IST-2001-33127

SciX

Open, self organising repository for scientific information exchange

D8: Technology: market watch, state of the art and requirements analysis

Responsible authors: Brian Clifton and Grahame Cooper (USAL)
Co-authors: Bo-Christer Bjork(SHH) , Tomo Cerovsek (FGG), , Almudena Fernandez Valero(INDRA), Gudni Gudnarson (IBRI), Turid Hedland (SHH), Bob Martens (TUW), Etiel Petrinja (LJU), Ziga Turk (LJU)

Access: public
Version: 1.0
Date: 31-Oct-2002

EXECUTIVE SUMMARY:

The presented report documents the baseline for the development of the SciX pilot software. It draws from the business process analysis (D1) and includes:

- Scix-related projects and research;
- state-of-the-art in electronic publishing;
- technology overview addressing core web services technologies, knowledge management, community building and security.
- overview of formal and industry standard related to the content of digital archives (metadata and full content).
- requirements analysis - what different players actually need and expect from electronic publishing and digital archiving.

It has been confirmed that that numerous standards and tools exist, some also available for free, that allow for building the pilot applications. XML based technologies are emerging as dominant for any kind of integration of several services. More specifically, the OAI Metadata Harvesting protocol and the Dublin Core metadata standards need to be supported.

Based on this corpus of knowledge, the business process analysis (D1) and the Content acquisition techniques (D5), the system architecture will be proposed in the deliverable D9.

RELEASE HISTORY

date	changes
22.7.2002	Outline of the document
9.8.2002	Initial draft
23.9.2002	Updated draft, based on discussion at the project meeting
16.10.2002	Final version pending some editorial polishing
31.10.2002	Version 1.0 released

TABLE OF ABBREVIATIONS

see Appendix 2.

TABLE OF CONTENTS:

1 INTRODUCTION

This document is Deliverable 08 of the SciX project. Its purpose is to investigate the existing systems which are available in the field of Electronic Publishing(EP). Secondly it provides a review of new and existing technologies that could be used in the development of the new system.

This report is structured into several sections:

- Section 2 reports on the projects that were found to be in some way related to the SciX project. Each is briefly presented and then its relevance and possible collaboration opportunity with SciX is addressed.
- Section 3 first presents the market watch- the existing free and commercial systems that deal with digital library issues. It then elaborates on the various business models for digital content. A special section deals with the content syndication, which is addressed in the WP5. The open source movement is addressed as separate phenomena.
- Section 4 presents the state of the art overview of the core technologies. Its main focus are the XML and related technologies, such as the UDDI and WSDL, as well as other component technologies.
- Section 5 addressing the community building tools. Digital content has a potential of building a whole community around it, much greater than a paper based content.
- Section 6 is looking at the whole problem of digital publishing through a knowledge management perspective.
- Section 7 is addressing the security issues.
- Section 8 addresses the repository content - the metadata standards, citation standard and tools, syndication standards etc. It is also looking at potential full text formats.
- Section 9 presents the requirement analysis of the key stakeholders of digital libraries - authors, readers, industry, journals, reviewers, librarians, service providers, etc.

The scope covered by this document is huge. It compiles the base knowledge required to build intelligent, self-managed digital libraries. Rather than including very detailed information and a few topics, the document is rather broad, encyclopaedic, and points the reader to numerous web resources for further reference.

We would therefore like to emphasize that this document is by its very nature an evolving document. The pace of technological change in the IT Sector is rapidly increasing and is changing on a monthly basis. Therefore, although the contents of this document reflect the state of the art as of October 2002, the reader should be aware that advances may have been made which render some of this information dated.

2 RELATED PROJECTS

Both in the US and in Europe several projects and initiatives are working on issues that are of relevance to the SciX project and its technical base. Each project is presented using the information on its Website. The last paragraph addresses the possible relation to SciX.

2.1 FIGARO

FIGARO (http://www.figaro-europe.net/index.html) is a European academic e-publishing initiative focused on the creation of an effective and affordable communication and publishing environment for scholars. It is intended as an information source for anyone and everyone who has an interest in the development of scholarly communication and electronic publishing.

The goals and baselines of Figaro are very similar to SciX with Figaro focusing more on the content creation workflow and SciX more on the business process modeling, content syndication and low barrier archive creation.

2.2 METAE - THE METADATA ENGINE PROJECT

METAe (http://meta-e.uibk.ac.at/) will develop application software focusing on the automatic recognition and extraction of metadata from printed material, especially books and journals, an omnifont OCR-engine for the recognition of "Fraktur" (a German style of black-letter text type) and other seldom type faces used in European printing history; the development of five historical dictionaries supporting the OCR-engine, an XML/SGML search engine and a open source library for a simple web-application for presenting digitised printed material.
Of particular interest to SciX and the content acquisition works is the automatic recognition and extraction of metadata from printed material, especially books and journals. SciX is well suited to test this tool in our engineering context.

2.3 TIPS - TOOLS FOR INNOVATIVE PUBLISHING IN SCIENCE

In the TIPS project (http://tips.sissa.it/) they "propose a new approach to scientific information production and dissemination. In this approach, a set of user-friendly and advanced tools and services are organized in an open system to support research information production, management, access, and use in a coherent manner. As an implementation and to make possible the evaluation of the system, these tools and services will be integrated on a web-based portal for the high-energy physics community.

The proposed system will support the activities of document writing, reviewing, publishing, searching, disseminating and reading, as well as the communication among members of the research community. The effectiveness of the implemented system will be demonstrated by an experimental evaluation. This approach is suitable for supporting a more productive research community, in which researchers can work in a more effective, inexpensive, and pleasant way: delays and costs due to paper documents can be considerably reduced, multimedia can be added to electronic documents, information access can be improved (and information overload decreased) by using advanced information retrieval and filtering techniques".

The TIPS project seems to nicely complement SciX, which is not tackling the issue of authoring and of the document formats. One could envision an architecture where the TIPS solution is used in the editing process while SciX is used to support the editorial workflow, storage, retrieval and organization of scientific archives.

2.4 KEPT - KNOWLEDGE E-PUBLISHING TOOLS

The aim of KePT Project is to develop a software package for publishing complex knowledge bases on the Internet, integrating innovative visualisation and navigation tools, based on the recent XML/TopicMap ISO standard. This is expected to provide innovative, highly intuitive visualisation and navigation tools on the web. The final tool will primarily address the publishing industry interested in promoting digital content on the Internet and many companies in ICT that aims at developing leading edge e-business applications.

SciX could use the KePT tool to provide an intuitive graphical interface to the topics in its repository.

2.5 OPEN ARCHIVES FORUM

The Open Archives Forum (http://www.oaforum.org/) provides a Europe-based focus for dissemination of information about European activity related to open archives and, in particular, to the Open Archives Initiative (see section 5.6). The aim of the Forum is to facilitate clustering of IST projects, national initiatives and other parties interested in the open archives approach. In order to do so, the Forum brings interested parties together to build a community of interest, enable exchange of information and establish a web-based European information source for open archives. In addition, the Forum undertakes comparative reviews of technical and organizational issues.

The particular relevance of this project is that SciX will most likely adopt the Open Archives Initiative standards for the service interfaces of its pilot software. SciX has subited its information to the OAF website.

2.6 ARION -AN ADVANCED LIGHTWEIGHT ARCHITECTURE FOR ACCESSING SCIENTIFIC COLLECTIONS

ARION (http://dlforum.external.forth.gr:8080/) is aiming to provide a new generation of Digital Library services for the searching and retrieval of digital scientific collections that reside within research and consultancy organisations. These collections contain data, programs and tools in various scientific areas and incorporate applications of different domains of knowledge.

ARION is focusing on scientific data (e.g. temperatures of the sea over the last century) and is therefore complementary to SciX, which is looking at scientific papers. They have developed a Workflow editor. Its possible use in SciX needs further study.

2.7 C WEB - COMMUNITY WEBS

The C-Web project (http://cweb.inria.fr/) was running in the year 2000 and aimed at designing a generic platform based on open standards, and the related methodology and know-how, to support community-webs structured along domain-specific ontologies, i.e.: formal representations of the knowledge shared by professionals working in a specific domain. It also aims at validating both the standards, the software platform and the methodology through experiments carried-out with several different user groups.

A limited number of C Web results was found on the Web. Contribution in the area of repository structuring through ontologies seems particularly strong. Work available on the community aspects which could be very beneficial to SciX.
(also refer to section 9 Information/Knowledge Management)

2.8 PSI3 - PERSONALISED SERVICES FOR INTEGRATED INTERNET INFORMATION

The main objectives of the project (http://www.psi3.org/) are:

- to specify a generic architecture for the development of personalized services for integrated Internet information;
- to develop generic components for such an architecture and to standardize the interfaces between these components;
- to search information through Internet by different approaches, and to evaluate mobile agents technology against more traditional robots-based technology.
- to specify and implement three personalized online services as pilot applications based on the generic architecture, namely an information service for the business sector "building and construction", an advanced Internet information search engine and an e-commerce application for e-learning.
- to evaluate the generic architecture and the pilot applications in user trials.

The interesting aspect of PSI is in the personalization, i.e. allowing the user to specify individual preferences useful for providing him / her with customized relevant information. The project is scheduled to end in 2002 and the results will be closely monitored.

2.9 OAI - OPEN ARCHIVE INITIATIVE

The Open Archives Initiative (http://www.openarchives.org) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. The fundamental technological framework and standards that are developing to support this work are, however, independent of the both the type of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials.

Support for Open Archives Initiative activities comes from the Digital Library Federation, Coalition for Networked Information, and National Science Foundation Grant No. IIS-9817416 (Project Prism). Participation in OAI has two dimensions:

- *Data Providers* A data provider maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata.
- *Service Providers* A service provider issues OAI-PMH requests to data providers and uses the metadata as a basis for building value-added services.

The current OAI technical infrastructure, which is specified in the Open Archives Initiative Protocol for Metadata Harvesting, defines a mechanism for data providers to expose their metadata. There is nothing in the OAI mission that restricts the work of the OAI to metadata alone. However, we are guided by the goal to define a low-barrier and widely applicable framework for cross-repository interoperability and believe that exposing metadata is plausible route to such a goal. We may, in the future, explore and define other mechanisms for interoperability.

2.9.1 OPEN ARCHIVES INITIATIVE PROTOCOL FOR METADATA HARVESTING - OAI-PMH

This defines a mechanism for harvesting XML-formatted metadata from repositories. The protocol does not provide a mechanism for harvesting data (content) that is not encoded in XML. The protocol also does not mandate the means of association between that metadata and related content. Since many clients may want to access the content associated with harvested metadata, data providers may consider it appropriate to define a link in the metadata to the content. The mandatory Dublin Core format(section 5.1.1) provides the *identifier* element that can be used for this purpose.

2.9.2 OAI AND THE DUBLIN CORE

Mapping among multiple metadata formats would place a considerable burden on service providers, who harvest the metadata and use it to build higher level services. While there is research work on creating services such as common search interfaces across heterogeneous metadata formats, a less burdensome and ultimately more deployable solution is to require repositories to map to a simple and common metadata format. The fifteen elements Dublin Core (see section 5.1.1) has over the past several years evolved as a de facto standard for simple cross-discipline metadata and is thus the appropriate choice for a common metadata set.

2.9.3 OAI AND INTELLECTUAL PROPERTY RIGHTS (IPR)

The OAI does not define or prescribe any rights management scheme. Issues of access restriction and management of intellectual property in exposed metadata are the responsibility of the data providers that adopt the protocol.

2.9.4 OAI AND JISC/ROMEO

http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html

The Joint Information Systems Committee (JISC) is a strategic advisory committee working on behalf of the funding bodies for further and higher education (FE and HE) in England, Scotland, Wales and Northern Ireland. It also works in partnership with the Research Councils.

The JISC promotes the innovative application and use of information systems and information technology education across the UK by providing vision and leadership and funding the network infrastructure, Information and Communications Technology (ICT) and information services, development projects and high quality materials for education. Its central role ensures that the uptake of new technologies and methods is cost-effective, comprehensive and well focused.

