the analysis. Of the unsupervised modelling approaches, clustering was identified as most suitable as it provides an indication of common combinations of attributes which contribute towards failure. A number of algorithms may be utilised, including Kohonen’s Self-

Organising Maps [61] and K-Means Clustering [62]. K-Means Clustering was selected due to the simplicity of implementation and of interpretation. As the categorisations for each instance were obtained via interrogation of the recorded query title, a significant number of fields remained blank. Various techniques exist to deal with missing data, Batista and Monard [63] propose three approaches; discarding instances with missing data, parameter estimation [such as through Expectation-Maximisation, 64] and imputation (the consideration of patterns within the data). Conklin and Scherer [65] note that discarding instances is the most common approach, however they argue that case-wise deletion, that of only removing those instances with missing values within the fields of direct interest in a given analysis, is preferable. This approach was taken in this research. The K-Means Clustering algorithm was used to identify clusters within the data using the failure mode and component as input. This was done using the second hierarchical level of the assembly facet and second and third hierarchical level of the failure modes facet respectively for two separate analyses (see Figure 5 inset for concepts at different hierarchical levels). Examples of concepts at each level of the Waypoint classification developed earlier are shown in Table 1. For both failure modes levels the total dataset size was reduced from 4,830 to 611 by case-wise deletion.

Figure 7 shows the clusters identified from K-means algorithm for the two analyses. The known redesign issue that has been resolved previously was identified from the records amongst a number of clusters. The assembly label was removed from the figure for confidentiality. For the specific issue, it may be seen that it occurs as the third most significant node in the first analysis (second level failure mode) but as the $11^{th}$ most significant node in the second analysis (third level failure mode). The clusters discovered were different, and for this particular case, the result from the first analysis was less interesting to the domain experts. This illustrates how the levels of granularity are important in identifying clusters of interest, as it is also possible that certain cases may only be reported at a less detailed level of description. Other clusters discovered were also presented to the domain experts to verify how insightful or informative they were but these will not be elaborated here. From this specific exercise, about 50% of the clusters discovered bear some useful meaning whereas others are not interesting, e.g. hydraulic and leak because this is a known common failure mode for hydraulic systems. This reflects that data mining in practice is a highly qualitative and subjective process and its successful outcome depends on factors such as domain knowledge, data, application and human issues [66].

**Figure 7 Clusters identified from the ISQ records with the component removed**

## 6 Discussion

The previous retrieval approach practised by the ISS engineers, based on contextual clues and keyword search, may not be efficient if the user is unfamiliar with the context or the domain vocabulary. The proposed faceted classification system allows users to form compound queries to facilitate retrieval, potentially expediting the retrieval of past ISQ instances based on the intersection of concepts to rapidly reduce the document sets. This functionality is extremely useful to ISS engineers in view of the very short time often required to respond to in-service queries. More importantly, it was shown that the faceted schemes may also provide evidence for repetitive and systemic issues that can facilitate design learning, which is currently subjective, either

through visually browsing the tree or using more structured data mining methods. From the outset, the intention of the research was not to deploy a computational tool but to demonstrate principles that can be taken up by industry. At time of writing, the research outcomes have been incorporated in an upgrade of the tool used by the ISS engineers as they see value in the approach.

Central to the effectiveness of the system are the information quality and the classification schemes used. From the ISQ records, common concepts are not readily identified due to variation in the way of reporting by engineers. Some of the issues include:

- Inconsistency in referring to the product, failure mode, operational environment etc. in the description
  - Information may be reported at different level of granularity e.g. damage/cracking
  - Alternative terms and synonyms e.g. damage/broken, cracking/chunking
  - Acronyms and abbreviation e.g. MLG/Main Landing Gear; assembly/assy
  - Spelling variation and errors, e.g. centre/center
- Missing or implicit information
  - "MLG Pin Assembly" – issue not reported (it is not clear what the mode of failure might be);
  - "Lower Bearing Corrosion" – main assembly not reported (there are numerous different assemblies with lower bearings).

