J Intell Manuf
DOI 10.1007/s10845-008-0141-1

Applying semantic web technologies to knowledge sharing in aerospace engineering

A.-S. Dadzie · R. Bhagdev · A. Chakravarthy · S. Chapman · J. Iria · V. Lanfranchi · J. Magalhães · D. Petrelli · F. Ciravegna

Received: 7 December 2007 / Accepted: 29 May 2008

Abstract This paper details an integrated methodology to optimise knowledge reuse and sharing, illustrated with a use case in the aeronautics domain. It uses ontologies as a central modelling strategy for the capture of knowledge from legacy documents via automated means, or directly in systems interfacing with knowledge workers, via user-defined, web-based forms. The domain ontologies used for knowledge capture also guide the retrieval of the knowledge extracted from the data using a semantic search system that provides support for multiple modalities during search. This approach has been applied and evaluated successfully within the aerospace domain, and is currently being extended for use in other domains on an increasingly large scale.

Keywords Information extraction · Semantic web · Aerospace engineering · Hybrid search · Ontology search · Information retrieval · Knowledge acquisition · Knowledge management · Knowledge reuse · Knowledge sharing · Ontology · Knowledge organisation · Usability evaluation · System evaluation

A.-S. Dadzie (✉) · R. Bhagdev · A. Chakravarthy · S. Chapman · J. Iria · V. Lanfranchi · J. Magalhães · F. Ciravegna
Department of Computer Science, The University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP Sheffield, UK
e-mail: a.dadzie@shef.ac.uk

R. Bhagdev
e-mail: r.bhagdev@shef.ac.uk

A. Chakravarthy
e-mail: a.chakravarthy@shef.ac.uk

S. Chapman
e-mail: s.chapman@shef.ac.uk

J. Iria
e-mail: j.iria@shef.ac.uk

V. Lanfranchi
e-mail: v.lanfranchi@shef.ac.uk

J. Magalhães
e-mail: j.magalhaes@shef.ac.uk

F. Ciravegna
e-mail: f.ciravegna@shef.ac.uk

D. Petrelli
Department of Information Studies, The University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP Sheffield, UK
e-mail: d.petrelli@shef.ac.uk

Introduction

Although significant effort has been made in the past to design tools that support the capture of organisational knowledge, much of the expertise that employees possess remains elusive. Data is routinely stored in unstructured, or at best, semi-structured form: images, numeric format and natural language are used in e-mails, word-processed documents, spreadsheets and presentations. Data also tends to be distributed geographically, so that organisational and language (terminology) differences magnify the barriers to retrieving its knowledge content.

Rolls-Royce plc (R-R) provides an industrial test bed where large amounts of knowledge contained in semi-structured, distributed data is routinely created and then retrieved for problem-solving. An example of knowledge creation in Rolls-Royce is an *event report*, which is a document created by a service representative each time a finding is recorded on a gas turbine during service. Event reports consist of arbitrarily formatted microsoft word files, sometimes based on user-created templates, and contain knowledge that is very relevant to both engine designers and service representatives, as they help to keep a memory of the issues experienced by customers during service.

A large number of unstructured documents are produced as a result (related to other similarly unstructured, previously existing and new reports), whose knowledge content is currently retrieved using keyword matching only; however keyword searching and metadata generation based on automated information retrieval (IR) yields low efficiency for such document types (Lanfranchi et al. 2007). Keyword matching systems are also not very useful in such circumstances because they only return documents, not knowledge; knowledge seekers have to perform the additional task of manually mining the content of documents retrieved (often stored across multiple media types) in order to extract aggregated data and perform the statistical analysis required to reach useful conclusions for resolving issues identified. This is an expensive and time consuming process, aggravated by the need to repeat the exercise on a regular basis, in order to capture and retrieve new knowledge. The ability to retrieve knowledge such that only the information relevant to a query is extracted from data, and presented to users in context, with information on uncertainty and provenance, would provide significant advantages for knowledge workers.

Capturing existing knowledge therefore places a significant load on users, in addition to the difficulty encountered in sharing and reuse, reducing efficiency, effectiveness and competitiveness. A traditional method for capturing knowledge so that it is easy to retrieve and reuse is to store data in structured databases; however centralised databases and their associated form-filling methods, being relatively inflexible, tend not to meet the requirements for knowledge capture and retrieval for specific workflows in different work areas. Design and service engineers, for instance, make use of similar resources for their work, but have different information needs, based on the knowledge required to perform each role: designers may focus on the wider picture, investigating issues that arise in order to develop new ideas for the next generation of gas turbines, whereas engineers may see the quick and effective resolution of issues that arise for gas turbines in service as more relevant to their role. Further, users often lack the technical skills or may not have access to the resources required to build and/or share databases that suit their information needs.

There is the need to reconsider the whole life cycle of information and knowledge management (KM), from the capturing of new knowledge to the recovery of information from legacy data for new applications, to the sharing and reuse of the knowledge retrieved. A new approach to KM is needed that uses an integrated system to provide the flexibility, efficiency and usability that would make organisational knowledge easily accessible and reusable. This paper describes our experience in using semantic web technology, text mining, knowledge representation, semantic and free-text search to support aerospace engineers in the timely retrieval of the knowledge required for their normal work:

1. *Knowledge acquisition:*
   a. automatic acquisition from legacy data in textual and multi- or cross-media format, using machine learning and matching of regular expressions, and document and feature processing,
   b. user-defined knowledge acquisition (KA) using dynamic web-based forms backed by updateable domain and task ontologies,
   c. supervised annotation of data to improve information storage and retrieval across text, images and cross-media documents.
2. *Knowledge storage:* structured and unstructured storage of keywords and semantic concepts using a proprietary tool.
3. *Knowledge retrieval:* employing multiple combinations of ontology- and keyword-based search, with support for basic statistical analysis of results and the ability to export data for use in external applications.
4. *Knowledge interaction:* effective support for KM without increasing users' cognitive load.

Following an extensive requirements analysis a prototype was implemented and an evaluation was carried out at R-R prior to its deployment for trials. User logs collected since demonstrate the benefits recognised by users, notable being significantly increased effectiveness in knowledge retrieval. (Bhagdev et al. 2007) describe an integrated system that demonstrates the use of our KA and search tools for the complete KM cycle. We also discuss the application of our research to other use cases and settings in industry.

Knowledge acquisition

In large organisations like R-R, the knowledge required to solve challenging problems is typically dispersed over several document repositories within and beyond the organisation, and also across several media. For example, to diagnose the cause of an issue identified concerning a specific component, R-R engineers may need to gather images of similar components, the reports that summarise past solutions, raw (numeric) data obtained from experiments on the materials, and so on. We find that automating the capture of semantic metadata from different repositories and media is an economic solution for the successful deployment of KM systems in such scenarios.

Sections "Semi-automated knowledge acquisition" and "User-defined knowledge acquisition" describe the technology we currently provide for KA from legacy data and new technology being developed to support structured KA at the data generation stage.

