Issues in moving to a semantic web for a large corporation.

Gary Wills[1], David Fowler[2], Derek Sleeman[2] , Richard Crowder[1], Simon Kampa[1], Leslie Carr[1] and David Knott[3].

[1] Intelligence, Agents, Multimedia Group, University of Southampton, Southampton, SO17 1BJ, England, UK
{gbw, rmc, srk, lac} @ecs.soton.ac.uk
[2] Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, Scotland, UK
{dfowler, sleeman} @csd.abdn.ac.uk
http://www.csd.abdn.ac.uk[2]
[3] Rolls-Royce plc, Derby, UK
david.knott@rolls-royce.com

Abstract. In many large engineering design organizations the information systems have developed over time into a set of heterogeneous resources. This makes it difficult for engineers to follow a trail through the resources. This situation becomes particular difficult when the Engineer is new to a company; unfamiliar with the systems and unaware of the history of the designs. This paper presents a demonstrator system developed with a major aerospace company to aid engineers, through the use of knowledge technologies, to locate the documentation they require. The paper presents the systems and lessons learnt to enable the organisation to move towards a more semantically enriched document repository.

1. Introduction

In many large organisations which perform a substantial amount of engineering design, information systems have developed over time into a set of heterogeneous resources. This makes it difficult for engineers to follow a trail through these resources. This situation becomes particular difficult when the engineer is new to a company; unfamiliar with the systems and unaware of the history of the designs.

The challenge for organisations is to develop an information system that is both comprehensive and will satisfy the increasing demands from industry for up-to-date and easily accessible information. In addition, technical documentation is frequently highly cross-referenced, often to documents in different formats. The process of locating information then becomes time-consuming, frustrating users as they open and close applications looking for essential information as they move between information systems. Conventional information management techniques are not considered sufficient to satisfy these requirements [18].

A hypermedia system allows associations to be made between information in different media in a manner similar to that naturally undertaken by people. The concept and requirements for industrial strength hypermedia were initially presented by Malcolm *et al* [16]. Due to the increase in virtual enterprises, lean and agile manufacturing, the demands for correct and easily accessible information have not abated. Wills *et al* have demonstrated the effective application of open hypermedia to industrial applications [22].

1.1 Background

The Advance Knowledge Technologies (AKT) project is one of the UK government's funded Interdisciplinary Research Collaborations (IRCs). The AKT project involves five universities from the UK. The AKT project aims to develop and extend a range of technologies providing integrated methods and services for the capture, modelling, retrieval, publishing, reuse and maintenance of knowledge [1]. As well as the academic expertise, AKT benefits from a close relationship with industrial collaborators. One such collaborator is Rolls-Royce, a major aerospace company based in the UK.

Rolls-Royce takes seriously the need to keep accurate records of design, manufacture and test; in part this is due to the nature of their aerospace business and in part due to the recognition that their main method of capturing corporate memory is through such documentation. As with many organisation's information systems, the system at Rolls-Royce has evolved over time, resulting in a number of document and information management systems.

Rolls-Royce is organised into Business Units (BU). Each BU is staffed with a number of engineers from different specialisations. A single component is usually designed by a number of specialists. This creates an additional requirement on any information system in that there is now a requirement for a federated (multi-perspective) view of the same information space. That is, the system should be able to provide different contextual slices (views) through the same information space to present appropriate information to the several designers/engineers who often have distinct perspectives.

2. Development of the Demonstrator

Like many organisations in the early 1990s Rolls-Royce outsourced the management of their IT systems. This has greatly influenced the scope of the type of demonstrator that could be build, as it was not possible for AKT to integrate prototype systems quickly in order to investigating alternative designs, or to carry out software trials. As a result AKT has focussed on using Rolls-Royce's data and providing demonstrators to show how advanced knowledge technologies can be applied in an international manufacturing company.

The AKT project aims to demonstrate how advanced knowledge technologies can be use to present a federated perspective on the documentation relating to a specific component, that is to provide Intelligent Document Retrieval.