As mentioned, stemming is done to deal with common lemma of words. Additionally, spelling variation, abbreviation and acronyms can be dealt with using dictionary and domain knowledge to link between similar concepts [67]. Some examples of common mapping required to deal with linguistic variation are shown in Table 4. In Waypoint, variation of terms can be coded into the rules/constraints within the concept map. For the Wing Structures database, facets related to the topological location of the fault were inherited from the original ISQ database. However, the records in these fields were only sparsely populated. Improving the records will help to improve the classification as the information is valuable for learning about the characteristics of the in-service issues. Information was extracted from the unstructured description field to populate the structured and more meaningful faceted schemes. This method requires the classification to have been defined and a dictionary of terms that are expected in each concept node to be built. The records with missing information are matched against those terms and are filled-in. Pre-processing operations will improve the accuracy of the association of the documents to the rules/constraints in the concept map.

**Table 4 Mapping to deal with variation**

The quality of some of the records was improved through information obtained from the ATA chapter, which is an industry-wide, agreed high-level breakdown of a generic aircraft into distinct systems. However, the complete information can only be obtained if one traces through the related records such as the original correspondence and technical reports that are linked to the database instances. Mainly to ensure that

signatures are maintained for legislative purposes, these documents are often in scanned PDF format that is not computer-interpretable. The textual content is also highly unstrucutreed, with various email correspondence and annotated pictures and diagrams. Improving the records by manually extracting relevant information from these essentially image-based records was considered to require excessive resource and to be subject to significant risk of transcription error.

It is argued that continuous evolution and refinement of the classification schemes and the reporting of newly identified key information is necessary to improve utility of the system. Although issues with viewpoint dependency and exhaustiveness of facets cannot be avoided, consensus achieved through involving the domain experts and users in a collaborative manner will guarantee usefulness of the schemes. This is not unlike collaborative ontology creation and maintenance methods that require iteration and revision [68, 69]. With current records as the training set, a taxonomy that is based on the literary warrant can be constructed which will in turn guide the creation of high quality records through prescribed information entry. The use of a controlled vocabulary or taxonomy such as through prescribed data entry (e.g. through selection lists) would significantly enhance the ability to classify, retrieve and mine the in-service information. It is anticipated that a transitionary provision to allow for variation to prescribed data entry will need to provided and analysed for improvement.

Maintaining the faceted schemes requires ongoing reconsideration and revision of the appropriateness of the schemes to classify the records. For example, new and emerging issues affecting composite structures may need to be introduced. The faceted classification approach allows changes to be made within the facet scheme independently, without necessarily affecting other facets. The schemes can be maintained easily because the conceptual and programming development can be separated, domain experts can relay changes conceptually to the IT support. In comparison to ontological based approaches such as that described in [27], which require higher level of rigour in expressing and formalising concepts and their relationships, the faceted classification approach may be more intuitive for human interaction, flexible and readily implemented.

In terms of data mining, other algorithms such as association rules may be useful to describe some interesting relationships between various aspects of the ISQ. Since the focus of this paper is not on the DM algorithms to use, modelling using other algorithms was not pursued. More importantly, further value can be attained from in-service information if the records can be linked to manufacturing-related information, such as the batch number, supplier, etc. For example, using the part and serial number to allow for cross referencing the ISQ records to the Bill of Materials and manufacturing information can enhance prospects of knowledge discovery. Patterns related to design and manufacture or to in-service performance and events (such as the frequency of maintenance) can be investigated that may result in insights and learning.

## 7 Conclusions

In-service information represents a significant resource for learning to the engineering companies. Although much information is gathered from in-service activities, currently the reuse of feedback information is still lacking. At present, service observation may be fed back through unstructured and disparate records making it

difficult to collate and analyse for inherent patterns or correlations. Therefore, learning-from-use relies predominantly on laborious manual trawling through information or through less formal communication channels. There are arguably more incentives for PSS companies to learn from use because such knowledge can allow them to improve their in-service support as well as the design of their future products.

This paper has described a codification approach based on faceted information classification to improve retrieval of in-service records that can potentially support greater learning from use. Computer implementation of the faceted schemes allows the user to arrive quickly at a highly relevant set of results by selecting concepts relevant to the query. In addition, the facets derived from the content of the ISQ records are useful to non-familiar users of the information system. As a result, the system becomes meaningful to others to interrogate, such as designers looking for typical issues raised on a particular component of the aircraft. More importantly, the faceted schemes also allow for patterns and trends in the records to be analysed, either by manually browsing the classification tree or automatically using suitable data mining algorithms. It has been demonstrated and verified with a known issue that the classification schemes can be used to discover various interesting relationships at different levels of granularity. Such capability can provide objective and evidence-based root-cause finding and can help in prioritising continuous development efforts. Ultimately, a closed-loop design learning from in-service experience can be facilitated more formally.