Semi-automated knowledge acquisition

Requirements for successful KA for the R-R industrial test bed described include the ability to perform extraction on a large scale, for both single and across different media, bearing in mind the ability to enrich knowledge extracted by fusing (related) information from multiple sources, taking advantage of existing domain and other relevant knowledge, and reporting and handling uncertainty in the information extracted. This section describes the technology developed for information extraction (IE) from text and from cross-media (or compound) documents. Figure 1 shows the structure of tables that may form part of an event report, one example of a semi-structured document from which knowledge must be extracted.

An analysis of a collection of such documents resulted in the following tasks for KA from text:

- *Acronym extraction (AE):* documents often contain acronyms such as *IFSD* and *NGV*. In order to obtain their definition, a manually built lookup table may be required. Some documents however include definitions of the acronyms they use; the application of acronym extraction techniques in such cases minimises human effort in providing such lookup tables (Xu and Huang 2006);
- *Field extraction (FE):* some document types have structured sections, in the form of tables or forms with fields. The content of these fields can be automatically extracted from documents and provided as metadata (Tengli et al. 2004);
- *Entity extraction/ontology population (EE):* many, if not all, R-R document types contain mentions of entities such as companies, engine parts and root causes of issues, that can be mined by ontology population methods into a knowledge base (KB), filling the KB with instances of an ontological type, or that can be highlighted in the original documents using EE methods (Giuliano et al. 2006; Iria et al. 2006);
- *Relation extraction (RE):* given entities identified in the documents described in the point above, RE can determine what kinds of relations exist between those entities, if any (e.g., a component *has description* component_description) (Giuliano et al. 2006).

Orthogonally to the above tasks, and looking beyond event reports to other report types generated at R-R, further analysis highlights the need to handle documents containing multimedia information, with content in text, images and/or numeric data. Figure 2 shows an example of such a document. Detailed analysis has revealed that in many cases, only when considering the different media in a document simultaneously can enough evidence be obtained to derive facts otherwise inaccessible to the knowledge worker using traditional methods for each single medium separately. Cross-media extraction for documents that contain multiple media (hereafter referred to as *compound* documents) is necessary if all the knowledge contained within them is to be extracted in context; IE from the text, image and numeric elements in isolation do not sum up to the whole.

In our approach, the tasks AE, FE, EE and RE, well studied for the case of single-medium acquisition, are generalised to perform cross-media acquisition, as described in the rest of this section. The design for the machine learning framework for cross-media IE presented in Fig. 3 receives as input a multimedia document, and produces semantic annotation with a set of inferred concepts. The process is divided into the following steps:

1. multimedia document processing,
2. integration of single- and cross-media information,
3. the use of background knowledge.

The first step in multimedia document processing is to extract single-medium elements and their relations from the compound document; document processing literature discusses several approaches for extracting layout information from PDF, HTML and other structured documents (see (Laender et al. 2002) for an overview). Single-medium KA algorithms process the content of the corresponding modality, while cross-media KA algorithms process both the content from the different modalities and the layout information. Following the extraction of single-medium content from compound documents, features are extracted from each media element. For image content, MPEG-7 low-level visual features (Manjunath et al. 2002) provide a rich description of the content in terms of colour, shape, texture, and histograms. From text we extract not only the traditional bag-of-words (Manning and Schütze 1999) but also named entities, relations and token-level features such as part-of-speech, orthography and lemma.

A multimedia document can express an idea across different modalities, with each text segment and image offering specific support to the knowledge content; however, determining which elements refer to the same idea or knowledge is not straightforward. The document layout and extracted cross-references (e.g., captions) may suggest how each text segment relates to each image, examples include (Arasu and Garcia-Molina 2003; Crescenzi et al. 2001; Rosenfeld et al. 2002). Arasu and Garcia-Molina (2003), Crescenzi et al. (2001) and Rosenfeld et al. (2002) approaches are based on (manually or semi-automatically extracted) templates that characterise each part of the document. Rosenfeld et al. (2002) implement a learning algorithm to extract information such as the author, title and date. They ignore text content and only use features such as fonts, physical position and other

Fig. 1 Sample table structure found in event reports. (Please note that this diagram describes structure, not the content of an actual report)

module/accessory details			
item	part number	s/n removed	s/n installed
xxxxxxxxx	pxxx-xx	0x-yyxxxx tsn/csn: xxx/xxx	0y-xxyxxxx tsn/csn: x/x

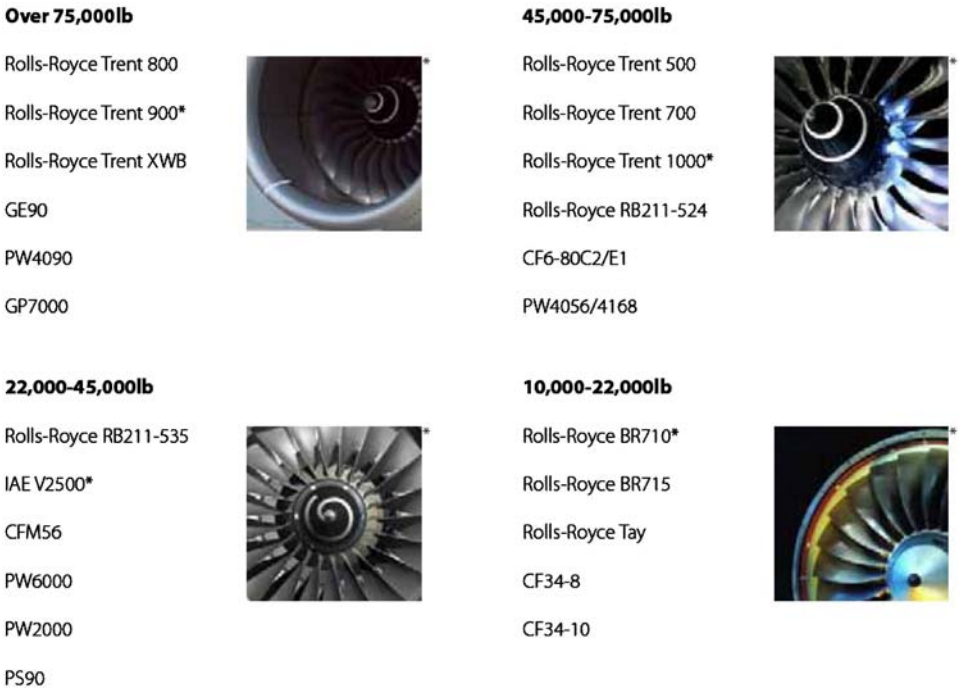
Part numbers
0y-0xxxxxx tsn/csn: xxx/xxx off
0y-1yyyyyy tsn/csn: x/x on

s/n removed	0y-0xxxxxx tsn/csn: xxx/xxx
s/n installed	0y-1yyyyyy tsn/csn: x/x

Parts/Components Removed or Installed (If Any):

On/Off	Part Number/ Serial Number	Part Description	Hours / Cycles		Qty	Destiny / Disposition
Installed	xxxxx	XX YYxx-XXyy)	yyy xxxx	TSN	1	

Fig. 2 An example of a (compound) document (extract from: http://www.rolls-royce.com/civil_aerospace/overview/market/outlook/downloads/outlook2006b.pdf) describing different categories of engines, which requires cross-media analysis



graphical characteristics to provide additional context to the information.