The federated representation (multi-perspective view) will be represented by different ontologies for each of the engineering specialisms:

- Designer,
- Stress engineer,
- Thermodynamics analyst,
- Methods specialist.

Rolls-Royce issues most design related documents from a document repository. Rolls-Royce also record the document type, authors, abstract, and keywords related to each document in a central database, for over 300,000 documents.

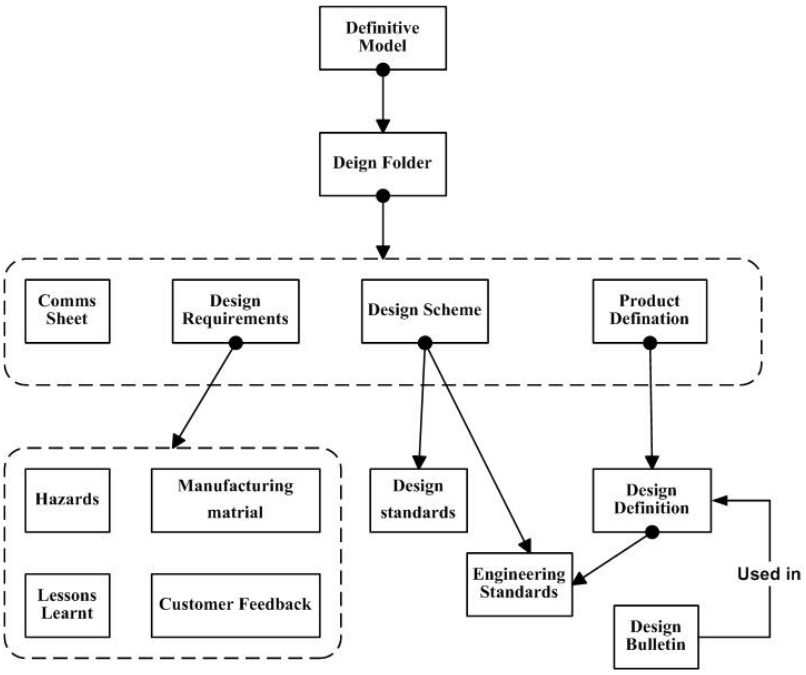

Figure 1 A example of the result of a card sort, detailing the documents referred during a partial task in the methods process.

2.1 Knowledge Acquisition

The first phase in building the demonstrator was to carry out a number of Knowledge Acquisition (KA) interviews at Rolls-Royce with carefully selected engineers (domain experts), from which a simple ontology of terms and concepts for each specialist was derived. During these interviews the 'card sort' technique was used to help the engineer show how they used different document types and the relationships between these documents [15]. In addition to the card sort, a number of semi-structured

interviews were carried out in order to understand the working environment and the difficulty encountered when designing a product.

The result of these interviews enabled the AKT team to identify, by specialism, the main concepts and the associated keyword for these concepts used by the particular type of engineer when searching for information. The interviews also enable the AKT team to produce the initial set of cards for the 'card sort' technique. Each expert was free to discard cards and to add any they felt was missing. A number of sessions were carried out with engineers from different specialisms within the same BU. Figure 1 represents the typical result from a card sort session.

2.2 Background technology

The term hypertext was first proposed by Ted Nelson in the late 1960s [7]. Nelson applied it to unstructured text, where associations between the text was made with links. Many '*closed*' hypermedia applications embed links into the structure of the document, i.e early and conventional world-wide-web pages. In contrast, an open hypermedia system can be defined as follows [9], [13]:

- A system which does not impose any mark-up upon the data that will prevent the data from being accessible to other processes that do not belong to the system.
- A system in which there is a separation of links from data objects.
- A system that can integrate with any tool that runs under the host operating system.
- A system in which data and processes may be distributed across a network and across hardware platforms.
- A system in which there is no artificial distinction between readers and authors.
- A system in which it is possible to add new functionality easily.