2.9.4.1 The RoMEO Project - Rights MEtadata for Open archiving

is funded by JISC for one year (1 August 2002 - 31 July 2003) to investigate the rights issues surrounding the 'self-archiving' of research in the UK academic community under the Open Archive Initiative's Protocol for Metadata Harvesting.

It will perform a series of stakeholder surveys to ascertain how 'give-away' research literature (and metadata) is used, and how it should be protected. Building on existing schemas and vocabularies (such as Open Digital Rights Language) a series of rights elements will be developed. A demonstrator system will then be created to show how rights metadata might be assigned, disclosed, harvested, and displayed to end users via the OAI Protocol for Metadata Harvesting.

This is relevant to sciX in terms of the IPR status of articles stored. The progress of this project should be carefully monitored.

2.10 ORIEL

The Online Research Information Environment for the Life Sciences is a project also funded by the IST Programme and coordinated by the European Molecular Biology Organization (EMBO). Oriel aims to provide research communities with tools to manage large, complex, and disparate digital information resources. With a view to making such technologies widely available, it will focus on the Life Sciences as a data-intensive and highly demanding testbed that will permit effective linking of different types of biological information displaying complex inter-relationships (literature, factual and multi-media databases), promote ease of navigation leading to creative exploration of the information landscape and facilitate user-friendly data presentation and information visualisation.

2.11 CYCLADES - AN OPEN COLLABORATIVE VIRTUAL ARCHIVE ENVIRONMENT

The main objective of CYCLADES (http://www.iei.pi.cnr.it/cyclades/) is to develop advanced Internet accessible mediator services to support scholars both individually and as members of networked communities when interacting with large interdisciplinary electronic (e-print) archives. Such archives are important vehicles for the dissemination of preliminary results and non-peer reviewed "grey literature". Most focus on information dissemination within disciplinary or institutional communities. However, scientific research is now oriented towards an interdisciplinary approach. Scientists thus need to easily retrieve information from diverse sources, and to communicate and collaborate across traditional community boundaries. CYCLADES aims at supporting the transition of e-print systems into genuine building blocks of a transformed scholarly communication model by developing a set of leading edge technologies providing innovative methods for information access, dissemination, sharing and collaborative work.

CYCLADES will base the development of the service environment on these specifications. In particular, a core set of cross-archive value-added services will be developed to constitute a federation of independent but interoperable services. According to this approach, a service provides a functionality and can either work independently or can communicate and collaborate with other services to offer a new value-added service. The Service Environment will provide OAi compliant functionality.

This is a very similar approach to what is planned in the pilot of the SciX project. The results of this work need to be monitored very closely. Current verion of the Website is not very informative.

2.12 NETWORKS

2.12.1 SCHOLNET

SCHOLNET aims at developing a digital library infrastructure to support the communication and the collaboration within networked scholarly communities. The digital library will provide traditional digital library services in addition to support for non-textual data types, hypermedia annotation, cross-language search and retrieval, and personalized information dissemination. This testbed will be used to demonstrate how an enhanced digital library can enable members of a networked scholarly community to learn from, contribute to, and collectively build upon the community's discipline-oriented digital collections.

Scholnet is another project whose goals and objectives are similar to those of SciX. Since the project was supposed to end in 2002 its findings could be used in SciX. Several deliverables, however, were not available on the Website.

2.12.2 RENARDUS

The Renardus service aims to provide a trusted source of selected, high quality Internet resources for those teaching, learning and researching in higher education in Europe. Renardus provides integrated search and browse access to records from individual participating subject gateway services (data providers) across Europe.

The Renardus service grew from a project funded 1 January 2000-30 June 2002 by the EU's Information Society Technologies 5th framework programme. Renardus exploits the success of subject gateways, where subject experts select quality resources for their users, usually within the academic and research communities. This approach has been shown to provide a high quality and valued service, but encounters problems with the ever increasing number of resources available on the Internet. Renardus is based on a distributed model where major subject gateway services across Europe can be searched and browsed together through a single interface provided by the Renardus broker.

SciX pilots should be compatible with services such as Renardus.

2.13 EPRINTS SOFTWARE

2.13.1 ERCIM TECHNICAL REFERENCE DIGITAL LIBRARY

The ERCIM (www.ercim.org) Technical Reference Digital Library (ETRDL) is a digital library service which has been set up to assist the scientists of the European Research Consortium for Informatics and Mathematics (ERCIM), to rapidly access, manage and disseminate technical reports and other reference material in the IT domain.
The software that supports this service could be similar to what we propose to implement in SciX, however, it does not seem to be placed under an open source license. Software is based on the "Dienst" protocol which is a forerunner of the Open Archives Harvesting protocol.

2.13.2 EPRINTS.ORG

EPrints (www.eprints.org) software has been created so that institutions can create OAI-compliant Archives quickly, easily and for free. OAI-compliance means all Archives created in this way are "interoperable." They use the same (OAI) convention for tagging their metadata (author, title, date, journal, etc.). That means the contents of all such Archives can be harvested integrated, navigated and searched seamlessly, as if they were all in one global "virtual" archive.

The primary purpose of the EPrints software is to help create open access to the peer-reviewed research output of all scholarly and scientific research institutions (mainly universities). Maximizing the access to research findings maximizes their usage and their impact on further research, to the benefit of researchers, theit institutions, the society that supports research, and to research itself. The EPrints software was designed primarily to be used by researchers and their institutions to maximize the access to -- and hence the impact of -- their research output.

While using the ePrints software is an option for SciX, some existing code base is believed to allow for advanced architectures and feature lists. Some related efforts such as the Open Citation Project would complement SciX.

2.14 HUBS AND PORTALS

In the previous sections only the projects found most relevant to SciX were presented. The services mentioned in this section are hubs to hundreds of others.

2.14.1 DIGICULT

Digital Culture (DigiCULT - www.digicult.info) is an IST Support Measure (IST-2001-34898) to establish a regular technology watch for cultural and scientific heritage over the period of 30 months (03/2002-08/2004). It is publishing reports, newsletters, thematic issues and technology watch reports related to the digital libraries with a special focus on cultural content.

2.14.2 DIGITAL LIBRARY FORUM

Digital library forum is another hub for digital library information - mainly for the German speaking audeince. SciX is present in their database.

2.14.3 EL.PUB - ELECTRONIC PUBLISHING R&D NEWS AND RESOURCES

The El.pub web site (http://www.elpub.org/) is published by the IST INFORM project and provides a focal point for news and resources about research and developments (R&D) in interactive electronic publishing. The effort is continued in the Epifocal project.

Their service provided important pointers in learning about related work in this area.

2.15 DATA MINING AND TEXT MINING PROJECTS

Traditionally, there have been two types of statistical analyses: *confirmatory analysis* and *exploratory analysis*. In confirmatory analysis, one has a hypothesis and either confirms or refutes it. However, the bottleneck for confirmatory analysis is the shortage of hypotheses on the part of the analyst. In "*exploratory analysis*", (Tukey, 1973), one finds suitable hypotheses to confirm or refute. Here the system takes the initiative in data analysis, not the user.

Data mining is the process of finding patterns in the data and is relevant to both mentioned techniques. There are different functions of data mining:

- sequence analysis for time dependent data,
- link analysis which tries to determine relations between the data,
- summarization which describes subsets of the datasets by computing the median and standard deviation,
- classification which map datasets to one or more predefined classes and
- cluster analysis which, similar to classification, groups datasets into clusters, by means of similarity metrics

- The process of clustering is applied to collection of documents or records. Following parameters represents the content of collection:
- Number of all records in whole collection
- Number of stem terms in whole collection
- Average term usage in whole collection
- Average term vector length
- Standard deviation of vector length
- Average frequency of terms in vector
- % of terms with freq. 1

Information retrieval techniques and algorithms are usually tested on large reference collections like TREC, CACM and ISI.

Various AI technologies for analysing text databases are known since the late 1970s (van Rijsbergen 1979). They became particularly popular after the explosion of the World Wide Web, when the search engines were looking for the technologies to increase the relevance of the searches (Zamir and Etizioni, 1999) or build some intelligence into Web browsing (Mladenic, 1999). An example of an "intelligent" interface to bibliographic data is for example the www.researchindex.com. Several Websites implement such technology, for example AltaVista and Google (Bring and Lawrence, 1998). Concerning sciX IR (information retrieval) issues there are several related research projects. Some of them are listed below in alphabetical order:

2.15.1 BAILANDO

Berkeley's BAILANDO series offer several solutions to rather specialized issues in IR. Especially interesting is for example LINDI (Linking Information for Novel Discovery and Insight). Major aims of LINDI project were: (1) to investigate how researchers can use large text collections in the discovery of new important information, and (2) to build software systems to help support this process. The research developments were primarily focused on information retrieval techniques covering following operations:

- Iteration of an operation over the items within a set. (This allows each item retrieved in a previous query to be used as a search terms for a new query.)
- Transformation, i.e., applying an operation to an item and returning a transformed item (such as extracting a feature).
- Ranking, i.e., applying an operation to a set of items and returning a (possibly) reordered set of items with the same cardinality.
- Selection, i.e., applying an operation to a set of items and returning a (possibly) reordered set of items with the same or smaller cardinality.

2.15.2 CORA

Cora is a special-purpose search engine covering computer science research papers. It allows keyword searches over the partial text of Postscript-formatted papers it has found by spidering the Web. Cora provides access to over 50,000 research papers on all computer science subjects. The construction of Cora has been greatly automated by the use of artificial intelligence and machine learning techniques. Efficient topic-directed spidering is performed using

reinforcement learning; papers (in Postscript format) are automatically categorized into the topic hierarchy by probabilistic techniques; papers' titles, authors, references, etc, are automatically extracted using hidden Markov models. Basically the software provides a means for automated creation and maintaining portals by the use of machine learning techniques. A detailed overview can be found in (McCallum, 2000)

2.15.3 DESIRE

Development of a **E**uropean **S**ervice for **I**nformation on **R**esearch and **E**ducation was a EU-funded Project in two phases (1996-1998 and 1998-2000). One part of the work, done by NetLab of Lund University Library (http://www.lub.lu.se/), especially by Traugott Koch, was to evaluate automatic classification and compare and combine these techniques with manual classification. The subject chosen was engineering and as basis for both, the manual and the automatic classification, the Ei thesaurus was used (http://www.ei.org). Data sources were HTML files and files with record-syntax format used in the Combine Harverster (http://www.lub.lu.se/combine). The pre-processing included weightings according to the location and deletion of stop words, etc. and optionally stemming with the Porters stemming algorithm. Then a simpler matching algorithm was used to compare the words of the texts with the words of the thesaurus. The final weighting of the results is done by term complexity/classification type, match location (metadata, headings, other text) and matching frequency.

2.15.4 GERHARD

GERman **H**arvest **A**utomated **R**etrieval and **D**irectory is a research project of Bibliotheks- und Informationssystem (BIS, http://www.bis.uni-oldenburg.de/) of the University of Oldenburg, Institut für Semantische Informationsverarbeitung of the University of Osnabrück (ISIV, http://www.isiv.uni-osnabrueck.de/) and Oldenburger Forschungs- und Entwicklungsinstitut für Informatik-Werkzeuge und -Systeme of the University of Oldenburg (OFFIS, http://www.offis.uni-oldenburg.de/) funded by the Deutsche Forschungsgemeinschaft (DFG). The project web-site is: http://www.gerhard.de.

GERHARD created a HARVEST based robot-generated index of Web resources in Germany using indexing and automatic classification and provides a search and a hierarchical browsing facility based on the Universal Decimal Classification (UDC) which is maintained electronically at the Library of the ETH Zürich. The UDC is three-lingual (English, German, French) and includes about 60.000 entries with 13 different relations. The automatic classification works in three steps:

- Preparation of the UDC, basically the creation of a thesaurus from UDC via deletion of stop words and morphological reduction of the words to their stems.
- Preparation of the HTML documents, deletion of stop words.
- Analysis of the notations and sorting the most relevant according to the UDC hierarchy. This is done twofold: one for the whole text and one only for the title. The later computation gets a higher relevance level. In average one document gets 6-7 UDC notations. The statistical analysis additionally relies on specific information about the UDC hierarchy which is stored in a database.