Our approach is similar to that proposed by Rosenfeld et al. (2002), developing a set of cross-media features for the types of documents to be processed. Examples of these features are layout structure, distance between segments, cross-references, font type and colour, and background colour or pattern. All these features can be extracted from PDF or HTML formats, providing the steps that follow with essential information about how elements relate to each other.

Once the features have been extracted from a multimedia document, a learning algorithm can be used to create knowledge models for all the single concepts. Each model is created according to the task being performed (of the aforementioned tasks AE, FE, EE and RE). Sparse feature data such as text and dense feature data such as images have very different characteristics; in cross-media KA the large diversity in the types of data raises the need to pre-process the data to produce a single common representation of all data. The feature processing step aims to estimate a representation that will ease the task of the learning algorithm: we follow Magalhães and Rüger (2007) and process text and images independently with probabilistic latent semantic indexing to produce a canonical representation of both text and image feature space. This allows statistical learning algorithms to more easily handle different types of data simultaneously; cross-media features therefore do not need to be pre-processed since they are extracted such that they can be used immediately.

The estimated single concept models are built using exclusively the concept's own examples. However, media annotations provide information about concepts' co-occurrence across different media. This type of background knowledge describes the semantic structure of the problem that cross-media KA algorithms can exploit to enhance the model of each individual concept. Approaches like those proposed by Naphade and Huang (2001) and Preisach and Schmidt-Thieme (2006) are able to capture the semantic structure of

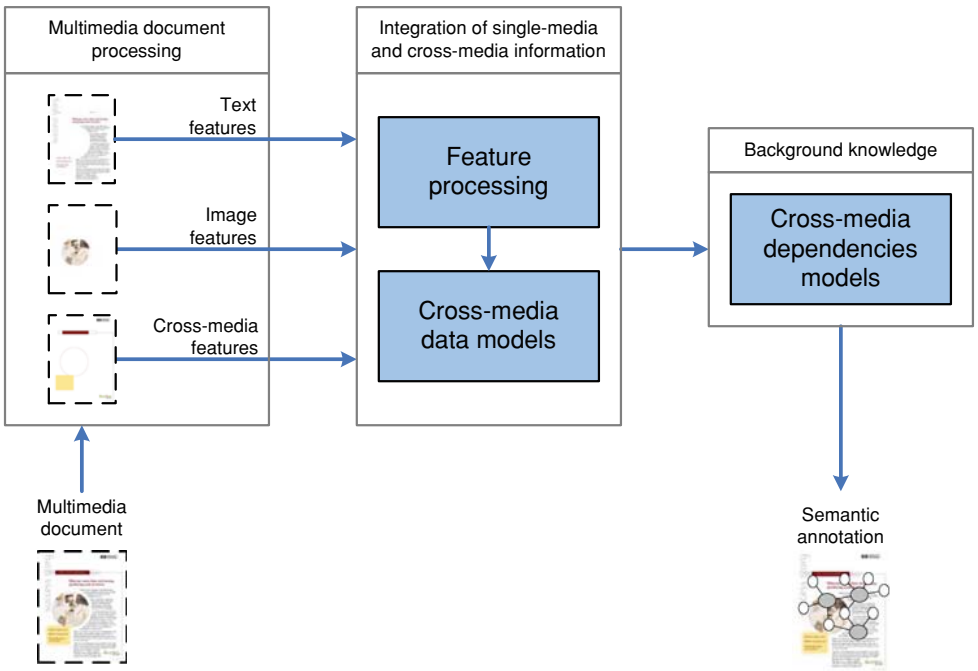


Fig. 3 KA framework for cross-media IE

the problem, thus improving the precision of information extraction algorithms. We adopt this latter method for IE.

User-defined knowledge acquisition

The methods for IE described in section "Semi-automated knowledge acquisition" mean a delay between the data generation and the KA stages. While supervised IE is the chosen solution for KA from large amounts of legacy data there are advantages to be gained by merging KA with data generation for the creation and storage of new knowledge. In *K-Forms* knowledge is captured at data creation time by generating exhaustive and explicit information without forcing users through a pre-defined set of steps. Capturing strategies are flexibly and declaratively designed, allowing each set of users to follow the acquisition strategy that best satisfies their requirements.

The data input is converted into semantic (RDF) statements, by capturing and classifying the knowledge contained using a domain ontology, so to make it available for sharing and reuse. Work flows are modelled via user and task ontologies that may be extended as required to capture new requirements for knowledge and different classes of users. Customised fields and features are obtained by declaring in the ontology the information required, relevant restrictions and how to acquire this information based on user profiles. Relationships among fields such as precedence, layout and associated services are established using the ontology. The actual form design and the document that results from filling in each form are obtained through the use of an ontology reasoner and interpreter.

Related work is that done by Dumas et al. (2002), Gupta et al. (2005) and Corcho et al. (2003). K-Forms provides several advancements, notable being the ability to support the efficient and effective creation of user-defined forms. Previous systems are reliant on technologies such as X-Forms[1] and the manual design of their web forms; our approach provides flexible support for the dynamic knowledge communities that are the target of the K-Forms KA system.

Ontology modelling

K-Forms reuses where possible, existing domain, user and/or task ontologies, in addition to existing templates for data capture, extending each as needed to satisfy requirements for knowledge capture and storage for each community. This serves a dual purpose: in addition to reducing the effort required to create or edit forms, it also provides connections between different but related schemas, improving data and knowledge exchange between communities of users, helping to achieve an important aim of the Semantic Web (Hendler 2001). Figure 4 illustrates the modelling of a concept, the class *component_description*, in a domain ontology used to describe the parts in a gas turbine. In addition to the terms defined, relationships between these terms and the rules governing these relationships may be expressed within the ontologies used in K-Forms. For example, the statement:

Component_description has form component p1: text area

can be expressed within the ontology language OWL DL,[2]

[1] http://www.w3.org/TR/xforms.

[2] http://www.w3.org/TR/owl-guide.

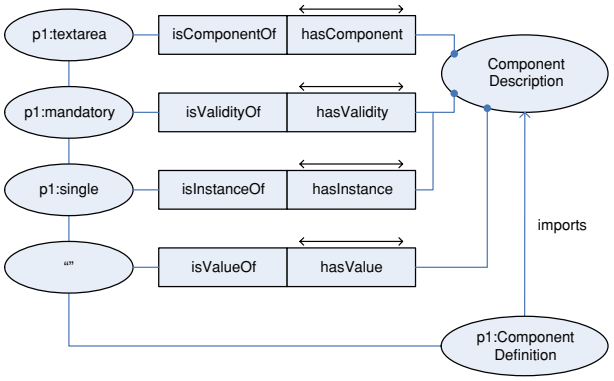


Fig. 4 Modelling restrictions in K-Forms, expressed using OWL DL

where $p1$ is a generic external ontology which models all the available HTML components available for a web form; *component_description* is the concept in the ontology, *has* represents a *has value* relation, and *form component* is an *object* type relation defined within the ontology. OWL DL is used because it supports those users who want maximum expressiveness without losing the computational completeness (all entailments are guaranteed to be computed) and decidability (all computations will finish in finite time) of reasoning systems.