Within the open hypermedia philosophy, the hypermedia links are themselves a valuable store of knowledge. If this knowledge is bound too tightly to the documents, then it cannot be applied to new data. Therefore, no information about the links is held in the document data files in the form of mark-up. Instead, all data files remain in the native format of the application that created them, whilst the link information is held in link databases (linkbases). Research into open hypermedia has been undertaken at the University of Southampton since 1989. The philosophy of open hypermedia was instantiated in the development of a software package, called Microcosm, which is still being used in teaching and research today [13].

With the advent of the Web, the open hypermedia principles and the Microcosm philosophy were embodied in a system called the Distributed Link Service (DLS) [4]. At the heart of the DLS is a proxy Web-server. All requests to and from and Web pages sent to the user must all go through the proxy server (see Figure 2). The proxy server can then manipulate the HTML and add additional links to the content of the web pages. The philosophy of Microcosm is also acknowledged to have influenced the new open linking standard for the Web, X-Link [10].

Hypertext is just one example of the use of a family of techniques that are intended to transcend the limitations of static, sequential presentations of text [19]. Hypertext

uses computer effects (such as linking, indexing and interaction) to improve familiar textual communication for human beings; it is a form of human communication augmented by computer-manipulated media, databases and links. By contrast, the Semantic Web is an application of the World Wide Web aimed at computational agents, so that *programs,* and not just humans, can interpret the meaning of the information stored in the WWW hypertext [2]. The basis of this interpretation is an ontology, a structure which forms the backbone of the knowledge interpretation for an application.

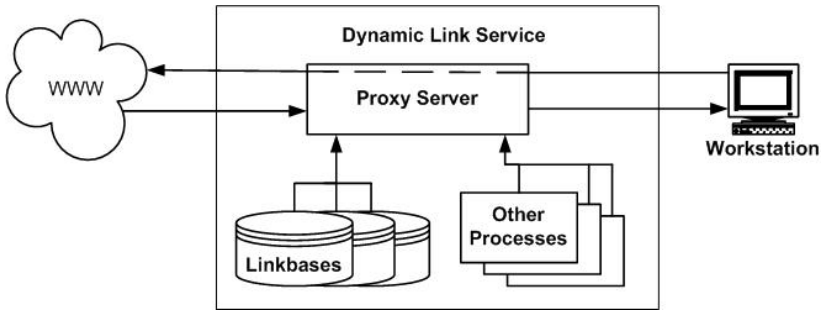

Figure 2 Dynamic Link Service

An ontology is "*a specification of a conceptualization*" [14]; Gruber explains that a common ontology defines the vocabulary with which queries and assertions are exchanged among agents (people or software). The ontology sets out all the entities (objects or concepts) that we are interested in and the relationships that connect these entities together. This is intended to be a *pragmatic* definition, i.e. it defines the vocabulary that is actually *in use*, and the concepts that are *useful* in problem-solving. It does not give the deep underlying philosophical vision of the fundamental entities in the field. Hence, in Knowledge Management (KM), an ontology is a tool, whose quality is entirely dependent on its usefulness.

Carr *et al* describes the use of ontologies to model and capture domain knowledge, and to provide a shared and commonly agreed understanding of a particular domain [5]. The COHSE project (Conceptual OHS Environment) produced an experimental ontological hypermedia system by combining an existing open hypermedia link service with an ontological reasoning service to enable documents to be linked via the concepts referred to in their contents. Previous attempts to improve the linking through simple lexical matching had serious limitations due to the uncontrolled method of adding links: many keywords turn up in many contexts and there is no simple lexical basis for discriminating important terms and significant links. The aim of the COHSE project therefore was to combine the OHS architecture with an ontological model to provide linking on the *concepts* that appear in Web pages, as opposed to linking on *simple uninterpretted text fragments*.

COHSE used a standard Web browser controlled by an adapted *link service* which in turn used three independent services to manipulate the exposed DOM of the Web page, resulting in the effect of ontologically-controlled hypertext.

An **ontology service** manages ontologies (sets of concepts related according to some schema) and answers specific queries about them. The **metadata service** annotates regions of a document with a concept, rather than the familiar case of annotating a document with a simple piece of text. The **resource service** is a simple librarian which is used to lookup Web pages which are examples of a particular concept (*i.e.* which can be used to *illustrate* a concept).