## Acknowledgement

The authors gratefully acknowledge the funding provided by the Engineering and Physical Science Research Council (EPSRC) for the KIM Project (http://www-edc.eng.cam.ac.uk/kim/) under Grant No. EP/C534220/1 and the I*d*MRC under Grant No. GR/R67507/01 for the research reported in this paper. The contribution of our industrial collaborator and Mr. Joe Cloonan is also gratefully acknowledged.

## References

[1] N. Doultsinou, R. Roy, D. Baxter, J. Gao. *Identification of Service Knowledge Types for Technical Product-Service Systems*. in *4th International Conference on Digital Enterprise Technology*, 2007, Bath, UK.

[2] KIM, *Immortal Information and Through-Life Knowledge Management (IITKM): Strategies and Tools for the Emerging Product-Service Paradigm*. 2005-2009, Engineering and Physical Science Research Council. Grant Reference, EP/C534220/1.

[3] N. Rosenberg, *Inside the Black Box: Technology and Economics*. Cambridge University Press, Cambridge, 1982.

[4] M. Cooley, *Architect or Bee? Human/Technology Relationship*. Langley Technical Services, 1980.

[5] D. Blockley, *Engineering from Reflective Practice*, Research in Engineering Design. 1992, **4**, 13-22.

[6] M.A. Maidique, Zirger, B. J., *The New Product Learning Cycle*, Research Policy. 1985, **14**(6), 299-313.

[7] A. Gunasekaran, E. Tirtiroglu, V. Wolstencroft, *An investigation into the application of agile manufacturing in an aerospace company*, Technovation. 2002, **22**(7), 405-415.

[8] M.T. Hansen, N. Nohria, T. Tierney, *What's your strategy for managing knowledge?*, Harvard Business Review. 1999, **77**(2).

[9] E. Wenger, *Communities of Practice and Social Learning Systems*, Organization. 2000, **7**(2), 225-246.

[10] C. McMahon, A. Lowe, S. Culley, *Knowledge management in engineering design: personalization and codification*, Journal of Engineering Design. 2004, **15**(4), 307-325.

[11] A.K.S. Jardine, D. Lin, D. Banjevic, *A review on machinery diagnostics and prognostics implementing condition-based maintenance*, Mechanical Systems and Signal Processing. 2006, **20**, 1483–1510.

[12] P. Fleming, V. Kadirkamanathan, H. Thompson, *Distributed Aircraft Maintenance Environment – E-Science GRID Demonstrator*. 2002-2005, EPSRC: Sheffield.

[13] M. Ong, Ren, X., Allan, G., Kadirkamanathan, V., Thompson, H. A., Fleming, P. J., *Decision Support System on The Grid*, in *Int'l Conference on Knowledge-Based Intelligent Information & Engineering Systems*. 2004: New Zealand.

[14] P. Koudal, Lee, H. L., Whang, S., Peleg, B., Rajwat, P., *OnStar: Connecting to Customers Through Telematics*, Global Supply Chain Management Forum. 2004.

[15] T. Alonso-Rasgado, G. Thompson, B.-O. Elfström, *The design of functional (total care) products*, Journal of Engineering Design. 2004, **15**(6), 515-540.

[16] A. Muller, A. Crespo Marquez, B. Iung, *On the concept of e-maintenance: Review and current research*, Reliability Engineering & System Safety. 2008, **93**(8), 1165-1187.

[17] B. Weir, *Computerised Maintenance Management Systems (CMMS): An Impartial View of CMMS Functions, Selection and Implementation*. accessed May 2009, Plant Maintenance Resource Center

[18] K. Frantzi, *Automatic recognition of multi-word terms:. the C-value/NC-value method*, International journal on digital libraries. 2000, **3**(2), 115.

[19] I. Korkontzelos, I. Klapaftis, S. Manandhar, *Reviewing and Evaluating Automatic Term Recognition Techniques*, *Advances in Natural Language Processing*, 2008, 248-259.