The software is implemented in C and Perl. The classification module is a client/server implementation in C. The UDC data is stored in an relational database. With the exception of a program library for stemming, which belongs to a company Lingsoft from Helsinki, the code is available for researchers. Currently GERHARD provides access to 1,284,819 documents with 6,182,891 relations between them. For more detailed information on GERHARD see (Wätjen et all, 1998).

2.15.5 OASIS

Open **A**rchitecture **S**erver for **I**nformation **S**earch and Delivery was a EU-funded project, the Consortium of which consisted University College Dublin, Ireland, St.-Petersburg State University, Russia, University of Tübingen, Germany, JSC Peterlink, St.- Petersburg, Russia, Valtek Ltd., Kiev, Ukraine and DSI Ltd., Irkutsk, Russia.

The OASIS software is intended to perform intelligent information search and delivery service based on artificial neural network techniques and methods. The cluster analysis method is called Hierarchical Radius based Competitive Learning (HRCL) and used a multidimensional neural network that is self learning and able to find clusters on different hierarchical levels. It thus is able to create an Internet catalogue automatically. The clustering is not dependent on the sequence of the analysed documents. No classification system or thesaurus is needed, since it will be created automatically. Since more than two dimensions are used it is not possible to visualize the relations of the cluster in forms of a map. OASIS system uses relevance feedback from a user not only to refine the user's query, but to refine its internal algorithms also. It can accumulate query and relevance feedback statistics, analyse it and use results of this analysis to increase the precision of search. LDAP is used for the storage of metadata information. The OASIS search engine is distributed. Several OASIS servers are connected via CORBA and store information about different document collections. The server perform a chaining of search requests transparently for the client. This intelligent forwarding is done via LDAP. For the merging of results from different servers again neural network technology is used. More information about the system is available in (Heuser & Rosentstiel, 2000).

2.15.6 SOL-EU-NET

The goal of this project is to enhance competitiveness and find new business opportunities in the global IT market by establishing a virtual European enterprise composed of companies and research laboratories with highly specialised expertise in two IT areas: data mining and decision support. The established Sol-Eu-Net enterprise will be organised as a flexible business structure made of cross-organisational, time-focused, task-driven work teams. It will work towards enhanced usage of data mining and decision support in industry, businesses and public services, contributing to improved quality, efficiency and effectiveness of their operations. This will be achieved through specific solutions to end-user problems, prototype project workshops, project monitoring and consulting, collaborative work and combination of problem solutions, as well as through education, training and spreading information Web-based information source.

2.15.7 THE INTERSPACE PROTOTYPE

Based on SOM(System Object Model) technology and on the research carried out within the Digital Library Initiative (DLI) project at the University of Illinois, partially funded by ARPA ITO, the Interspace project aims to bring a new level in network information management based on correlation of knowledge and "intended to be universal to all repositories of all types". Its semantics are based upon statistical clustering techniques from information retrieval and image processing. Concept spaces, collections of abstract concepts generated from concrete objects are seen as independent of the physical objects they represent. The statistical analysis is not based on single terms but on up to 5-word phrases. It relies on and enhances the results of the TIPSTER project (see http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/). Different statistical algorithms had been evaluated and performance-wise evolved: concept spaces (co-occurrence matrices), category spaces (Kohonen maps), metadata generation (Hopfield nets) and meta-map generation.

A lot of software had been designed and implemented in the project using Java, Smalltalk and CORBA technology. The Interspace system is a layered system with a kernel layer providing the Interspace Analysis Environment and an underlying service layer with different modules organized by a domain management and fed by the Multimedia Concept Extraction unit. The modules are:

- Concept Assigner that performs automatic subject indexing based on a variant of a Hopfield network.
- Concept Space Generator that provides automatic generation of thesauri.
- Category Map Generator that classifies an information corpus into different conceptual categories using a variant of Kohonen's (Kahonen, 1997) self-organizing feature map (SOM).
- Validation Service that evaluates the results of the other modules.

3 MARKET WATCH

3.1 COMMERCIAL SYSTEMS

3.1.1 LEXIS-NEXIS

Lexis-Nexis is a leading decision-support information-and-services provider to professionals in the legal, business, government and academic markets. Provides authoritative legal, news, public records and business information, including tax and regulatory publications, in online, print or CD-ROM formats.

Its 12.000 employees worldwide serve customers in more than 60 countries. The division that comprises the Group's publishing assets is regionally organized, in Europe, Asia-pacific and Latin America. These include the market-leading Butterworths companies in the UK and British Commonwealth, Les Editions du Juris Classeur in France, and many other companies that are household names in their markets.

LexisNexis Group offers targeted Web information solutions that can be integrated into customer business processes and systems. Lexis-Nexis Web and dial-up online solutions combine searchable access to more than three billion documents from thousands of sources. More than 2.5 million documents are added each week.

Lexis-Nexis makes this data available to customers through its full-text search system, witch stores data in a proprietary format designed to optimise the full-text search service.

Its services among others include: “Search Advisor” that locates the appropriate legal materials or sources; “Case Summaries” that provides caselaw summaries written by its expert law attorneys; “Core concepts” or headnotes that link directly to review relevant cases, the “Get & Print” service that retrieves and delivers full-text documents to the printer or browser.

Lexis-Nexis offers a variety of pricing options, including subscription and pay-as-you-go.

3.1.2 OCLC

The OCLC Online Computer Library Centre Inc. is a nonprofits membership organization serving 41.000 libraries in 82 countries and territories around the world. It has been a key player since the very beginning of digital content. Starting in 1967 as a small cooperative known as the Ohio College Library Consortium (hence the "OCLC" acronym), it pioneered online library catalogues and shared library cataloguing. Many years and many millions of bibliographic records later, this work has grown into the giant WorldCat union library catalogue, which is used by thousands of libraries for cataloguing, research and interlibrary loan.

More recently, OCLC has turned its attention to the new opportunities of digital content. It developed the FirstSearch service for searching online research databases, and the Electronic

Collections Online service for permanently storing and searching online journals. It is continuing its leadership in library cataloguing by developing technique and standards for cataloguing electronic documents.

The services offered by OCLC are:

- Collection Management, Cataloguing & Metadata
- Reference
- Resource Sharing
- Digitisation & Preservation
- Databases

OCLC WorldCat (the OCLC Online Union Catalogue) is the world's largest database of bibliographic information created with the electronic catalogues shared by the librarians along nearly three decades. WorldCat offers over 47 million bibliographic records (representing 4000 languages) and holdings information vital for collection development, cataloguing, authority control, and retrospective conversion services. Through the OCLC FirstSearch service, users can access 70 databases including familiar names from leading information providers as well as resources provided exclusively by OCLC. OCLC databases include: WorldCat, ArticleFirst, Electronic Collections Online, NetFirst, PAIS International, PapersFirst, ProceedingsFirst, and the OCLC Union Lists of Periodicals.

Becoming an OCLC member enables global access to all these services and databases.

3.1.3 SCIENCEDIRECT

ScienceDirect is a Web database for scientific research that offers access to:

- More than 1600 scientific, technical and medical peer-reviewed journals from Elsevier Science as well as from societies and STM publishers, including top titles such as The Lancet, Cell, Tetrahedron, Journal of Molecular Biology, Brain and Language, etc.
- Over 40 million abstracts (free to all users).
- Over two million full-text scientific journal articles.
- An expanding suite of Bibliographic databases. A selected range of the most popular academic databases per scientific field, representing journal coverage of around 10.000 titles.
- Another one million full-text articles richly inter-linked to another publishers' platforms.

ScienceDirect covers a wide variety of subject areas and disciplines, including: Biological Science, Business & Management Science, Chemistry, Clinical Medicine, Earth Science, Economics, Engineering & Technology, Environment Science, Materials Science, Mathematics & Computer Science, Neurosciences, Pharmacology & Toxicology, Physics and Social Sciences.

Search functionality is based on title, authors' names or keywords within the article abstract and/or the full text of the article.

Selected articles can be viewed by a number of retrieval options:

- Summary: An outline including an abstract, hyperlink to references, thumbnail images and tables.
- Full-text article in HTML, with hyperlinks to the full-text and/or abstracts of cited articles and references.
- Full-text article in PDF format for fast downloading or printing.
- Export Citations and Abstracts, allows access to reference lists, exportable in file formats for use with citation management programs.

Users can also create a personal profile that provides access to additional features including email alert when new data is added, save their searches and personal journal list.

ScienceDirect offers a variety license options, from a complete access through a selection of a subset of the entire ScienceDirect journals database. Access to abstracts of articles and journal tables' contents is free to all users.

3.1.4 ICONDA

the International CONstruction Database sponsored by the Council for Building Research, Studies and Documents (CIB), and the International Union of Building Centre (UCIB), contains approx. 500.000 references to worldwide technical literature on civil engineering, urban and regional planning, architecture and construction.

Essential subjects are: architectural design and planning, construction of industrial, commercial, and residential structures, structural design, material properties and testing, public works and major construction projects, regional and municipal planning, project planning, finance and management, soil mechanics and geotechniques, maintenance, restoration, and conservation techniques, construction equipment and methods, and Computer Aided Design (CAD).

Sources are scanned for ICONDA by international organizations in 23 countries. Sources include over 600 periodicals, books, research reports, conference proceedings, business reports, theses, and non-conventional literature, from 1976 to the present.

Searching for information is possible by title, author name, journal name, publication date, publication year, ISBN, ISSN, starting date, estimated completion date or frequency of search terms.

3.1.5 COMPENDEX

is an interdisciplinary engineering database with over six million summaries of journal articles, technical reports, and conference papers and proceeding in electronic form, dating from 1970. Abstracts from over 5.000 international journals, conferences papers and technical reports are included. Each year, over 220.000 new abstracts are added from 175 disciplines and major specialties. Its Thesaurus contains over 9.000 terms and synonyms.

Topics covered range from Aerospace to Ocean & Underwater Technology.

Compedex database is available on ScienceDirect, others system databases and soon on LexisNexis.

3.1.6 ISI - WEB OF KNOWLEDGE

ISI Web of Knowledge encompasses both multidisciplinary and specialized content, covering journals, books, proceedings, patents, chemical structures, evaluated Web content, grant funding, and preprints. This research environment also provides research tools to help search, analyse, and manage information, as well as access a community of influential researchers.

Multidisciplinary content covered:

- Nearly 8.500 authoritative, high impact journals, including access to current and retrospective bibliographic information, author abstracts, and cited references found in:
  - Approximately 5.900 world's leading scholarly and technical journals covering more than 150 disciplines
  - Over 1.700 world's leading scholarly social science journals covering more than 50 disciplines.
  - Nearly 1.130 of the world's leading arts & humanities journals.

- Tables of contents from about 7.600 journals and 2.000 books, including links to more than 3600 ISI-evaluated Web sites and over 170.000 searchable, full-text Web documents, available from selected Web sites.
- Coverage of 20 million patents in all technologies from 40 international patent issuing authorities.
- Global coverage of proceeding from the most prestigious conferences and meetings in the sciences, social sciences, and humanities.

Specialized content covered:

- High quality, synthetic chemical information with precise query capabilities.
- Renowned life science and biomedical research index from the publisher of Biological Abstracts.
- Coverage of international applied life science research produced by CABI Publishing.
- A comprehensive index to literature in physics, electrical/electronic engineering, computing, control engineering, and information technology produced by the Institution of Electrical Engineers (IEE).

The tools offered by ISI Web of Knowledge enable to search bibliographic references on the Internet, organize references in a database, and quickly create and format bibliographies.

Optional, weekly e-mail alerts automatically keep researchers up to date with the latest search results.

There is a variety of subscription options from a selection of content by field of research through a limited time of subscription (e.g. $99 per year).