Form template design

In industrial organisations the target end users, who possess the advanced domain knowledge to be captured, typically do not make use of ontology modelling tools such as Protégé,[3] as this does not map directly to user roles and tasks. We attempt to resolve this in K-Forms by using an interactive web form template designer. Users may add forms and populate the content of each form (transparently) using AJAX[4]-based interactive services. Figure 5 shows an example of a web form template being created.

Ontology concepts may be associated with a form field; this adds an extra statement in the RDF document written during knowledge capture, enabling semantic search engines such as that described in Section "Combining keyword and semantic search in a hybrid approach" (see also (Lanfranchi et al. 2007)), to retrieve the knowledge captured. K-Forms enables the input of data via multiple interface strategies, such as drop-down menus for enumerated fields and free text fields. Value checking is performed using an ontology reasoner and interpreter. For single text fields a terminology recogniser, based on string distance metrics (Chapman 2004), aids the identification of the values intended; this is particularly important for fields with hundreds of possible values (there are, for e.g., around 300,000 parts in a gas turbine). The terminology recogniser, constrained by the user input, proposes the correct term or a set of matching options. For example, entering *valve* will return options including *IP bleed valve* and *air start valve*. Free text fields containing (unstructured) descriptions may also be used, and their content semantically annotated (manually) by the user with a tool such as AKTiveMedia (Chakravarthy et al. 2006) or automatically by an IE module (such as described in Section "Semi-automated knowledge acquisition"). Images may be annotated using the same strategy.

In addition to storing the information input into the form a PDF file is generated, using the ontology and the reasoner to explicitly control the generation and layout of the document. Different documents can be created according to the intended usage, by composing the extracted information and knowledge as required. Figure 6 shows the web form that is generated after realisation of the template in Fig. 5.

Knowledge storage and access

Structured vs. unstructured information storage

Captured knowledge, both that extracted from legacy documents (refer Section "Semi-automated knowledge acquisition") and during data generation (Section "User-defined knowledge acquisition"), must be stored such that it enables wider and more extended usage. Traditionally two distinct approaches have been used:

Structured information storage

Here knowledge is formally organised into a predefined rigid structure used for direct knowledge access, the most prevalent example of structured knowledge being a database. When an ontology is used, concepts in the document can be related through logical statements, e.g., *discolouration* on *component_number_1245* allows the retrieval of all instances where *component_1245* was found to be *discoloured*. Ontology-based, structured storage can be used to associate formal metadata to text, making the document content (as opposed to its keywords) available for automatic processing (Berners-Lee et al. 2001). Storing structured knowledge has several advantages: quantitative analysis, reasoning, automatic manipulation and direct access to knowledge are all useful tools to data analysis and knowledge discovery. However, structured data storage is inflexible as access to it is constrained to that of the storage mechanism itself; knowledge within the document that is not represented in the data structure is lost. In addition, designing structured resources for holding knowledge is costly, and although covering the most typical cases, may exclude specialist usage. Even the best

[3] http://protege.stanford.edu.

[4] http://developer.mozilla.org/en/docs/AJAX.

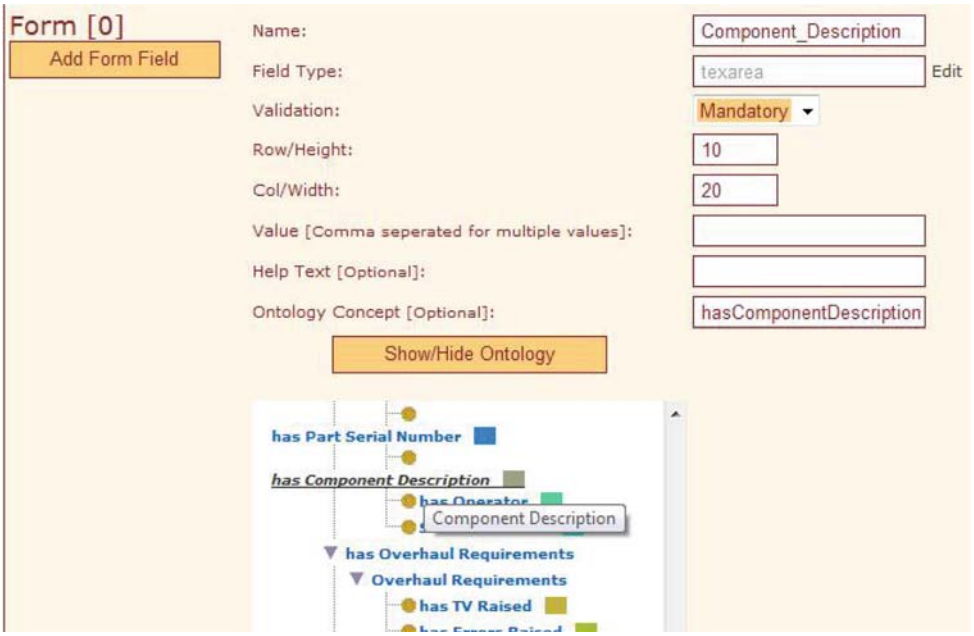


Fig. 5 Form template design in K-Forms

Fig. 6 Form realisation based on the underlying ontology (see also Fig. 5)

ontology is unlikely to cover all user information needs, as these may change with time and in ways that may not conform to the ontology design; further, ontology updating is rarely done as it tends to be a complex and expensive activity.

Unstructured information storage

In this case data is not coded into a predefined structure, but instead is indexed to allow free-text retrieval. An example of unstructured data access is provided by search engines that index and retrieve knowledge from enterprise archives or the world wide web. At access time the user expresses their information needs by entering a query; the search engine uses its index to retrieve the data relevant to the current query. While this approach provides a high degree of flexibility, as the data retrieved is that which matches the query, there is no guarantee the retrieved result set contains all and only the relevant data existing in the repository. As a consequence quantitative analysis is unreliable (see section "Effectiveness of hybrid search system" for the discussion on recall and precision). Much of the effectiveness of this technique depends on the text source; engineers' reports are generally extremely succinct and the language very specialised, two conditions highlighted as critical for traditional unstructured, keyword-based retrieval. A further complication in technical text is the context and ambiguity of keywords used. In the example *find all cases in which a component was changed due to discolouration*, the fact that *discolouration* was observed on a *component* is fundamental to answering the question. A report featuring a *component* and *discolouration* without revealing the connection between the terms fails to answer the question. Another limitation is in the retrieval of entire documents; reading of all documents' content is required to extract the specific knowledge and make optimal use of the data. It must be noted that although unreliable, with contextual information missing and issues of conceptual ambiguity, unstructured knowledge repositories support all possible combinations of knowledge contained in the indexed documents.

Structured and unstructured data both have advantages and disadvantages. At a closer look they are complementary, one providing order and precision, the other flexibility and multiplicity. To take advantage of both, a dual store that simultaneously holds structured and unstructured information was created, the *K-Store*. For unstructured information a normal keyword index is used, in this case SOLR,[5] alongside a semantic index[6] storing structured triples. This technique differs from CSail which indexes semantic stored instances as text; in our approach the entire document content is indexed, meaning complete coverage in the keyword index, without being limited to concepts only.