[20] M. Delgado, M. Martín-Bautista, D. Sánchez, M. Vila, *Mining Text Data: Special Features and Patterns*, *Pattern Detection and Discovery*, 2002, 175-186.

[21] MIL, *Military Handbook: Reliability Prediction of Electronic Equipment*. 1995, Department of Defense. MIL-HDBK-217F.

[22] OREDA, *Offshore Reliability Data Handbook 2002*, 4 ed. Det Norske Veritas, 2002.

[23] I.J. James, J. Marshall, L. Walls, *Improving Design For Reliability With In-Service Data Analysis*, in *Proc. Annual Reliability and Maintainability Symposium*. 2002. 182.

[24] C.W. Chan, L.-L. Chen, L. Geng, *Knowledge engineering for an intelligent case-based system for help desk operations*, Expert Systems with Applications. 2000, **18**(2), 125-132.

[25] Y. Cheng, H.G. Melhem, *Monitoring bridge health using fuzzy case-based reasoning*, Advanced Engineering Informatics. 2005, **19**, 299-315.

[26] J. Hendler, *Agents and the Semantic Web*, IEEE Intelligent Systems. 2001, **16**, 30-37.

[27] S.C. Wong, R.M. Crowder, G.B. Wills, N.R. Shadbolt, *Informing Preliminary Design by Incorporating Service Knowledge*, in *International Conference on Engineering Design*. 2007: Paris, France.
[28] A.C. Foskett, *The subject approach to information*. Library Association Publishing, London, 1996.
[29] J. Rowley, J. Farrow, *Organising Knowledge*, 3rd ed. Gower, Aldershot, Hants, UK, 2000.
[30] A. Taylor, *Introduction to Cataloguing and Classification*. Libraries Unlimited, Westport, CT, 1992.
[31] B.C. Vickery, *Classification and Indexing in Science*. Butterworth, London, 1975.
[32] L. Spiteri, *A Simplified Model for Facet Analysis*, Canadian Journal of Information and Library Science. 1998, **23**(1-30), April-July 1998.
[33] M.D. Giess, P.J. Wild, C.A. McMahon, *The generation of faceted classification schemes for use in the organisation of engineering design documents*, International Journal of Information Management. 2008, DOI: 10.1016/j.ijinfomgt.2007.10.001.
[34] C. McMahon, R. Crossland, A. Lowe, T. Shah, J.S. Williams, S. Culley, *No zero match browsing of hierarchically categorized information entities*, Artificial Intelligence for Engineering Design, Analysis and Manufacturing. 2002, **16**(243-257).
[35] A.T. Lowe, *Studies of information use by engineering designers and the development of strategies to aid in its classification and retrieval*, in *Mechanical Engineering*. 2002, University of Bristol: Bristol.
[36] Gemstone, *Faceted Classification and Adaptive Concept Matching*. 2006, Gemstone Business Intelligence Ltd. Technical Whitepaper.
[37] A. Dong, A.M. Agogino, *Text analysis for constructing design representations*, Artificial intelligence in engineering. 1997, **11**(2), 65-75.
[38] Y. Reich, S. Konda, S.N. Levy, I. Monarch, E. Subrahmanian, *New roles for machine learning in design*, Artificial intelligence in engineering 1993, **8**(3), 165-181.
[39] W3C, *Extensible Markup Language (XML)*, in *http://www.w3.org/XML/*. 2008.
[40] C. Beghtol, *Semantic validity: Concepts of warrant in bibliographic classification systems*, Library Resources & Technical Services. 1986, **30**(2), 109-125.
[41] E.W. Hulme, *Principles of Book Classification*, Library Association Record. 1911, **13**(October), 354-358.
[42] S. Boag, Chamberlin, D., Fernández, M. F., Florescu, D., Robie, J. & Siméon, J. . *XQuery 1.0: An XML Query Language - W3C Recommendation*. [cited 2007 Jan]; Available from: http://www.w3.org/TR/xquery/
[43] Apache (2006) *Apache Lucene -- Overview*.
[44] F. Ciravegna, *Understanding messages in a diagnostic domain*, Information Processing and Management. 1995, **31**(5), 687-701.
[45] E. Stoica, M. Hearst, M. Richardson, *Automating Creation of Hierarchical Faceted Metadata Structures* in *Proc. of NAACL-HLT*. 2007: Rochester NY.
[46] U. Priss, *Formal Concept Analysis in information science*, Annual Review of Information Science and Technology (ARIST). 2006, **40**.