When a web page is loaded, the ontology service provides a complete listing of all the language terms that are used to represent the concepts in the relevant ontology. Each language term is searched for in the document, and, if found, its associated concept is looked up. Having identified the significant concepts in the document, the resource service provides a list of documents that are about instances of this concept.

At this point, a number of potential link anchors and destinations have been identified for the page and decisions can be taken about whether the document contains too many or too few links. In those circumstances, alternative links may be chosen from the broader or narrower concepts in the ontology in order to expand or cull the set of link anchors. The decisions about link culling and presentation are controlled by behaviour modules which define the navigation and interaction semantics of the resulting ontological hypertext.

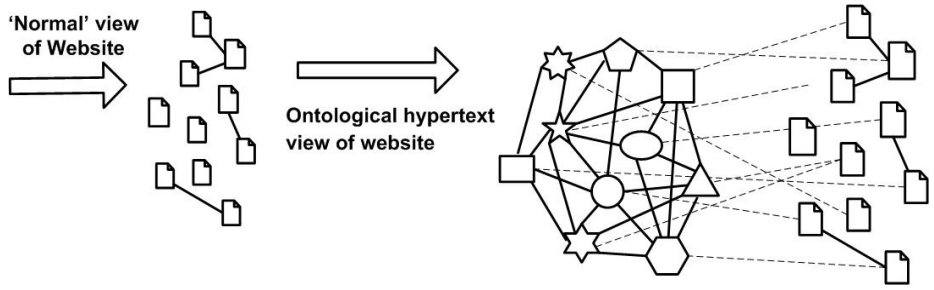

Figure 3 An ontology allows a weakly linked Website to be enriched through the meta-layer provided by the ontology.

The Ontoportal framework was initially developed in the Ontoportal project, and was used to build a web portal for the metadata research community, Metaportal [6]. It has now been extended to a generic application framework for building web portal applications based on domain ontologies. The Ontoportal system is the result of integrating ontologies, as conceptual models of knowledge, with hypertext to provide a powerful application environment. Which taking advantage of the various benefits of ontological hypertext and which also augments the breadth of hypertext with more meaningful links.

Once a domain ontology has been created, specific *ontoportals* are generated by populating the knowledge base with specific metadata, this process is referred to as instantiation of the ontology. That is, the resources have been identified as belonging to one or more of the concepts within the ontology. New *portals,* each representing a different view on the same set of resources, are generated by the Ontoportal framework from the instantiated domain ontology.

Woukeu *et al* have shown how ontologies and open hypermedia can be used to enrich a weakly linked resource, see Figure 3, using the Ontoporal framework [21].

2.3 Demonstrator

A web-based demonstrator was built that showed how, by using a simple ontology, appropriate documents can be retrieved from the document repository. The demonstrator used techniques developed from:

- The ontological hypertext system Conceptual Open Hypermedia Services Environment (COHSE)
- A framework for developing ontologically driven portals (Ontoportal) which allows different ontologies to be used on the same document set

The resulting list of document is ordered according to the engineering task and the related concepts (identified by keywords).

Depending on the designers main responsibility (Stress analysis, Thermodynamics, manufacturing Methods, or Design, see Figure 4) being undertaking, they are presented with appropriate job functions from the ontology, see Figure 5.

The AKT team also recognized that sometimes people found a particular document extremely useful for a certain task, so the system allowed Engineers to recommend certain documents to their colleagues. This can be considered as another form of knowledge elicitation. Similarly it was recognized that the engineers may also want to browse the key words when they are hunting for information. Therefore the system allows engineers to look through all the key words regardless of design function. The search could be further refined by looking for the occurrence of the concept (keyword) in any combination of Title, Abstract and Keyword fields. Once the engineer has chosen the job function, as in the case of Figure 4 and Figure 5 the role is *Design* and the function is *Input requirement*s, a list of tasks is returned, Figure 6.