[5] http://lucene.apache.org/solr.

[6] This project uses Sesame (http://www.openrdf.org) but this can easily be exchanged for alternative (appropriate) stores.

A semantic store holding extensible graph-based triples does not suffer from the same flexibility issues as do databases; data within a semantic store is associated with an ontology which expresses taxonomic typing/classification and extendable relational interactions between graph-based knowledge instances. Such representations facilitate higher level inference as well as allowing additions to the underlying data without impairing use of the repository. These inbuilt possibilities allow querying to extend beyond that which databases can offer. In particular stores such as 3-store,[7] Allegrograph[8] and Sesame facilitate flexible query through the SPARQL[9] query language. This freedom along with the improved coverage of keyword approaches allows more flexible knowledge access than that offered by a single mechanism.

In summary, in order to combine structured and unstructured querying our search system, K-Search,[10] performs three offline steps:

- indexing documents using keywords,
- defining a domain ontology,
- gathering structured knowledge using an ontology (see sections "Semi-automated knowledge acquisition" and "User-defined knowledge acquisition").

The dual storage thus created supports more flexible interaction as the user is able to choose between differing search modalities for the task in hand. These search modalities are described in section "Combining keyword and semantic search in a hybrid approach".

Combining keyword and semantic search in a hybrid approach

To take advantage of the double storage facility in K-Store we have designed a new method for information search. *K-Search* combines the flexibility of unstructured retrieval with semantic structure, making synergistic use of the strengths of both techniques, and supporting users in focusing on relevant issues with faster retrieval and more accurate results.

From a user point of view, structured queries must be formulated in a logical language that has to be learned and remembered. Conversely, unstructured retrieval has the advantage of being all encompassing—any term can be searched for independently of previous processing—and straightforward to use,—terms are simply entered into a (natural language) query. The interface of K-Search supports the user in formulating queries in whichever way suits their skills and information requirements.

The hybrid approach (HS) adopted in K-Search uses and fuses keyword search (KS) and ontology search (OS). Each mode may be used independently, or OS and KS may be combined within the same query (to form a hybrid query), depending on the purpose of the search. The user interface of K-Search (see Fig. 7) has been designed to support the composition of HS queries as well as quick changes from one mode to another, KS only or OS only. At retrieval time, K-Search performs the following steps:

- the user query is parsed and the types of searches are identified: KS (unstructured), OS (structured) and combinations of the two (HS);
- keywords are sent to K-Store (the dual storage system described in section "Structured vs. unstructured information storage") for unstructured retrieval; this will return the identifiers (URIs) of all the documents containing the keywords;
- queries about concepts (and their relations) are matched with the triples stored in the triple store, K-Store, using the query language SERQL (Broekstra and Kampman 2003),[11] with support being extended to SPARQL also;
- if the user has formulated a query using both KS and OS, the results of the different queries are merged to obtain HS results.

When a query is performed, the result set contains the reports where the concepts and the keywords in the query co-occur. Figure 7 shows the result set displayed as a list in the mid-right panel of the interface; each item in the list shows the name of the document and the values of the fields used for OS. Individual reports are displayed on the bottom right on request (by clicking on the file name for a list item). Multiple documents may be opened simultaneously, each displayed in a different tab. The original layouts of the documents are maintained. Annotations are made evident through colour highlighting, and are the means to advanced features or services (Lanfranchi et al. 2005); for example, clicking on a concept results in query expansion using the selected term.

One of the advantages of structured data is the support for quantitative analysis of the retrieved result set using graphs and charts. K-Search allows users to select concepts from the ontology to display the retrieved set with respect to the selected parameters. The graph on the right in Fig. 7 groups the results retrieved for technical variance requests submitted in the year 2004, where *component_description* contains the term *valve*, by (the concept) *engine_family*. Each plotted

[7] http://www.aktors.org/technologies/3store.

[8] http://agraph.franz.com/allegrograph.

[9] http://www.w3.org/TR/rdf-sparql-query.

[10] Our hybrid search system was initially named X-Search; it has however been renamed to K-Search as it was discovered that the name had been previously copyrighted.

[11] http://www.openrdf.org/doc/sesame/users/ch06.html.

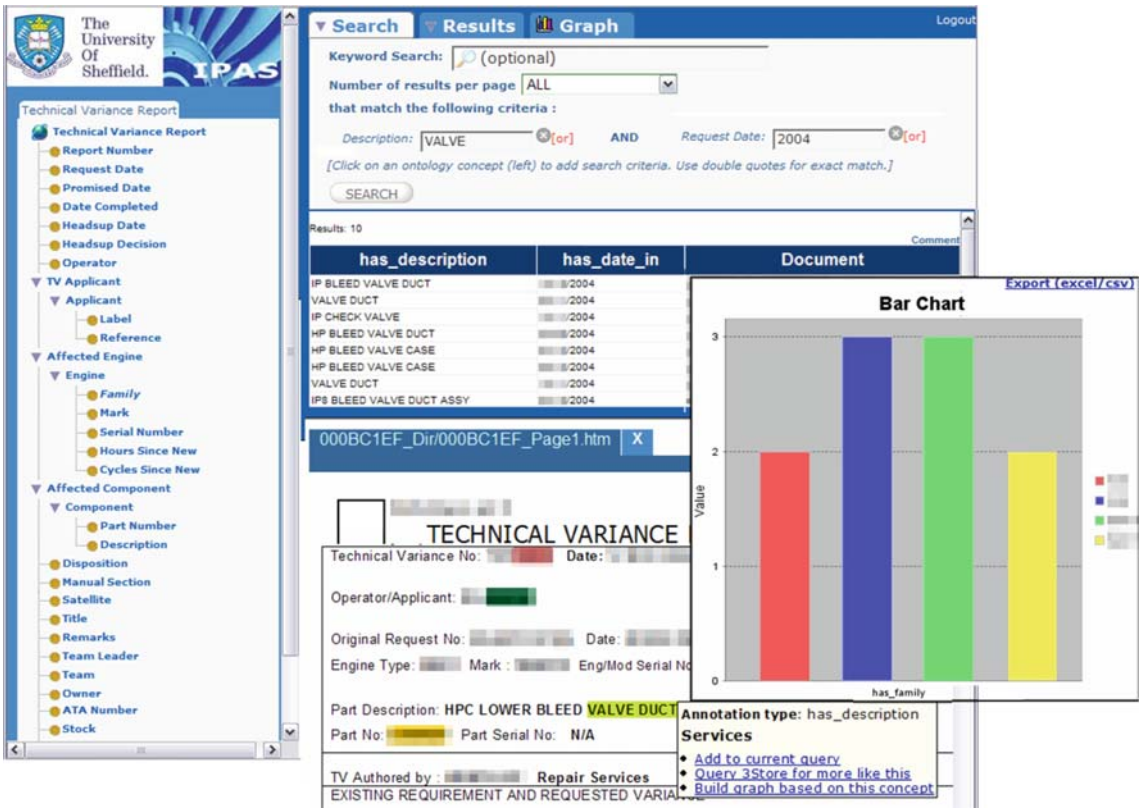


Fig. 7 K-Search query interface and visualisation of results (Please note that non-relevant sections of the image have been deliberately obscured)

engine_family (each bar in the example in Fig. 7) is active and may be clicked on to focus on the sub-set of documents that contains that specific occurrence.