3.2 FREE AND OPEN SYSTEMS

3.2.1 SPARC

Scholarly **P**ublishing and **A**cademic **R**esources **C**entre. SPARC is an alliance of universities, research libraries, and organizations built as a constructive response to market dysfunctions in the scholarly communication system. These dysfunctions have reduced dissemination of scholarship and crippled libraries. SPARC serves as a catalyst for action, helping to create systems that expand information dissemination and use in a networked digital environment while responding to the needs of scholars and academe. Membership in SPARC numbers approximately 200 institutions in North America, Europe, Asia, and Australia. SPARC members pledge through a purchase commitment to support the SPARC-endorsed journals that fit their collection development agenda.

The high and fast-rising cost of journals has had a devastating effect on the flow of scientific communication, the research community, and library collections. The situation is especially dire for journals in the scientific, technical and medical (STM) fields. SPARC was created to offer a constructive response to this issue. It works to find common ground among libraries, publishers and scientists who share the goal of making scientific communication responsive to the goals of science.

Data gathered by the Association of Research Libraries shows that libraries are spending more and getting less. This study showed that serials spending was 152 percent higher in 1998 than a dozen years earlier -- yet there has been a seven percent decline in the number of titles libraries are getting for their money. Journals in the sciences rated the highest average journal cost.

The strain of rising journal prices is compounded by the availability of new media -- such as Web editions of existing journals -- and ever-more-specialized journals competing for available budgets.

Researchers are in search of society (or otherwise non-commercial) journals to which to submit their work -- journals motivated by service to the research community rather than by profit. SPARC works to facilitate the development of such journals and in the process stimulate competition in the realm of scientific communication. SPARC members support SPARC through annual membership dues.

SPARC is not a publisher. SPARC helps stimulate competition in the market by nurturing high-quality, low-cost journals published by researchers, societies or publishers with scientist -- and library-friendly values and practices

SPARC aims to introduce top-quality STM journals at a significantly lower price than those currently available. Ultimately, SPARC-endorsed journals give libraries the opportunity to more

than offset their costs with reductions in the number of high-priced journals to which they subscribe.

SPARC library subscription support gives journals a strong readership from the first few issues forward. This subscription base is the foundation upon which a new journal can build prestige, attract authors, and become a true alternative.

3.2.2 CUMINCAD

http://cumincad.scix.net

CUMINCAD, an acronym for "Cumulative Index on CAD", is a bibliographic index compiling papers related to computer- aided architectural design and helps focusing on future CAAD education and research activities. CUMINCAD supports the search and the dissemination of CAAD-related publications.

Each entry includes (1) typical bibliographic data, (2) extended bibliographic data, (3) Internet links and (4) administrative data. The bibliographic section includes information on authors, title, where and when it was published. This information is not structured into very small chunks as in most bibliographic libraries but broken into the four fields only. This facilitates the input that can be done quickly by simply using *cut* and *paste* of fairly large chunks of data. Rather smart full text searching techniques compensate for the lack of structure. Extended bibliographic data includes summary of the publication and the keywords (if available). The keyword set is not controlled, which is appropriate for a fast moving field such as CAAD. Internet related fields include possible links to full text of the publication and to author's co-ordinates. In case these are not defined explicitly, search into the Web can be launched and there are quite good chances of finding the full text or the authors.

In the framework of annual conferences organized by regional CAAD-Associations (ACADIA in North America, eCAADe in Europe, Sigradi in South America and CAADRIA in Australasia) thousands of papers have been published. Rarely were the proceedings published by a professional publisher, therefore, the texts were neither entered into commercial indexes, nor were they sold commercially. The full texts were not broadly available; only conference attendees had copies. On the other hand, the associations retained in most cases the copyright to this work and could therefore allow its publication/archiving in the CUMINCAD. Thus this work is available on the net and rescued from oblivion. At the time of writing, CUMINCAD includes 4301 records with abstracts. 883 papers are available in full text as well.

3.2.3 LOS ALAMOS

http://arxiv.org/new/

The trends in electronic scientific publishing indicate that more disciplines are setting up pre-print archives following the example of the Los Alamos physics pre-print server.

The Los Alamos pre-print achive in physics is one of the oldest (starting in 1991 and online from 1996) and established still working archive. The success of the archive can be traced back to the high energy physics community that did have a pre-existing hardcopy preprint habit, and had journals as their primary communication medium. However the Los Alamos pre-print archive initiative has grown also into other areas of physics and mathematics and even to computation and linguistics Ginsparg, P. (1996) Winners and losers in the global research village http://xxx.lanl.gov/blurb/pg96unesco.html

The submission rate statistics for the server shows that around 3000 pre-prints are being submitted monthly for the year 2002. The total number of documents on the server amount to (September 2002) 209.000. http://arxiv.org/show_monthly_submissions. The usage statistics is also very high, roughly 120.000 – 140.000 connections each day.

3.2.4 ICAAP

http://www.icaap.org/

The International Consortium for the Advancement of Academic Publication is a research and development organization devoted to the advancement of electronic scholarly communication. The mission of ICAAP includes technological support, publication, and enhancement of scholarly journals and educational resources, with the goals of greater accessibility, recognition and communication within the academic community.

ICAAP (International Consortium for the Advancement of Academic Publication) is a non-profit venture hosted by Athabasca University, Canada. Since 1998, ICAAP has produced a number of scholarly journals as well as maintaining a list of free scholarly publications. Access to the ICAAP collection is free and unrestricted.

In addition to providing services to scholars wishing to start their own independent scholarly journals and maintaining a database of free scholarly journals and resources, ICAAP also provides a service of tracking manuscripts through the peer review and publication paper path.

This service allows the scholar to:

- Register and make himself available for news and announcements
- Become a scholarly journal peer reviewer by simply registering as a user and enter areas of interest and/or specialization. Journal editors using the system are able to search the user database for appropriate reviewers.
- Become a scholarly journal editor by registering as a user and creating a new journal. The service provides the scholar with a default web page and full access to powerful back office editorial software to help manage the new journal.

3.2.5 ATMOSPHERIC CHEMISTRY AND PHYSICS

http://www.atmos-chem-phys.org

Offers a new journal concept to inprove scientific discussion by interactivity, providing an interactive scientific journal, Atmospheric Chemistry and Physics. The journal is launched recently in September 2001. The publication process concists of two stages and involves a scientific Internet-based discussion forum. The aim is to foster discussion, enhance effectiveness and transparency of scientific quality control through a rapid publication process and free accessibility.

The first stage envolves a rapid peer review stage with a quick publishing on the discussion website of the journal. During this interactive public discussion phase the referees comments as well as comments from the scientific community and the authors' replies are published.

The second stage completes the publication process and the final revised papers are published in the journal. The publication statistics http://www.copernicus.org/EGS/acp/ACPStats2.htm report an average total time from submission to final paper publication of 25 weeks.

3.2.6 CITESEER

http://citeseer.nj.nec.com/cs/

CiteSeer, also known as ResearchIndex, is a scientific literature digital library that aims to improve the dissemination and feedback of scientific literature, and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness. CiteSeer is a free citation index for computer science and other disciplines utilizing technology available at. Unlike most search engines, this database indexes postscript and PDF documents in addition to HTML files. The database parses the citations, identifies citations to the same paper, determines the context of citations within the documents, and indexes the full text. This combination of actions allows the user to search for articles by keyword, title, or author, find scholarly documents that cite a particular article, and look at the context of citations made within and to a particular article. (http://www.library.ucsb.edu/istl/01-winter/databases.html modern Electronic Publishing.

3.3 BUSINESS MODELS FOR DIGITAL CONTENT

3.3.1 INTRODUCTION

In the most basic sense, a business model is a method of doing business by which a company can sustain itself, that is, generate revenue.

From a more comprehensive point of view a business model is a description of how a company intends to create value in the marketplace. It includes the combination of products, services, image, and distribution that this company carries forward. It also includes the underlying organization of people, and the operational infrastructure that they use to accomplish their work.

Some models are quite simple. A company produces goods or service and sells it to customers. If all goes well, the revenues from sales exceed the cost of operation and the company realizes a

profit. Other models can be more intricately woven. Broadcasting is a good example. Radio, and later television, programming has been broadcast over the airwaves free to anyone with a receiver for much of the past century. The broadcaster is part of a complex network of distributors, content creators, advertisers (and their agencies), and listeners or viewers. Who makes money and how much is not always clear at the outset. The bottom line depends on many competing factors.

Reading the literature one finds business model categorized in different ways. One excellent example is this taxonomy offered by Michel Rappa (Alan T. Dickson Distinguished University Professor of Technology Management at North Carolina State University in Raleigh), which include:

- Brokerage
- Advertising
- Infomediary
- Merchant
- Manufacturer
- Affiliate
- Community
- Subscription
- Utility

A firm or a corporation may combine several different models as part of its overall Internet business strategy. For example, it is not uncommon for content driven businesses to blend advertising with a subscription model.

In the following paragraphs each business model is described, laying emphasis on those that are involved or can involve digital content.

3.3.1.1 Brokerage

Brokers are market-makers: they bring buyers and sellers together and facilitate transactions. Brokers play a frequent role in business-to-business (B2B), business-to-consumer (B2C), or consumer-to-consumer (C2C) markets. Usually a broker charges a fee or commission for each transaction it enables. The formula for fees can vary. A business model for digital content within this category could be:

- Business Trading Community or vertical web community -- is a comprehensible source of information and interaction for a particular vertical market. A community may contain product information, daily industry news and articles, job listings and classifieds. Example: VertMarkets.
- Transaction Broker. This model provides a third-party payment mechanism for buyers and seller to settle a transaction. An example is PayPal(section 3.2.2.13). Really, this model is used in combination with any other.
- Bounty Broker – offers a reward for finding a person, thing, idea, or other desired, but hard to find item. The broker may list items for a flat fee and a percent of the reward for items that are found.

- Search Agent – is an agent (i.e. a software agent) used to search-out the price and availability for a good or service specified by the buyer, or to locate hard to find information.

3.3.1.2 Advertising Model

The web-advertising model is an extension of the traditional media broadcast model. The broadcaster, in this case, a web site, provides content (usually, but not necessarily, for free) and services (like e-mail, chat, forums) mixed with advertising messages in the form of banner ads. The banner ads may be the major or sole source of revenue for the broadcaster. The broadcaster may be a content creator or a distributor of content created elsewhere. The advertising model only works when the volume of viewer traffic is large or highly specialized.

- Portal -- is a point of entry to the web, usually a search engine that includes diversified content or services. The high volume makes an advertising profitable and permits further diversification of site services. An example is Yahoo!.
- Personalized Portal – allows customization of the interface and content. This increases loyalty as a result of the user's own time invested in personalizing the site. See My Yahoo!
- Niche Portal – cultivates a well-defined user demographic. For example, a site that attracts home buyers, young women, or lawyers, can be highly sought after a venue for certain advertisers who are willing to pay a premium to reach that particular audience.
- Registered Users – content-based sites that are free to access but require users to registers (other information may or not be collected). Registration allows inter-session tracking of users' site usage patterns and thereby generates campaigns.
- Query-based Paid Placement – self favorable link positioning (i.e. sponsored links) or advertising keyed to particular search terms in a user query. Ex: Google.

3.3.1.3 Infomediary Model

Data about consumers and their consumption habits are valuable, especially when that information is carefully analysed and used to target marketing campaigns. Independently collected data about producers and their products are useful to consumers when considering a purchase. Some firms function as infomediaries (information intermediaries) assisting buyers and/or sellers understand a given market.

- Advertising Networks – services that feeds banner ads to a network of sites, thereby, enabling advertisers to deploy large marketing campaigns. By using cookies, de Ad Network operator collects data on web users that can be used to analyze marketing effectiveness. Example: DoubleClick.
- Audience Measurement Services – online audience market research agencies. Example: Nielsen//NetRatings.
- Metamediary – is a neutral, central on-line hub that automates transactions, aggregates information, improves market reach, and provides related services. Metamediaries help their participants reduce both product and process costs by resolving information-based inefficiencies, acting as a catalyst to compress time, slash costs, and improve processes. Example: Edmunds.