By selecting a task, the system will look-up the appropriate concepts (keywords) and returns a list of appropriate documents (title, author and date) from the database. The documents are returned by document type and the order of the document types is governed by the results of the card sort obtained during the knowledge acquisition phase. That is the document types the engineers most commonly used to undertake the task. In addition to ensure that information is not missed documents types not identified in the card sort are listed by type alphabetically at the end. In addition peer recommended documentation is listed first. The engineer can get further information on any document by clicking on the document in the list and the full record including the abstract/summary is returned.

Figure 4 The engineer selects the role they are undertaking.

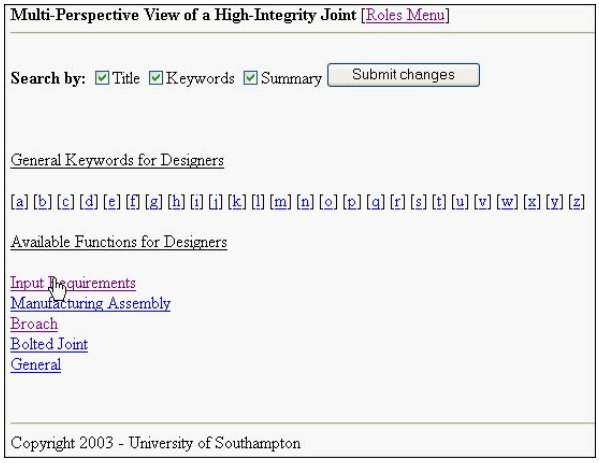

Figure 5 The engineer can choose which function to follow when undertaking the design role.

Occasionally, more documents are returned than is practical to browse through. Therefore it is necessary to cull the list in a meaningful way; the method chosen is to present the engineer with an interactive image of the card sort, see Figure 7, which allows the engineer to view the returned list by just one document type at a time and then switch to another type. By using the pictorial representation of the card sort the engineer can browse the results with some understanding as how the document type relates their role or their perceived role.

Figure 6 The engineer selects a requirements task associated with the design role.

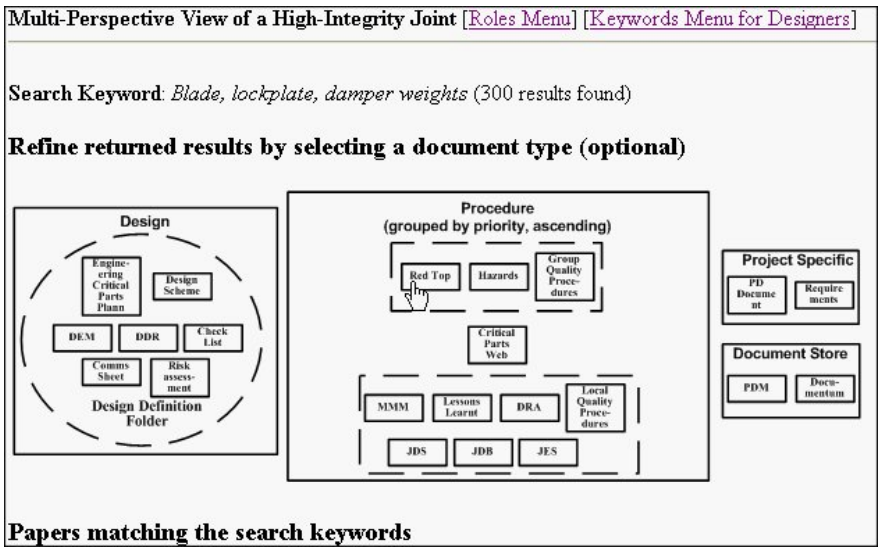

Figure 7 Interactive image of card sort to filter the large number of documents returned by type.

3. Lessons learnt

As with many knowledge management projects, the knowledge acquisition phase took a considerable time. We were aided in this process by having a team that con-

sisted of mechanical engineers (who acted as domain experts), computer scientist and knowledge management engineers.

One of the aims of the system was to provide intelligent document retrieval, for example retrieval based more on the semantics than simply keywords. In many ways this can be likened to data mining [3], in that to facilitate this knowledge mining it was important to ensure that there is a clearly defined and well structured warehouse. This was achieved by using ontologies to give the warehouse structure and a triple store to hold the content [20].