Results from tests and field trials

Two pilots for K-Search have been installed at R-R, to retrieve knowledge from event reports and technical variances, with simple graphical analysis of the information retrieved. The pilots are currently at the beta stage, after improvements made based on in-house system and stress testing, and on-site user evaluation.

Figure 8 shows the interaction between the components used to build a fully integrated system for the complete KM lifecycle (Bhagdev et al. 2007), demonstrating the flow of knowledge extracted from legacy data (unstructured), and at document creation time (structured), into document stores based on Sesame and 3-store. IE was performed using T-Rex (Iria et al. 2006; Iria and Ciravegna 2006), which performs adaptive IE and document classification, and Saxon,[12] (Greenwood and Iria 2008) a rule-based annotation tool.

[12] http://nlp.shef.ac.uk/wig/tools/saxon.

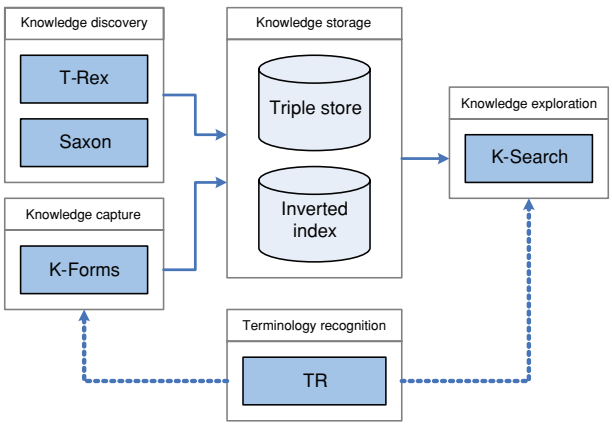


Fig. 8 Integrated system architecture, showing interaction, knowledge and data flow, between the different applications developed to support the KM lifecycle using semantic web technologies (described in further detail in (Bhagdev et al. 2007))

The terminology recogniser based on string distance metrics (Chapman 2004) feeds into the knowledge acquisition and sharing stages, to disambiguate terms used. The assertions retrieved, stored as triples and using an inverted index, are made available for use in K-Search.

Effectiveness of hybrid search system

The quality of the metadata generated by the adaptive IE system, T-Rex, was evaluated on a corpus containing irregularly structured tables in 400 documents: precision was seen to increase with the number of documents in the training set, from 76% for 40 documents to 90% for 240. Recall remained constant, at 100%. Overall results for a two-cross fold test for 200 documents were 95.12% for precision and 97% for recall, with an F-measure of 95.84%, confirming that the metadata generated was of very high quality (refer (Lanfranchi et al. 2007) for more detail).

The standard equations used to calculate precision, recall and the F-measure are:

$$Precision = \frac{Correct\ System\ Answers}{System\ Answers},$$

$$Recall = \frac{Correct\ System\ Answers}{Expected\ Answers},$$

$$F_\alpha = (1-\alpha)\frac{Precision \cdot Recall}{Precision + Recall},$$

where *Expected Answers* is approximated with the cardinality of the set of all the relevant documents returned by any of the three modalities, standard practice in evaluations on large sets of documents. We used a value of 1 for α, to obtain a weighted harmonic mean.

The HS system was then tested using 21 queries generated based on real tasks carried out by users at R-R, performing independent tests for KS and OS, followed by tests to evaluate the effectiveness of HS, for a corpus of 18,097 event reports and using a domain ontology built by Aberdeen University as part of the IPAS[13] project.

A sample query that requires a combination of OS and KS is: *what events were identified during maintenance in 2003 with a cause due to control units*. For the purposes of this evaluation, our implementation of HS uses OS where information is covered by the ontology and metadata is available, and KS for all other instances. Standard measures for precision (retrieval of only those documents relevant to a query) and recall (retrieval of all documents relevant to a query) were calculated for the first 20 and 50 documents returned for each search type (KS, OS and HS).

Figure 9 summarises the results of the user evaluation, showing, for the F-measure for HS, an increase of 49% for KS and 55% for OS:

- Precision: equal to that for OS and an increase of 51% over KS,
- Recall: an increase of 109% over OS and 46% over KS.

Query	POS	Keyword 20			Ontology 20			Hybrid 20 General		
		COR	ACT	EXP	COR	ACT	EXP	COR	ACT	EXP
Q1	84	16	20	20	20	20	20	20	20	20
Q2	22	16	20	20	0	0	20	16	20	20
Q3	25	1	20	20	11	20	20	11	20	20
Q4	63	19	20	20	19	20	20	19	20	20
Q5	27	9	20	20	12	20	20	12	20	20
Q6	5	4	8	5	0	0	5	4	8	5
Q7	7	6	6	7	0	0	7	6	6	7
Q8	1	1	1	1	0	0	1	1	1	1
Q9	5	3	3	5	0	0	5	5	5	5
Q10	83	12	20	20	0	0	20	20	20	20
Q11	2	1	1	2	0	0	2	1	1	2
Q12	3	3	3	3	0	0	3	3	3	3
Q13	7	6	6	7	0	0	7	6	6	7
Q14	145	19	20	20	19	20	20	20	20	20
Q15	40	8	20	20	0	0	20	20	20	20
Q16	11	1	16	11	11	11	11	11	11	11
Q17	13	3	20	13	0	0	13	4	4	13
Q18	7	1	4	7	0	0	7	4	20	7
Q19	25	10	17	20	0	0	20	11	11	20
Q20	53	3	20	20	20	20	20	20	20	20
Q21	37	18	20	20	0	0	20	20	20	20
TOTAL	665	160	285	281	112	131	281	234	276	281
		PREC	REC	F-MEAS	PREC	REC	F-MEAS	PREC	REC	F-MEAS
		0.56	0.57	0.57	0.85	0.40	0.54	0.85	0.83	0.84

Fig. 9 Comparative evaluation of KS, OS and HS on 21 queries

A user evaluation was carried out with 32 employees at R-R from the design, service and business departments, to obtain information on usability, including among others, measures of efficiency and effectiveness, and how well users understand the HS paradigm and what benefits they perceive over KS or OS in isolation. Each evaluation involved an (assisted) familiarisation exercise, followed by a set task and a user-suggested task completed without assistance, to allow users to define their own knowledge retrieval strategies. The evaluation was concluded with a user satisfaction questionnaire and a brief interview.

An analysis of the evaluation results showed users to have understood the HS concept, with different strategies adopted, confirming that our implementation of HS is able to support multiple approaches to knowledge retrieval. Learnability, ease of use, system accuracy and speed all recorded very positive results (see Fig. 10).