3.3.1.4 Merchant Model

Wholesalers and retailers of goods and services. Sales may be made based on list prices or through auction. In some cases, the goods and services may be unique to the web and not have a traditional "brick-and-mortar" storefront.

- Virtual Merchant -- or "e-tailer", a merchant that operates over the web. Eg : Amazon.com.
- Click and Mortar -- traditional brick-and-mortar retail establishment with web storefront. Eg: Barnes & Noble.
- Bit Vendor -- a merchant that deals strictly in digital products and services and, in its purest form, conducts both sales and distribution over the web. eg: Eyewire.

3.3.1.5 Manufacturer Model

A model predicated on the power of the web to allow a manufacturer (i.e., a company that creates a product or service) to reach buyers directly and thereby compress the distribution channel. The manufacturer model can be based on efficiency, improved customer service, or a better understanding of customer preferences.

- Brand Integrated Content -- Traditionally, manufacturers rely on advertising to build customer awareness. Commercials via broadcasters like radio, television and mass market publishers (newspapers and magazines), or through product placement in TV and motion pictures, has been a mainstay of modern business. The Web enables a manufacturer to integrate their brand more intimately with the content. The company's bmwfilms is a creative blend of advertising with entertainment that paves the way for a new approach that might be called "advertainment" -- taking the idea of product placement advertising to the extreme.

3.3.1.6 Affiliate Model

In contrast to the generalized portal, which seeks to drive a high volume of traffic to one site, the affiliate model provides purchase opportunities wherever people may be surfing. It does this by offering financial incentives (in the form of a percentage of revenue) to affiliated partner sites. The affiliates provide purchase-point click-through to the merchant. It is a pay-for-performance model -- if an affiliate does not generate sales, it represents no cost to the merchant. The affiliate model is inherently well-suited to the web, which explains its popularity. Variations include, banner exchange, pay-per-click, and revenue sharing programs.
Examples: Barnes & Noble and Amazon.com.

3.3.1.7 Community Model

The viability of the community model is based on user loyalty. Users have a high investment in both time and emotion in the site. In some cases, users are regular contributors of content and/or money. Having users who visit continually offers advertising, infomediary or specialized portal opportunities. The community model may also run on a subscription fee for premium services.

- Voluntary Contributor Model -- similar to the traditional public broadcasting model -- the listener or viewer contributor method used in not-for-profit radio and television broadcasting. The model is predicated on the creation of a community of users who support the site through voluntary donations. Not-for-profit organizations may also seek funding from charitable foundations and corporate sponsors that support the organization's mission. The web holds great potential as a contributor based model because the user base is more readily apparent. Example: WCP.org
- Knowledge Networks -- or expert sites, that provide a source of information based on professional expertise or the experience of other users. Sites are typically run like a forum where persons seeking information can pose questions and receive answers from (presumably) someone knowledgeable about the subject. The experts may be employed staff, a regular cadre of volunteers, or in some cases, simply anyone on the web who wishes to respond. Example: AllExperts.
- Discussion Miners - consists of generate revenue by aggregating relevant content from various community sources These niche players crawl, spider, or index through communities, gathering content from a variety of sources, like Usenet newsgroups, message boards, chat archives, customer-provided ratings and reviews. The posting from various sources are aggregated and analyzed using software that “reads” all of the postings and finds the themes or clusters. The general themes and clusters of information are packaged into reports and sold to companies as an adjunct to traditional market research, i.e. focus groups

3.3.1.8 Subscription Model

Users are charged a daily, monthly or annual fee to subscribe to a service. It is not uncommon for sites to combine free content with "premium" (i.e., subscriber- or member-only) content. Subscription fees are incurred irrespective of actual usage rates. Subscription and advertising models are frequently combined.

- Content Services -- beyond newspapers and magazines, the Web has encouraged the use of the subscriber model for music and video, as well. Example: Listen.com
- Person-to-Person Networking Services -- are conduits for the distribution of user-submitted information, such as individuals searching for former school mates. Example: Classmates.
- Trust Services -- an independent third party that engenders trust between unfamiliar parties entering into a transaction. The need of trust increases with the value and complexity of the product or service that is sold. Trust services typically come in the form of membership associations that abide by an explicit code of conduct, and in which members pay a subscription fee. Example: Truste.
- Internet Services Providers -- offer Internet connectivity and related services on a monthly subscription. Example: American Online.

3.3.1.9 Utility Model

The utility model is based on metering usage, or a "pay as you go" approach. Unlike subscriber services, metered services are based on actual usage rates. Traditionally, metering has been used

for essential services (e.g., electricity water, long-distance telephone services). Internet service providers (ISPs) in some parts of the world operate as utilities, charging customers for connection minutes, as opposed to the subscriber model.

3.3.2 SYSTEMS PROVIDED FOR PAYMENT

3.3.2.1 Overview

A close relationship between e-commerce and e-payment systems has been assumed as a matter of course. Conventional wisdom was that “there could be not thriving e-commerce without robust, secure and standardised e-payment infrastructure.” Many specialists saw e-payment as a killer application. Yet, the actual development proved very different. While electronic commerce, in both business-to-consumer (B2C) and business-to-business (B2B) segments, has been growing more rapidly than the initial market forecasts, the development of Internet-based payment systems has been quite disappointing and practically all the systems have run into difficulties, sometimes fatally.

There are a good number of systems provided for payment. Some solutions are:

3.3.2.2 Digital cash

- Flooz.com a central account based payment system allowing user-to-user payments
- Internet Cash Corp is a prepaid card that is purchased from a real-world store and spent on-line. A temporary anonymous account is setup from the unique card ID (which looks something like: 3842 F932 J283 7832 PRXZ), and its value is decremented as purchases (as small as 50 cents) are made on-line

3.3.2.3 Virtual Points

- Mypoints.com is the ultimate destination for free rewards--on the Internet and beyond. With MyPoints, you can earn rewards from name-brand merchants like Blockbuster Video, Barnes & Noble Booksellers and Bloomingdale's. You can even earn vacations and frequent flyer miles.

3.3.2.4 Person-to-person payments:

- PayPal allows paying with a credit card or checking account through a recipient's email address.
- BillPoint allows person-to-person payments from a credit card. Originally targeted at eBay customers.
- Yahoo! PayDirect allows person-to-person payment via e-mail.

3.3.2.5 Virtual escrow is a third party which holds a buyers money in trust until a vendor delivers purchased goods

- I-Escrow Inc

- Escrow.com

3.3.2.6 Digital wallets:

- Yahoo! Wallet is a service offered by Yahoo that securely stores credit card, billing, and shipping information. Only is necessary sign in with the Yahoo! ID and a Security Key
- Passport is the same service but in this case offered by Microsoft.
- Earthport offers a digital wallet, allowing seamless transactions across a digital environment - over the Internet via PC's, digital TV and mobile commerce formats including WAP and SMS. Facilitates two-way transactions in multiple currencies from micro payments to large fund transfers using cash or payment cards.

3.3.2.7 Credit cards:

- DataCash provide secure credit card authorization over SSL.
- Merchant Commerce Inc is a credit card and electronic check processing for Internet businesses.
- CyberCash recently acquired by VeriSign, enables companies to authorize, process, and manage multiple payment types (including credit cards, debit cards, purchase or procurement cards, Internet checks and automated clearing house transactions), multi-currency options, and different payment models online.

3.3.2.8 Virtual credit cards.

All information necessary to complete a transaction online is provided via the account reference card. No physical card is provided. The specific account number is issued to the customer by VISA, American Express or MasterCard.

3.3.2.9 Credit and debit cards

Despite numerous attempts aimed at offering innovative alternatives, credit and debit cards are currently the main payment instrument for B2C transactions, used in over 95% of purchases. The most frequently used approach relies the Internet only at the initial stage, when a customer communicates his card number to the merchant's server. Once the number is communicated (using SSL), the remainder of the payment settlement uses the existing payment network and procedures. This requires having merchant accounts with the authorization networks. However, these merchant accounts are unreasonable for small and home business with small profit margins. (A monthly fee, plus approximately 2% of purchase amount).

3.3.2.10 Digital or Electronic cash

Digital or Electronic cash is similar to paper currency and coins. The idea is that customers posses electronic tokens or funds that can be exchanged immediately for goods and services. The possibility of anonymous use of this tokens introduce two important goals to achieve:

- Systems are needed to secure that the tokens are authentic.
- They are spent only once.

- Generally, the initial balance of funds for users of these systems is paid into an account that is unique to the system. In other words, the electronic funds that make up the initial balance are purchased by credit card, wire transfer, or some other means.

3.3.2.11 Electronic Checks

An electronic check contains all of the information that is found on a traditional paper check but can be transferred by email. This kind of e-payment is not widely used in Europe, but it is very important in the U.S. in particular in B2B environments.

3.3.2.12 Smart Cards

Smart card technology makes use of a chip that stores information, from digital signatures to financial information. For financial transactions, the technology can be put to use in much the same way as rechargeable phone cards. The chip can be "charged up" with a specific amount of money that can be used as cash for purchases. When the funds are depleted, the chip can be recharged with more funds. While many implementations rely on putting smart card chips on plastic cards (like ATM or credit cards), the chips can also be put into electronic devices such as cell phones and computers.

3.3.2.13 PayPal

This is an independent service that allows electronic payment person-to-person. Paypal acts as a neutral intermediary offering low risk to both seller/receiver and buyer/sender of the money
PayPal, uses e-mail to inform the receiver that a payment has been made. It accepts money from the purchaser in one of three ways:

- Charging the purchaser's credit card for any transactions (payments)
- Debiting a checking account for any payments
- The purchaser sending a check to create a positive balance in his account at PayPal, and having any payments deducted from the account Payment recipients can use the money in the account for online purchases or payments, can receive the payment from PayPal by check, or can have PayPal directly deposit the money into their checking account.

3.3.2.14 Advertising supported

Advertising supported business model consist of:

1. Build a website with interesting content
2. Attract visitors who want to read that content
3. Sell space on that website to advertisers

This, that seems to be a simple idea, it has proved extremely difficult in practice, and only a few have been able to make a profit through selling advertising alone. Those that have been successful tend to share the following attributes:

- Excellent content attracting a large, specialist, group of visitors of particular interest to advertisers.
- Innovative and varied advertising options.

- Competitive pricing and service.

The Paid directory model is one of the most used and successful ways chosen as an advertising-supported program. The model works just the way it sounds: advertisers pay to have a link. The link usually leads either to a single web page about the advertisers or directly to the advertisers themselves.

The model gives the advertisers exposure across the site's multiple ad vehicles, including everything in the price. Give them plugs in email mailings. Give them banner ads on the site. Give them the ability to insert their messages into any communication with the site audience.

The offer can be like this:

- A listing in the directory of sponsors.
- A mini web page on the site (with a link or a request-for-more-information form).
- Rotation of their banner throughout the site.
- A plug in the site's email newsletter.

There are different methods to charge costs to the advertisers that are well known definitions for the Advertising Industry:

CPM	Cost per thousand impressions (banner ad views)
CPC	Cost per Clickthrough (every time someone clicks on the banner)
CPA	Cost per Action (sign ups, filled out forms, purchases, clicks)

3.3.2.15 Content Licensing

Content licensing is an agreement between two parties, a content creator and an organization or company, to use the content without transferring the ownership. The content can be licensed for a variety of purposes including web feeds, reprints, or email distribution.

The licence encompass a series of elements which describe clearly the agreement terms:

- Parties: a clear statement using the full contractual names of the parties to the licence.
- Definitions: essential terms must be defined (eg Authorised Users, Licensed Material, Access Options etc).
- Agreement: a broad statement of the type of licence, time frame, and licence purpose(s).
- Coverage: detailed description of the materials being licensed including current and retrospective files; content (abstracts and full text) etc.
- Variations: content providers often insert clauses allowing them unilaterally to vary aspects of the licence, typically with regard to the addition, withdrawal or suppression of content or the raising or lowering of subscription costs.
- Permitted uses and prohibitions: this is the important detail on what is essentially regarded as fair use of licensed materials including access, use, display, downloading, storage, re-use, incorporation of copyright statements on downloaded materials, special terms on networked use of downloaded materials etc and any further restrictions on their use.