Intelligent document retrieval relies on machine readable data that is fairly clean (nothing is ever perfect and the system must allow for this). Hence, it is important to maintain information discipline throughout the organization. This is partly a social issue as well as an engineering matter; people often do not understand the importance of it or see it as their responsibility to ensure that the data entered is correct. From an engineering perspective a system can be designed to help maintain the discipline. For instance, analysis of the database showed that the *document type field* was a free text field, hence many documents types had misspelt descriptors; a simple drop down list would have been sufficient in order to maintain the consistency of the database.

A similar situation arose with the keyword field, while it was understandable to have a field that allows the author or authors to enter free keywords, it needs to be coupled with a taxonomy or a list of higher level keywords to help give the document some context within the organization. There are several approaches that can be used here from using an ontology to using existing taxonomies from other professional institutions or library classifications. Again discipline is required in any free text field, for instance analysis of the keyword field showed that only a minority of fields had keywords or phrases, most had sentences. Hence a semi-automatic process was applied to produce sensible keywords (a set of 9000 in total) from the meaningless information held in the keyword field.

Some of the above problems encountered could have been alleviated if much of the metadata for a document did not have to be manually entered in the database. Most if not all the information could be 'stripped out' of document automatically as most of the information is on the front sheet of the document. This was achieved practically by using document templates.

It is important to provide a controlled but flexible means for people to share their knowledge eg through a company wide intranet. As this is not yet available many of the BU's have their own website, which is used to collate knowledge on how to carry out the design. These pages are manually maintained and all links are button type links. Although only a few years old these several intranet are growing larger by the day and already are becoming unmanageable; dangling links are appearing [8], with knowledge being lost as pages disappear into hyperspace never to be found again.

4. Conclusions and Future Work

To help overcome the difficulty engineers find locating the correct information whilst designing components for the aerospace industry, a system was developed that demonstrated how knowledge technologies can be used to help solve this problem. While

the system was only a demonstrator, feedback from the focus group was very encouraging; we had to frequently remind them that this was a demonstrator and it could not be brought into service next week. There are considerable benefits to the organisation [17], other that the engineer getting less frustrated at not being able to locate the information they know is there. We discovered during the KA interviews that experiments and associated costs (design and build of test rigs) were being repeated as engineers did not know that similar experiments had been carried out a few years earlier.

The lessons learnt from developing this system are now to be applied to the next system; a semantic web service environment [2][11]. As many of the Business Units now have their own website, we believe that semantically enriching the metadata of the Web pages and documents held on these sites, would provide the organisation with a method of improving the interconnectivity of the knowledge. Thereby stopping the growth in small disconnected islands of knowledge within the organisation. In addition the ontologies developed will aid automatic linking of the concepts held in the documents and Web pages.

The engineers require a flexible and easy method of capturing knowledge and sharing this with colleagues. This was the reason given for the development of these group websites. It is our aim in the next phase to develop Intelligent Editors to provide tools to support the capture, reuse and maintenance of this knowledge, For instance the Designers Workbench [11]. Fundamental to this being achieved is the warehousing of the knowledge, using the ontologies and structure, knowledgebases as the containers and a set of tools that will:

- Provide a means by which engineers can refine the information extraction techniques and retrieval mechanism, with respect to an ontology for a Business Unit or by specialisation.
- Provide searchable semantically enriched meta-data for each BU's intranet pages.
- Aid the maintenance of the ontologies.
- Allow Designers to customize their User Models.

5. Acknowledgements

The author thanks Rolls-Royce engineers Gary Nicholson and Tamsyn Thorpe, for participating, advice and organizing the interviews.