The comparison of free text and ontology retrieval may be seen as both a new and an old research area; previous research has shown that the combination of keyword search with controlled terminology leads to an increase in both precision and recall. Using ontologies allows automatic reasoning and other advanced features, some systems have been developed to take advantage of the benefits recognised (see Kiryakov et al. 2003, 2004; Ducatel et al. 2006); there is, however, significant work still to be done in this research area.

K-Search was initially developed for a case study as part of the IPAS project, and its hybrid search mechanism has been tested only on R-R corpora. Knowledge models for KA using the cross-media algorithms have been built on TRECVID[14] data, we are now implementing these models using data provided by R-R for the X-Media project. As part of the tool integration exercise currently being performed for X-Media the KA tools for text and cross-media are being run on other

[13] http://www.3worlds.org.

[14] http://www-nlpir.nist.gov/projects/trecvid.

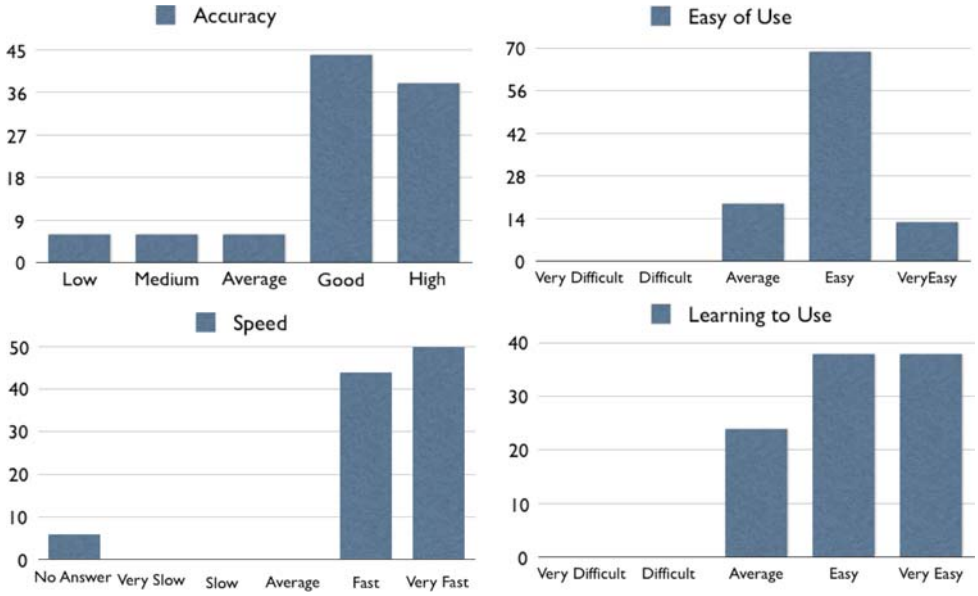


Fig. 10 Results of the evaluation of K-Search by 32 users (values are in %)

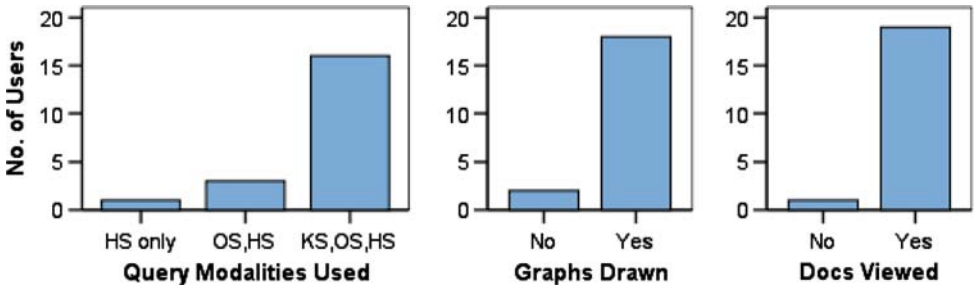


Fig. 11 Summary of usage for K-Search pilot for event reports

corpora, both within R-R and also using other benchmark datasets such as that provided by Reuters.[15] The results of our tests will be published as they become available.

Analysis of user logs

An analysis of the usage logs for the beta version of the first pilot of the HS system, (knowledge retrieval from event reports), for 20 users revealed varied use of the system, with a small number of users trying it out for periods of up to half an hour to the majority trying out the system for over an hour on a single day. A quarter of the users made repeated use of the pilot over a number days. Figure 11 summarises usage of the first pilot.

The most commonly occurring queries included (on average 1 or 2, but up to 4 in a single query) concept searches on *engine_type*, *part_removed*, *airport_location*, (report) *date* and *operational_effect*. KS only and keyword-in-concept searches were also mixed with OS. Usage appears to follow the general trend:

1. run a series of queries;
2. open a set of documents (for viewing) and/or attempt to draw a graph (graph requests tend to start with only one element to group on, then are repeated with a second, i.e., a sub-group);
3. run more queries (often a refinement of previous queries or similar queries using alternative strategies);
4. open another set of documents (based on new search results or go back to view previously unopened results).

User feedback concerned missing elements in results (user expectations based on experience—the pilot had access to a sub-set of the event report corpus at R-R), additional features and options for statistical analysis (e.g., the option to draw scatter plots). Users appear to appreciate the benefits provided for KM; a recurring request was to include other types of corpora as event reports represent only a fraction of the documents made use of in typical knowledge retrieval tasks.

Application to other industrial settings

A third pilot is currently undergoing system testing and heuristic evaluation, prior to installation for evaluation by end users at R-R, to support structured KA during the creation

[15] http://www.daviddlewis.com/resources/testcollections/reuters21578.

of module condition reports. Knowledge acquired from the web-based forms is fed into our HS system, so that the new knowledge is immediately available for retrieval; the separate step, performing (supervised) IE, is no longer required. This pilot presents a good case for evaluating the benefits gained by integrating the different technologies we provide; up to this point we had provided the K-Search tool on its own for knowledge retrieval from (legacy) corpora that had been indexed and annotated using supervised machine learning methods. Providing support for KA during document creation (using K-Forms) completed yet another step in the KM lifecycle, giving end users greater control over the creation, storage, sharing and re-use of the knowledge required for their daily work. The provision of a single, integrated system prevented the increase in cognitive load associated with switching between different tools, due to the break in the process flow.

Benefits immediately recognised by users during an informal presentation to key personnel included the ability to capture information using a more structured process, ensuring that all required data is recorded, in addition to the metadata used to enrich the formal information captured. At the community level the ability to customise knowledge capture to the specific needs of these smaller working groups removes the restrictions posed by data capture using centralised databases. The process followed in creating new K-Forms also suggests relationships between fields declared on a form and existing concepts in domain and other related ontologies, so that explicit connections can be made that provide greater support for knowledge sharing across communities of practice.