- Undertakings, warranties, indemnities and liabilities: the licensee should ensure that certain undertakings are included in the licence - such as the licensor guaranteeing a suitable hardware and software platform to provide acceptable access times; 24 hour availability; compliance with security and privacy requirements for users accessing remote sites; the frequency of provision of usage data etc. It is also essential that warranties confirm that the licensor has a legal right to license use of the copyright material, and that no third party copyrights or other intellectual property rights are likely to be infringed. An indemnity to this effect should be included in the licence.
- Termination: procedures for and conditions under which the licence can be terminated by either party must be specified.
- Applicable law: a statement indicating which law governs interpretation of the licence, especially important if a dispute arises, is also needed.
- Cost and payment: the subscription price, statements relating to VAT and payment procedures.

3.3.2.16 Sponsoring

The sponsoring model consists of selling an advertiser the exclusive right to advertise on one or more particular piece of content.

This business model requires firstly that the content would be changed with some frequency to keep the audience interested. And second, someone has to work with the advertiser to make sure its brand is properly represented through the content area.

The offer would be something like this:

- Exclusive sponsorship of a content area.
- The sponsor's logo is incorporated in every banner, button, and graphic in the content area.
- Text ads included in any email communications with readers of the content area.

3.3.3 FREE SYSTEMS

Free systems are those systems that are free licensed, allow free redistribution and also can include shared development and free access to the source code. These can also allow a free information interchange between peers, like P2P technology.

3.3.3.1 Napster/Gnutella

Napster and Gnutella are the two most known examples of Peer-to-Peer (P2P) technology. This technology helps people to collaborate with others to exchange information and to efficiently use the processing power of their network. P2P networking is a type of network in which each workstation has equivalent capabilities and responsibilities. This differs from client/server architectures, in which some computers are dedicated to serving the others. At its most basic level, P2P is simply one computer node talking directly to another without any intermediary.

Napster was born as a solution to share the millions of MP3 music files stored by the people on their PCs. As a central directory for available files, the Napster server provides information

about where the song is located, the user sharing the song, and details helpful which file to download. Napster differs from pure P2P technology because these MP3 files are cataloged in a central directory hosted by a server. No files are actually stored on this server; it exists only to index the files of its users, to become a search engine for locating files.

Along the same lines, Gnutella is able to share files that other Gnutella users have set up to make available. Whereas Napster is limited to MP3 music, Gnutella allows downloading any type of file. Unlike Napster, which relies on centralized servers for searching, Gnutella lets clients interact directly with one another in a decentralized network, without the use of a central indexing system. There is no a central server, so all network topology information must be discovered dynamically.

The cost effectiveness of P2P can be found in a reduction of centralized management and server storage resources and in optimised computing resources. In terms of the financial services industry, many feel that P2P represents a new future for transmitting research and deal-making in a virtual environment.

3.4 WEB SYNDICATION

3.4.1 INTRODUCTION

The exponential growth of the Internet, heterogeneity of technologies, sites and content types has complicated maintaining a transparent cyberspace. In this scenario the Web Syndication Model has gained an increasing importance for automating self-organising information repositories providing up-to-date content from independent sources. The separation of roles, content creation and content has been a business model widely used in the print and broadcasting industry but is becoming increasingly relevant in doing business on the Internet, in development of new business models and B2B relationships.

3.4.2 THE MODEL

The basis of the web syndication model “eSyndication” is sharing of content according to a predefined business relationship, for example an eMarketPlace that receives catalogue updates from suppliers. The supplier syndicates his catalogue data, based upon an agreement with the eMarketplace, which then aggregates catalogue data from several sources and delivers it to the consumer as a single virtual catalogue. The supplier gains a wider web presence by syndicating to several eMarketplaces, while the customer gets a single point of entry to several suppliers simultaneously. In the above scenario the key feature is the automation of the data exchange process by dynamic syndication controlled by scheduling or business-rule triggers. “eSyndication is the process where a syndicator (content producer or distributor) delivers content to a subscriber (content aggregator or destination) according to an agreed-upon recurring schedule, based on time or business rules. The syndicator is said to syndicate content to the subscriber, and the subscriber is said to subscribe to content from the syndicator.”

The web syndication model defines roles of stakeholders the content creator, the syndicator, the content distributor and the consumer and additionally a number of possible intermediary partners.

- Content creators are the originators of the content and the endpoint for business transactions. Through syndication, originators can extend their market reach and decrease time-to-market. Additionally small transaction businesses and small commercial sites have easier access to volume markets for relatively little cost.
- Syndicators provide for transformation and packaging of content and management of content distribution and relationships between content creators and distributors and can provide intermediary services (aggregators), technology solutions or software applications that facilitate syndication in complex business environments
- Distributors act as aggregators of syndicated content and provides the interface to the client and handles all interaction on behalf of the content owner, including business transactions e.g. subscriptions and eCommerce.

The above separation of concerns provides for flexibility in dynamic business relationships and type of business models. It facilitates functional specialisation and a much more efficient use of resources and opportunities in doing business on the web.

3.4.3 METADATA STANDARDS

The following are some of the metadata standards used in Web Syndication (see section 5 – Metadata)

- Information and Content Exchange (ICE)
- Rich Site Summary or RDF Site Summary (RSS)
- Open Content Syndication (OCS) Directory format
- Channel Definition Format (CDF)

3.5 OPEN SOURCE MOVEMENT

In general, open source refers to any program whose source code is made available for use or modification as users or other developers see fit. Open source software is usually developed as a public collaboration and made freely available.

Open Source is a certification mark owned by the Open Source Initiative (OSI). Developers of software that is intended to be freely shared and possibly improved and redistributed by others can use the Open Source trademark if their distribution terms conform to the OSI's Open Source Definition. To summarize, the Definition model of distribution terms require that:

- The software being distributed must be redistributed to anyone else without any restriction.
- The source code must be made available (so that the receiving party will be able to improve or modify it).

- The license can require improved versions of the software to carry a different name or version from the original software.

While it is true that an open source business may not make money directly from its products, it is untrue that open source do not generate stable and scalable revenue streams.

The open source business model removes the commercial value away from the actual products towards generating revenues from ancillary services like systems integration, support, tutorials and documentation.

Open source also cuts down on essential research and development cost while at the same time speeding up delivery of new products.

4 STATE OF THE ART IN TECHNOLOGY

4.1 OVERVIEW

This chapter presents a general overview of a group of technologies and tools of general purpose that can be used to support the design of the SciX architecture including:

- eXtensible Markup Language (XML);
- Universal Discovery Description and Integration (UDDI);
- Web Services Description Language (WSDL);
- Simple Object Access Protocol (SOAP);
- Component Technology;
- Client/Server and Web Applications.

4.2 XML

Nowadays, it is widely accepted inside the World Wide Web community that XML, the e**X**tensible **M**arkup **L**anguage, is the standard to support the exchange of information among heterogeneous systems from different organisations. XML, developed by the W3C in 1996, is actually a *metalanguage* (a language for describing other languages) due to the fact that it allows the creation of customised markup language for unlimited different types of documents. XML allows designers to create their own customised document formats, thus enabling the definition, transmission, validation, and interpretation of documents among heterogeneous applications.

The aim of XML is to exchange structured documents in an open way between heterogeneous systems in a simpler way than SGML and in a more advanced way than HTML. The structure of an XML document can be given by a Document Type Definition (DTD) or by an XML-Schema, which are described next.

4.3 WEB PUBLISHING

4.3.1 OVERVIEW

It can be expected that within few years all major application will be web based or at a minimum have a web front-end. The rapid move to the web as a response to user requirements for multiple presentation formats for different presentation media are moving application developers from static HTML based presentation as front-ends to XML based publishing frameworks. Portability of data and access by different user-agents in the rapid changing business environment set demand for highly flexible application interfaces and presentation look and feel.

HTML and CSS have in the past, primarily been used by web developers for separating content from style in web applications, but the need for frequent and regular updating presents challenges in maintaining consistency across web-sites. More and more developers have

adopted architectures for server side generation of dynamic web pages. Server side generated web pages (Server Pages) combine static HTML code and dynamically generated data at run-time. Server Pages act as HTML templates, which include scripting code for inserting dynamically generated content at the right places within the web page. The scripting languages provide interfaces to component architectures such as Java Beans or COM for instantiating component instances during the processing of the server page. Scripting code instantiate objects for data generation or interacting with the file system, databases and other data sources. This type of architecture provides a powerful and flexible way of separating between content generated by scripting code and presentation by the HTML coded template page.

Technology providers have developed their own flavour of server side architectures, Microsoft the Active Sever Pages (ASP), the Java community Java Server Pages (JSP) and the open source community Hypertext Pre-processor (PHP) and the Common Gateway Interface (CGI). JSP, ASP and PHP all share the same approaches in handling server pages. However JSP unlike ASP and PHP, which are interpreted at run-time, is compiled to byte code called Servlets before invocation. Once the Servlet is available subsequent requests for web pages will invoke the ready Servlet in memory saving computing resources.

In the open source community several solutions have evolved around Java and XML that attempt to obtain clearer approach for presenting web content. Because Java can run in any platform it has been the driver for many new and emerging web technologies using XML as the common standard for data portability and integration of separate existing applications. Java and XML web publishing relay on Servlets as the technology for delivering dynamic content. The Servlet technology has proven advantage over other that share similar concepts like CGI in that they are instantiated as process threads rather then separate processes with substantial saving in resources and performance. Additionally data persistence and process state can be handled cleanly through implicit objects in the Servlet API.

WebMacro (www.webmacro.org) is a 100% Java open-source HTML template engine and back end servlet development framework that enables programmers and designers to work together while promoting the model-view-controller pattern. The aim of WebMacro is to allow web designers and code programmers to work together on projects but apart in parallel thereby achieving separation of content and presentation.

Tea (http://opensource.go.com/Tea/TeaArchitecture.pdf) is a simple yet powerful template language. Tea is most commonly used for creating dynamic web pages in the TeaServlet. Tea is a strongly typed, compiled programming language, designed to work within a Java-based hosting environment..

“The key features of Tea include:

- Simplicity. The language is easy enough for someone without a programming background to use and understand.
- Separates data acquisition from data presentation. All of the data used by a template written with Tea is accessible from information passed to the template at execution time.

- Protects critical data from executing templates. An executing template written with Tea is not able to directly access data from sources such as a database, so there is no need to worry that a template writer will be able to accidentally damage mission critical data.
- Protects the system in which a template runs from the executing templates. Malformed templates are not able to affect the system in which they run."

Active Server Pages (ASP), the JSP equivalent from Microsoft is now superseded by ASP.NET a part of the .NET framework for web application development on Windows platforms.

Hypertext Pre-processor (PHP) has gained a lot of support in the resent years. PHP is a project of the Apache Software Foundation and is a type-less general purpose HTML embedded scripting language for web development and as acclaimed by the authors, a language that is what scripting languages from the major vendors should have been. The latest release now includes a new Zend scripting engine, "which greatly improves the object model, adds exception handling and provides a much better infrastructure for the integration of external technologies like Java or .NET. (http://www.php.net)". PHP is available for most platforms and web servers that make it a proven alternative to JSP and ASP for server side web development.

XML APIs - SAX (Simple API for XML) and DOM (Document Object Model) are both means to access content in XML documents.

Technology providers such as Microsoft, Sun and IBM provide parsers that are accessible from most development environments and support both types of interfaces.

JDOM is a Java specific API for accessing XML Documents and as such provides a more intuitive and straightforward interface for Java programmers then DOM. JDOM is not a XML parser. The JDOM provides a set of classes for manipulating XML and must relay on an external parser to supply the raw XML input, DOM tree or SAX events.