6. References

[1] AKT Manifesto, http://www.aktors.org/publications/

[2] Berners-Lee, T., Hendler, J., Lassila, O. (2001) The Semantic Web, *Scientific American, May 2001 34-43*

[3] Berson A, Smith S, thearling K. (2000) Building Data Mining Applications for CRM, McGraw-Hill ISBN 007134446

[4] Carr, Les A. and De Roure, David C. and Hall, Wendy and Hill, Gary J. (1995) The Distributed Link Service: A Tool for Publishers, Authors and Readers. *In Proceedings Fourth International World Wide Web Conference: The Web Revolution, Boston, Massachusetts, US).*, pages 647--656.

[5] Carr L., Bechhofer S., Goble C., Hall W. (2001)Conceptual Linking: Ontology-based Open Hypermedia. *In Proceedings WWW10, Hong-Kong, May 2001.*

[6] Carr. L, Kampa S, and Miles-Board T. (2001) MetaPortal Final report: Building Ontological Hypermedia with the Ontoportal Framework. Technical Report,ECSTR-01-005 University of Southampton, May 2001 ISBN 085432 7371

[7] Conklin J. Hypertext: An Introduction and survey. *IEEE Computing Vol. 20, 17-41, 1987*

[8] Davis HC. Data integrity problems in an Open Hypermedia Link . *PhD Thesis University of Southampton November 1995*.

[9] Davis, HC, Knight, S. and Hall, W.. Light Hypermedia Link Services: A Study of Third Party Application Integration. *Proceedings of the ACM European Conference on Hypermedia Technology. ECHT'94, Edinbough, Scotland, September 18-23, 1994, pp 41-45*

[10] DeRose S, Maler E, Orchard D. 2001 XML Linking Language (XLink) Version 1.0 W3C Recommendation 27 June 2001,

[11] de Roure, D., Jennings, N. R. and Shadbolt, N. (2003) *The Semantic Grid: A future e-Science infrastructure*, in Berman, F., Fox, G. and Hey, A. J. G., Eds. *Grid Computing - Making the Global Infrastructure a Reality*, pages pp. 437-470. John Wiley and Sons Ltd.

[12] Fowler D, Sleeman D, Wills G, Lyon T, Knott D. (2004) The Designers' Workbench: Using Ontologies and Constraints for Configuration. *AI-2004. The Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence Queens' College, Cambridge, UK 13th-15th December 2004*

[13] Fountain AM, Hall W, Heath I, Hill G. MICROCOSM: An Open Model for Hypermedia with Dynamic Linking. *In Rizk A, Streitz N & Andre J. eds. Hypertext: Concepts, Systems and applications. The Proceedings of the European Conference on Hypertext, INRIA, France, pp 298-311. Cambridge University Press 1990.*

[14] Gruber T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5(2):199-220*

[15] Kelly G, (1955)The Psychology of Personal Constructs, W.W. Norton, New York, 1955

[16] Malcolm K., Poltrock S., Schuler D. Industrial Strength Hypermedia: Requirements for a Large Engineering Enterprise. *In Hypertext '91. The Third ACM Conference on Hypertext, San Antonio, Texas, December 1991, pp 13-24. ACM Press 1991*

[17] Mika P, Iosif V, Sure Y, Akkermans H. (2004) Handbook on Ontologies, Eds. Staab S, Studer R. Springer-Verlag.

[18] Marinheiro R M N, Hall W. "Expanding a Hypertext Information Retrieval System to Incorporate Multimedia Information", *Proceedings of the 31st Annual Hawaii International Conference on System Sciences, Vol. II, pp 286-295, 6-9 January, 1998, IEEE Computer Society.*

[19] Nelson, T. (1987). Literary Machines. 87.1 edn. Computer Books.

[20] schraefel, m. c., Shadbolt, N. R., Gibbins, N., Glaser, H. and Harris, S. (2004) CS AKTive Space: Representing Computer Science in the Semantic Web. In *Proceedings of World Wide Web Conference 2004* .

[21] Woukeu A, Wills G, Conole G, Carr L, Kampa S, Hall W. Ontological Hypermedia in Education: A framework for building web-based educational portals. *Ed-Media, Hawaii 23-28 June 2003*

[22] Wills GB, Sim YW, Crowder RM, Hall W (2002) Open Hypermedia for Product Support *International Journal of Systems Science* 33(6):421-432