[47] K. Yang, Jacob, E., Loehrlein, A., Lee, S., Yu, N., *Organizing the Web: Semi-automatic construction of a faceted scheme*, in *IADIS International Conference WWW/Internet*. 2004: Madrid, Spain.
[48] G.M. Sacco, *Dynamic Taxonomies and Guided Searches*, American Society for Information Science and Technology. 2006, **57**(6), 792-796.
[49] Y. Tzitzikas, N. Armenatzoglou, P. Papadakos, *FleXplorer: A Framework for Providing Faceted and Dynamic Taxonomy-Based Information Exploration*, in *19th International Conference on Database and Expert Systems Application*. 2008: Turin, Italy. 392-396.
[50] D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*. MIT Press, Cambridge, USA, 2001.
[51] U. Fayyad, P. Stolorz, *Data Mining and KDD: Promise and Challenges*, Future Generation Computer Systems. 1997, **13**, 99-115.
[52] M.S. Chen, J. Han, P.S. Yu, *Data Mining: An Overview from a Database Perspective*, IEEE Transactions on Knowledge and Data Engineering. 1996, **8**(6), 866-883.
[53] P. Smyth, *Data Mining: data analysis on a grand scale?*, Statistical Methods in Medical Research. 2000, **9**, 309-327.
[54] CRISP-DM. *Cross-Industry Standard Process for Data Mining*. Step-by-step data mining guide 2000 1 Jan 2005]; Available from: http://www.crisp-dm.org.
[55] I.K. Sethi, *Data Mining: An Introduction*, in D. Braha, Editor, *Data Mining for Design and Manufacturing: Methods and Applications*. Dordrecht, The Netherlands, 2001, Kluwer Academic Publishers.
[56] C. Westphal, T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. John Wiley, USA, 1998.
[57] Y. Reich, S.V. Barai, *Evaluating Machine Learning Models for Engineering Problems*, Artificial Intelligence in Engineering. 1999, **13**, 257-272.
[58] I.H. Witten, E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 2000.
[59] R. Grossman, S. Bailey, A. Ramu, B. Malhi, P. Hallstrom, I. Pulleyn, X. Qin, *The Management and Mining of Multiple Predictive Models Using the Predictive Modeling Markup Language*, Information and Software Technology. 1999, **41**, 589-595.
[60] SPSS, *Clementine 7.0 User's Guide*. 2002: Chicago.
[61] T. Kohonen, *Self-Organising Maps*, Springer Series in Information Sciences. Springer, Heidelberg, 1995.
[62] J.B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. 1967: Berkeley, University of California Press. 281-297
[63] G.E.A.P.A. Batista, M.C. Monard. *An Analysis of Four Missing Data Treatment Methods for Supervised Learning*. in *1st International Workshop on Data Cleansing and Preprocessing*, 2002, Maebashi, Japan.
[64] A.P. Dempster, N.M. Laird, D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, Series B (Methodological). 1977, **39**(1), 1-38.
[65] J.H. Conklin, W.T. Scherer, *Data Imputation Strategies for Transportation Management Systems*. 2003, Centre for Transportation Studies, University of Virginia: Virginia, USA. 127.

[66] C.E. Brodley, P. Smyth, *Applying classification algorithms in practice*, Statistics and Computing. 1997, **7**(1), 45-56.

[67] J.D. Wren, J.T. Chang, J. Pustejovsky, E. Adar, H.R. Garner, R.B. Altman, *Biomedical term mapping databases*, Nucl. Acids Res. 2005, **33**(suppl_1), D289-293.

[68] M. Cristani, R. Cuel, *A Survey on Ontology Creation Methodologies*, International Journal on Semantic Web and Information Systems. 2005, **1**(2), 49-69.

[69] L. Jiahui, M.G. Daniel, *Between ontology and folksonomy: a study of collaborative and implicit ontology evolution*, in *Proceedings of the 13th international conference on Intelligent user interfaces*. 2008, ACM: Gran Canaria, Spain.

**Figure captions**



**Table captions**