JAXP from Sun Microsystems is a Java API or an abstraction layer for manipulating XML. JAXP provides a vendor-independent API (i.e. it handles vendor-specific parser issues behind the curtains) to DOM and SAX as well as providing a clean Java interface to DOM and SAX and methods for handling of namespaces and validation from Java perspective known to be cumbersome using DOM or SAX.

4.3.2 WEB PUBLISHING FRAMEWORKS

To achieve complete separation of content and presentation in web publishing a more sophisticated content publishing system is required then this can be achieved with the XML/XSL approach. These systems are normally referred to as web publishing frameworks.

Web publishing frameworks do address the need for consistently and uniformly styled web content. As normal URL requests are made to web servers a web-publishing framework is responsible for responding to similar requests with a published version, which is the response, transformed to the format required by the user-agent issuing the request. The publishing engine

will automatically determine the correct transformation either by inquiring about the user-agent type or by resolving requests for different document formats. This implies that a publishing frameworks support media independent publishing, serving of clients on any platform, generation of pure content from row internal data, apply uniform presentation and style in a consistent way over a set of documents and apply transformations to content to suite user-agent and consumer needs. The above requirements generally translate to that web publishing frameworks are now exclusively bases on XML and related technologies such as XSL.

Web publishing frameworks vary widely in implementation, features and conventions. Many depend on platform specific technologies that limit portability and effect stability such as use of a vendor specific XML parser, support for different XML specifications and implementation of latest version of W3C specifications. Other considerations include support for integration with existing technological solutions and existing data resources. As well as application specific concerns such as role-based publishing, content from various authors and presentation and style from web-artists.

There exist numerous systems on the market both commercial and open source that claim to be web-publishing frameworks. Some are mere web content management systems that offer some means for separation of concerns (i.e. content creation and styling) and some XML capability for generation of content and application of XSLT for transformation. JSP, ASP, PHP etc based systems all share the inherent problem with these architectures, intermixed logic and presentation and to some extent content generation. A new technology was needed to obtain the web publishing framework objectives. One of few such technologies that have been successfully implemented and that have gained wide support is the Apache Cocoon, a sub-project under the XML Apache development project.

Cocoon is a powerful and flexible XML/XSL web-publishing framework built according the Separation of Concern principle (SofC). Cocoon architecture introduces a pyramid model where a management node sits at the top of pyramid and logic, content, and style nodes at the bottom. Interaction between individual nodes is enforced by contracts.

The strength and scalability of Cocoon stems from a pluggable architecture into which Generators, Transformers and Serialises can be added at will without any modification to the Cocoon framework. Also, another feature of Cocoon is that it compiles Generators and Transformers into directly executable form, achieving considerable performance gain over event-driven run-time interpretation used in normal parsing. Generators are compiled from “logicsheets” and XSP code (see below) and Transformers from XSL stylesheets.

XSP builds on concepts from JSP namely the "taglib". Cocoon user manual defines logicsheets as; “An XSP logicsheet is a "tag library" in the sense that it defines a set of custom XML tags which can be used within an XSP program to insert whole blocks of code into the file”. Cocoon comes with several pre-defined taglibs. By using taglibs, business logic can be hidden from the developer, as he just needs to refer to defined tags to produce content.

Similarly as JSP technology generates Servlets, XSP technology produces Generators. The

content, logic and style. As XSP is XML code, unlike JSP or ASP, it can be transformed and manipulated just like any other XML code and is easily portable across applications.

Cocoon leverages other Apache XML technologies like Xerces, Xalan and FOP to provide a comprehensive next generation web publishing framework. Cocoon is based around XML and XSL and targeted to sites of medium - high complexity.

JPublish (www.jpublish.org) “is a simple web publishing system which uses the Velocity template engine (a subproject of Jacarta-Apache project, http://jakarta.apache.org/velocity/) in combination with a content management framework to build dynamic web sites. JPublish was designed to ensure a clean separation of content, programming logic, and presentation logic”.

JPublish separates the different parts of web applications so that the whole application is easier to develop and maintain. JPublish defines several roles that determine the design of JPublish.

- Designers will typically work with templates to create a uniform design for a site or sections of the site.
- Content producers work with documents (XML) in content repositories
- Domain Specific Programmers will typically develop domain-specific Java code, either as EJBs or as normal Java classes.
- Integration Programmers will integrate the domain-specific code with the web site by creating "glue" logic. This glue logic is represented by the JPublish's action system.

4.3.3 PUBLICATION FORMATS AND XML VOCABULARIES

This section will focus on XML vocabularies that have emerged for publication of media-independent structured XML word processing type of documents. Several initiatives are underway to enable the creation of complex word processing type of documents in appropriate XML vocabularies. These XML vocabularies specify elements to mark-up documents such as a book, a article, a paper or a technical documentation and provide semantics for common features such as headings, sections, paragraphs, table of content, indexes etc -. elements that normally make-up written publications.

The **DocBook** DTD specification (www.docbook.org and www.oasis-open.org) dates back to 1991 and originated in the document type definition (DTD) of SGML. DocBook content model (XML/SGML vocabulary) was developed to provide a common consensus for computer and software technical documentation and to make documents interchangeable between parties. In 1994 DocBook development was moved to the Organization for the Advancement of Structured Information Standards (OASIS). Although not a general authoring vocabulary it has however been used by number of orgainsation for variety of publications. DocBook is now an XML vocabulary, used for semantic markup of documents such as books, articles, and technical documentation. Currently the DTD version is the only officially released version of the vocabulary but a W3C XML Schema is under development and is currently available in a preliminary version. Several XSL styleheets are available for transformation of DocBook sources into other formats, including HTML, XHTML, Formatting Objects (FO) for producing various target formats such as PDF, RTF, PS, etc and JavaHelp and HTMLHelp.

OpenOffice (www.openoffice.org) specification. OpenOffice originated from the release of Sun Microsystems to the open source community of the source code and specifications of StartOffice, a complete Office desktop suite, consisting of a word processor-, a spreadsheet-, presentation- application and more to complete the package. Although not all of StarOffice components such as the spell checker, fonts, clipart and database access where released, OpenOffice is never the less, a fully functional office suite to be soon comparable and a viable alternative to Microsoft Office suite. Among its strong features is that it persists all files in native XML format. StarOffice and OpenOffice share the same XML format specification. In its design reuse of XML based standards, with elements and attributes borrowed as possible from HTML, XSL-FO, Xlink, Dublin Core or SVG and inclusion of MathML. The StarOffice XML document is a compound document where individual (streams) parts are packaged into a single document through the use of the ZIP format. The constituent parts of an StarOffice document are text content, layout and styles, meta data and embedded images, graphics and objects. The above format enables separation of content, style and meta-data that can be read and manipulated independently of each other.

The StarOffice XML specification is designed in a way to be easily process, read and understand but not necessarily easy to implement (e.g. like native binary documents). This however, allows the format to abstract application specific issues and be highly portable. The XML specification is open for extensions and supplemental information. Custom style attributes can be added at will, which allows arbitrary information to be added to the document, furthermore complete streams can be added to the package adding supplemental information to the document.

Being an XML document offers generic advantages like being fully accessible by growing number of XML tools (e.g. viewers, editors, transformers and XML native databases), easily portable between applications, easily transformed to application dependent formats, version interoperability and archiving for long-term storage e.g. no out of date versions. One can on the other hand see the ZIP packaging of the StarOffice document as a drawback at first (e.g. not being able to parse the document right off), but ZIP is a widely available as an API which is merely adding an extra processing step in the pattern for extracting the needed parts from the ZIP archive before final processing.

Another important feature of StarOffice/OpenOffice is adaptability and openness. User developed Import and Export components can easily be integrated in to the office suite enabling integration with wide variety of applications. It also exposes an open API for integration with custom solutions as well as an editor API, which allows the StarOffice editor component to be used directly in custom applications and controlled through the API.

OpenOffice has just released its 1.0.1 version. It hasn’t been robust enough to gain wide usability, but it would be fair to say that when stable releases will become available it will be a contender as an alternative to Microsoft Office as well as a candidate standard for open office documents of the future.

The **Open eBook Forum** (OeBF) is an international trade and standards organization. Its members consist of hardware and software companies, publishers, authors and users and include heavy players such as Microsoft, Adobe and Overdrive Inc.

The common goal of OeBF members is to establish specifications and standards for electronic publishing, foster the development of applications and products that will benefit content creators, makers of reading systems and consumers. The following scope of the Open eBook is obtained from the Open eBook Publication Structure specification at the OeBF web site (http://www.openebook.com).

- "The purpose of the Open eBook Publication Structure is to provide a specification for representing the content of electronic books.

W3C InfoSet Specification (taken from http://www.w3.org/TR/xml-infoset/) "This specification defines an abstract data set called the XML Information Set (Infoset). Its purpose is to provide a consistent set of definitions for use in other specifications that need to refer to the information in a well-formed XML document

An XML document's information set consists of a number of information items; the information set for any well-formed XML document will contain at least a document information item and several others. An information item is an abstract description of some part of an XML document: each information item has a set of associated named properties. In this specification, the property names are shown in square brackets, [thus]. The types of information item are listed in section 2.

4.4 UDDI & WSDL

4.4.1 UDDI

The Universal Description, Discovery and Integration (UDDI) is a specification defining a way in which businesses can publish and discover information about "Web services". A Web service describes specific business functionality, offered to other companies or programs for their use. UDDI project is working to enable companies to quickly, easily, and dynamically find and transact among themselves. UDDI enables a company to:

- Describe its business and its services;
- Discover other companies that offer desired services; and
- Interoperate with these other companies.

The approach adopted by UDDI is to have a distributed registry of businesses and their services, implemented in a common XML format. The *UDDI Business Registry* (UBR), which is the implementation of the specification developed by the **uddi.org**, is a core element of the infrastructure that supports Web services. It provides a place for a company to register its business and the services that it offers. People or businesses that need a service can use this registry to find a business that provides the service. The UBR is operated as a distributed service. An Operator's Council sets policy and quality of service guidelines for the operators. A

"node operator" is a company that runs an instance of the public UBR. The operators replicate the registrations across all nodes on a regular basis thus resulting in a complete set of registered records available to all. The operators support a common set of APIs that will ensure the integrity and availability of the information provided.The UBR contains information about businesses and the services they offer.

4.4.2 WEB SERVICES

A Web service is a self-describing, self-contained, modular unit of application logic that provides some business functionality to other applications through an Internet connection. Applications access web services via ubiquitous web protocols and data formats, such as HTTP and XML, with no need to worry about how each web service is implemented. Web services can be mixed and matched with other web services to execute a larger workflow or business transactions.

Access to and from the UBR is performed using the Simple Object Access Protocol (described later in this document). However, a service registered in the UBR can expose any type of service interface. A service interface is the programmatic interface that is used to invoke the service. A web service interface can be implemented using an Internet protocol, such as SOAP, ebXML Message Service, E-speak, XML-RPC, CORBA, Java RMI, and COM+.

A *tModel* and a Binding Template point to specifications that describe the web service interface. UDDI does not dictate any specific technology or methodology to describe a web service interface. A web service interface can be described in a number of different ways. It can be described using simple prose, or it can be described using more formal description languages. An interface message format can be described using an XML schema, or using a service interface description language, such as the Web Services Description Language (*WSDL*). The WSDL specification provides a simple XML-based vocabulary for describing XML-based Web Services that are available over the network. The services themselves communicate using the Simple Object Access protocol (SOAP), HTTP, SMTP, or by some other means; WSDL, however, gives the user the metadata required to set up the communications. WSDL itself says nothing about how to publish or publicise such service descriptions.

As communications protocols and message formats are standardised in the web community, it becomes increasingly possible and important to be able to describe the communications in some structured way. WSDL addresses this need by defining an XML grammar for describing network services as collections of communication endpoints capable of exchanging messages. WSDL service definitions provide documentation for distributed systems and serve as a recipe for automating the details involved in applications communication.