An opportunity to test the scalability of our technologies, especially for cross-media KA, is participation in the X-Media[16] project, whose aim is to provide innovative solutions for KM on a very large scale in complex, distributed environments and across different media. The tools being built for IE from text and cross-media documents (in addition to IE tools for images and numeric data developed by other X-Media consortium partners) are being applied to four of the X-Media use cases developed to map different aspects of the KM life cycle in two industrial test beds, R-R and the automotive manufacturers, Fiat S.p.A. The knowledge extracted from the various data sets is retrieved using SERQL and SPARQL queries, to be fed into search and presentation systems, one of which is the hybrid search system, K-Search, to allow end users to retrieve knowledge based on their individual requirements. Tool development for the X-Media project is currently at the module integration stage: the tools developed to meet the requirements of a specified stage in the KM lifecycle are being configured to talk to the X-Media *Kernel*, the layer in the architecture that provides standard methods for communication between the different underlying technologies and the user interface. Detailed information on the results of unit and complete system testing, resolution of issues encountered during integration and the user evaluations to be performed will be published as they become available.

[16] http://www.x-media-project.org.

Conclusion

This paper describes the exploitation of semantic web technologies to support the KM life cycle, using examples of the applications of our research to real cases in industry. The processes followed for IE from text and cross-media documents are described, followed by technology for performing KA during data generation, by modelling users' information flows during their normal work. We then describe our hybrid search system, which complements keyword search with semantic search based on domain ontologies, to support knowledge retrieval and reuse.

Our technologies have been tested in real-world applications at R-R, and after extensive user evaluation, been successfully released to end users in a pilot scheme. Benefits over existing systems for KM have been recorded, notable being a significant reduction in time and user effort, in addition to higher precision and recall in knowledge retrieval. Demonstrations of the new system for KA during data generation to higher level management and other key personnel in industry have been received very well; the potential for improved KM, and the resultant increase in efficiency, effectiveness and competitiveness was recognised.

Finally, we are preparing for the evaluation of these applications on a much larger scale, looking toward the increased requirements for KM in the future, working also with KM tools developed by other members of the X-Media consortium, and applied to other use cases in industry, as part of the X-Media project.

Acknowledgements The technologies described in this paper have been funded by the X-Media, IPAS and AKT projects. X-Media is sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978. IPAS is part-funded by the UK Department of Trade and Industry under the Technology Program and by Rolls-Royce plc., DTI Reference TP/2/IC/6/I/10292 IPAS. The AKT project (http://www.aktors.org) was sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/N15764/01).

References

Arasu, A., & Garcia-Molina, A. H. (2003). *Extracting structured data from web pages*. San Diego, California: ACM SIGMOD international conference on management of data.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The semantic web*. Scientific American.

Bhagdev, R., Butters, J., Chakravarthy, A., Chapman, S., Dadzie, A.-S. et al. (2007). Doris: Managing document-based knowledge in large organisations via semantic web technologies. *6th international semantic web conference ISWC 2007* (Semantic Web Challenge Track), Busan, Korea.

Broekstra, J., & Kampman, A. (2003). *SeRQL: A second generation RDF query language*. Amsterdam, Netherlands: SWAD-Europe workshop on semantic web storage and retrieval.

Chakravarthy, A., Lanfranchi, V., & Ciravegna, F. (2006). Cross-media document annotation and enrichment. *1st semantic authoring and annotation workshop, Proc., ISWC 2006*.

Chapman, S. (2004). SimMetrics: A similarity library of metric algorithms for integration and comparison, http://www.dcs.shef.ac.uk/~sam/simmetrics.html.

Crescenzi, V., Mecca, G., & Merialdo, P. (2001). RoadRunner: Towards automatic data extraction from large web sites. *International conference on very large data bases*.

Corcho, Ó., Gómez-Pérez, A., López-Cima, A., López-García, V., & Suárez-Figueroa, M. C. (2003). ODESeW. Automatic generation of knowledge portals for intranets and extranets. *Proceedings of the International Semantic Web Conference, ISWC 2003*, Sanibel Island, Florida, USA.

Ducatel, G., Cui, Z., & Azvine, B. (2006). Hybrid ontology and keyword matching indexing system. *Proceedings of the 15th International World Wide Web Conference*, WWW 2006: Workshop IntraWebs 2006, Edinburgh, Scotland.

Dumas, M., Aldred, L., Heravizadeh, M., & ter Hofstede, A. H. M. (2002). Ontology markup for web forms generation. *Proceedings, WWW'02 Workshop on Real-world Applications of RDF and the Semantic Web*.

Giuliano, C., Lavelli, A., & Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy.

Greenwood, M. A., & Iria, J. (2008). *Saxon: An Extensible Multimedia Annotator (to appear in) Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco.

Gupta, S., Hawker, J. S., & Smith, R. K. (2005). Acquiring and delivering lessons learned for NASA scientists and engineers: A dynamic approach. *ACM Southeast Regional Conference, 2*, 370–375.

Hendler, J. (2001). Agents and the semantic web. *IEEE Intelligent Systems, 16*(2), 30–37.

Iria, J., & Ciravegna, F. (2006). A methodology and tool for representing language resources for information extraction. *Proceedings of the LREC 2006*.

Iria, J., Ireson, N., & Ciravegna, F. (2006). An experimental study on boundary classification algorithms for information extraction using SVM. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Kiryakov, A., Popov, B., Ognyano, D., Manov, D., Kirilov, A., & Goranov, M. (2003). Semantic annotation, indexing, and retrieval. *Proceedings of the International Semantic Web Conference, ISWC 2003*, Sanibel Island, Florida, USA.

Kiryakov, A., Popov, B., Dimitar, M., Ognyanoff, D., Marinov, R., & Terziev, I. (2004). Automatic semantic annotation with KIM. *Proceedings of the 3rd International Semantic Web Conference, ISWC 2004*: Demo Papers, Hiroshima, Japan.

Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *ACM SIGMOD Record, 31*, 84–93.

Lanfranchi, V., Ciravegna, F., & Petrelli, D. (2005). Semantic web-based document editing and browsing in AktiveDoc. *2nd European Semantic Web Conference ESWC*.

Lanfranchi, V., Bhagdev, R., Chapman, S., Ciravegna, F., & Petrelli, D. (2007). Extracting and searching knowledge for the aerospace industry. *Proceedings of the ESTC 2007*.

Magalhães, J., & Rüger, S. (2007). *Information-theoretic semantic multimedia indexing*. Amsterdam, Holland: ACM conference on image and video retrieval.

Manjunath, B. S., Salembier, P., & Sikora, T. (2002). *Introduction to MPEG 7: Multimedia content description language*. Wiley.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Naphade, M. R., & Huang, T. S. (2001). A probabilistic framework for semantic video indexing filtering and retrieval. *IEEE Transactions on Multimedia, 3*, 141–151.

Preisach, C., & Schmidt-Thieme, L. (2006). *Relational ensemble classification*. Hong Kong, China: IEEE international conference on data mining.

Rosenfeld, B., Feldman, R., & Aumann, J. (2002). *Structural extraction from visual layout of documents*. McLean, Virginia, USA: ACM CIKM.

Tengli, A., Yang, Y., & N. Ma, L. (2004) *Learning table extraction from examples*. Geneva, Switzerland: (COLING'04).

Xu, J., & Huang, Y. (2006). Using SVM to extract acronyms from text. *Soft Computing, 11*(4), 369–373.

The WIT tools page: http://nlp.shef.ac.uk/wig/tools.