4.4.3 WSDL DOCUMENT STRUCTURE

WSDL provides a way for service providers to describe the basic format of web service requests over different protocols or encoding. WSDL is used to describe **what** a web service can do, **where** it resides, and **how** it is invoked. While the claim of SOAP/HTTP independence is made in various specifications, WSDL makes the most sense if it assumes SOAP/HTTP/MIME as the

remote object invocation mechanism. UDDI registries describe numerous aspects of web services, including the binding details of the service. WSDL fits into the subset of a UDDI service description.

In short, WSDL is a template supporting how services should be described and bounded by clients.

4.5 SOAP - SIMPLE OBJECT ACCESS PROTOCOL

The Simple Object Access Protocol (SOAP) is a lightweight protocol for enabling heterogeneous applications to exchange information in a distributed environment. SOAP can be considered as a firewall friendly protocol as well as a platform and programming language agnostic technology. The SOAP native domain is the Internet and, as such, SOAP relies on heavy-weight Internet standards: XML and HTTP/SMTP. SOAP uses XML to represent the data to be exchanged due the fact that it is an extensible framework that is easy-to-use and has a low-cost of entry. SOAP takes advantage of HTTP/SMTP because more than any other application protocols, they both connect the world. Using anything else than HTTP/SMTP to connect over the Internet allows one to find out that other application protocols are just tolerated (in the best case) and usually blocked out by firewalls. Besides that, HTTP/SMTP are industry accepted transport protocols that are already supported by enterprise servers and are friendly with firewalls.

The SOAP messaging protocol uses HTTP to carry messages that are formatted with XML. The stated goal of the SOAP specification is two-fold:

- Provide a standard object invocation protocol built on Internet standards using HTTP as the transport and XML for data encoding;
- Create an extensible protocol and payload format that can evolve over time.

SOAP is simple. The client sends a request to a server to invoke an object, and the server sends back the results. SOAP works with existing Internet infrastructure, which means that it is not necessary to make any special accommodations on routers, firewalls, or proxy servers to use SOAP.

4.6 J2EE & CORBA

The Java[tm] 2 Platform, Enterprise Edition (J2EE) defines the standard for developing multitier enterprise applications. J2EE simplifies enterprise applications by basing them on standardized, modular components, by providing a complete set of services to those components, and by handling many details of application behavior automatically, without complex programming.

The Java 2 Platform, Enterprise Edition, takes advantage of many features of the Java 2 Platform, Standard Edition, such as "Write Once, Run Anywhere" portability, JDBC[tm] API for database access, CORBA technology for interaction with existing enterprise resources, and a security model that protects data even in internet applications. Building on this base, Java 2 Enterprise Edition adds full support for Enterprise JavaBeans[tm] components, Java Servlets

API, and JavaServer Pages[tm] technology. The J2EE standard includes complete specifications and compliance tests to ensure portability of applications across the wide range of existing enterprise systems capable of supporting J2EE.

Today's enterprises gain competitive advantage by quickly developing and deploying custom applications that provide unique business services. Whether they're internal applications for employee productivity, or internet applications for specialized customer or vendor services, quick development and deployment are key to success. Portability and scalability are also important for long term viability. Enterprise applications must scale from small working prototypes and test cases to complete 24 x 7, enterprise-wide services, accessible by tens, hundreds, or even thousands of clients simultaneously. However, multitier applications are hard to architect. They require bringing together a variety of skill-sets and resources, legacy data and legacy code. In today's heterogeneous environment, enterprise applications have to integrate services from a variety of vendors with a diverse set of application models and other standards. Industry experience shows that integrating these resources can take up to 50% of application development time.

As a single standard that can sit on top of a wide range of existing enterprise systems -- database management systems, transaction monitors, naming and directory services, and more -- J2EE breaks the barriers inherent between current enterprise systems. The unified J2EE standard wraps and embraces existing resources required by multitier applications with a unified, component-based application model. This enables the next generation of components, tools, systems, and applications for solving the strategic requirements of the enterprise.

Some examples of J2EE Application Servers are Websphere from IBM and Weblogic from BEA Systems. In the open source arena there is JBOSS.

4.7 COMPONENT TECHNOLOGY

4.7.1 ENTERPRISE JAVA BEANS

Enterprise Java Bean (EJB) facilitates the development of distributed Java applications, providing an object-oriented transactional environment for building distributed, component-based, multi-tier enterprise applications. EJB is used to encapsulate business rules and entities. A remote client can invoke the public method of a bean, which typically updates a database or executes a service.

EJB is not a product; rather it is a specification leading and driven by Sun with participation from many key vendors from the industry. This specification defines the EJB component/container architecture and the interfaces between EJB technology-enabled server and the components. Vendors such as IBM, Sun/Netscape Alliance, BEA, Oracle, Inprise, Iona, Fujitsu, Sybase, Gemstone, Persistence and others are providing commercial products that implement the EJB specification.

The EJB container is an environment in which EJB components (the so-called *beans*) are executed. Its first role is to serve as a buffer between a *bean* and the outside world. The container requests from a *bean*, it forwards requests from a client to the *bean*, and it interacts with the EJB server. Furthermore, the EJB Container provides services to the *beans* such as support for transactions, persistence, security, and management of multiple instances of a given bean.

The features of the EJB technology are as follows:-

- Beans are server-side components written entirely in the Java programming language.
- Beans contain business logic only – no system-level programming.
- System-level services such as transactions, security, life-cycle, threading, and persistence, are automatically managed by the EJB Container on behalf of the Beans.
- EJB architecture is inherently transactional, distributed, portable, multi-tier, scalable, and secure.
- Beans are declaratively customised. The customisable issues are the following: transactional behaviour, security features, life-cycle, state management, persistence, and so on.
- Beans are fully portable across any EJB server and any OS.

4.7.2 CORBA COMPONENT MODEL

When developing distributed applications most developers are trying to provide or acquire many of the same underlying services, including security, event notification, persistence, and transactions. These services are critical for giving along-term value for a distributed application. The question that arises is: *how can these services be packaged so that they are easy to use, easy to learn and easy to distribute*?

The CORBA Component Model (CCM) specification was written to address this issue in the CORBA object model. The CCM is a server-side component model for building and deploying CORBA applications. It is very similar to Enterprise Java Beans because it uses accepted design patterns and facilitates their usage, enabling large amounts of code to be generated. This also allows system services to be implemented by the container (the execution framework) provider rather than the application developer. The benefit and need for these types of containers can be observed through the growth of Application Server software. The CCM extends the CORBA object model by defining features and services in a standard environment that enable application developers to implement, manage, configure and deploy components that integrate with commonly used CORBA Services. These server-side services include transactions, security, persistence, and events.

4.7.3 .NET COMPONENT MODEL

.NET is perceived as very different things. Some see .NET as Microsoft's next-generation Visual Studio development environment. Some see it as yet another new programming language (C#). Some see it as a new data-exchange and messaging framework, based on XML and SOAP. In reality, .NET wants to be all of these things, and a bit more.
Here's an itemized list of the technical components making up the .NET platform:

- C#, a "new" language for writing classes and components, that integrates elements of C, C++, and Java, and adds additional features, like metadata tags, related to component development.
- A "common language runtime", which runs bytecodes in an Internal Language (IL) format. Code and objects written in one language can, ostensibly, be compiled into the IL runtime, once an IL compiler is developed for the language.
- A set of base components, accessible from the common language runtime, that provide various functions (networking, containers, etc.).
- ASP+, a new version of ASP that supports compilation of ASPs into the common language runtime (and therefore writing ASP scripts using any language with an IL binding).
- Win Forms and Web Forms, new UI component frameworks accessible from Visual Studio.
- ADO+, a new generation of ADO data access components that use XML and SOAP for data interchange.

4.8 AGENTS AND MULTI-AGENT SYSTEMS

The term agent[1] is often used in database, operating systems, and networking literature to mean a proxy for a computation or a site – a transaction, a process, a network router, that could be used to interact with the underlying entity (Honavar 1999). Agents are computer systems that are capable of autonomous actions (and/or reactions) in some environment in order to meet their design objectives. An agent will typically sense its environment (by physical or software sensors) and will have an available *repertoire* of actions that can be executed to modify the environment, which may appear to respond non-deterministically to the execution of these actions (Wooldridge 1999).

It is important to notice that the agents and the environment they interact with are deeply coupled. Simple agents can often display apparently complex behaviours when they interact with sufficiently rich environments. On one hand, the agents can be reactive or deliberative, pro-active, goal-driven, transient or persistent, autistic or social, rigid or adaptive, pre-programmed or instructive, collaborative or competitive, rational, autonomous or controlled, stationary or mobile, believable, ethical, self-replicating, and so on. On the other hand, the environments can have the following characteristics: observable, controllable, predictable, episodic or nonepisodic, static or dynamic, discrete or continuous, open or closed, and so on.

The topic agent has been evolving since the very beginning of the Artificial Intelligence. Some advances in computer science (e.g. multi-tasking, distributed computing, real-time systems, and networked environments) have paved the ground for the development of agent-based systems. Agent design has been the focus of study in AI over the past four decades. Agent taxonomies have been proposed. Agent languages have been developed. With the explosion of the Internet mobile agents came to play an important role in the agent-related research. Multi-Agent Systems – **MAS**, defined as a loosely coupled network of problem solvers that interact to solve

[1] There is no a precise and consensual definition for agent in the relevant literature – please see (Franklin and Graesser 1997) and (Honavar 1999) for a long list of definitions.

problems that are beyond the individual capabilities or knowledge of each problem solver (Durfee and Lesser 1989) – have started to be designed and implemented. And above all, an international standardisation consortium specifically devoted to agents was created: the FIPA.
The Foundation for Intelligent Physical Agents (FIPA) was formed in 1996 to produce software standards for heterogeneous and interacting agents and agent-based systems. The core mission of the FIPA standards consortium is to facilitate the interworking of agents and agent systems across multiple vendors' platforms. This is expressed more formally in FIPA's official mission statement:

> *The promotion of technologies and interoperability specifications that facilitate the end-to-end interworking of intelligent agent systems in modern commercial and industrial settings.*

The emphasis here is on practical commercial and industrial uses of agent systems. However, FIPA is also focussing on intelligent or cognitive agents, that is, software systems that may have the potential for reasoning about themselves and/or other systems that they encounter. FIPA believes that through a combination of speech acts, predicate logic and public ontologies, it is possible to offer standard ways of interpreting communication between agents in a way that respects the intended meaning of the communication.

To support this, FIPA has adopted and is working on specifications that range from architectures to support agents' communicating with each other, communications languages and content languages for expressing those messages and interaction protocols which expand the scope from single messages to complete transactions. In the future, there are plans to extend this even further to cope with longer term relationships between agents.

The FIPA's Agent Management Specification provides the normative framework within which FIPA agents exist and operate. It establishes the logical reference model for the creation, registration, location, communication, migration and retirement of agents. Several platforms have been developed based on this specification (e.g., Zeus, JATLite, and JADE[2]) and interoperability between them has been demonstrated.

Among these platforms, JADE (Java Agent Development Framework) is becoming quite widely and worldwide used (already being used in IST projects, namely CoMMA and LEAP). It is a software framework fully implemented in Java language. It simplifies the implementation of multi-agent systems through a middleware that claims to be compliant with FIPA specifications and through a set of tools that supports the debugging and deployment phase. The agent platform can be distributed among heterogeneous machines (which not even need to share the same OS) and the configuration can be controlled via a remote GUI. The configuration can be even changed at run-time by moving agents from one machine to another one, as and when required.

Another platform is named JATLite (Java Agent Template, Lite). JATLite is a package of programs written in the Java language that allow users to quickly create new software agents

[2] http://www.btexact.com/projects/ibsr/technologythemes/agentplatforms.htm, http://liawww.epfl.ch/~calisti/ACL-LITE/, http://sharon.cselt.it/projects/jade/.

that communicate robustly over the Internet. JATLite provides a basic infrastructure in which agents register with an Agent Message Router facilitator using a name and password, connect/disconnect from the Internet, send and receive messages, transfer files, and invoke other programs or actions on the various computers where they are running.

JATLite facilitates especially construction of agents that send and receive messages using the standard communication language, KQML (see http://www.cs.umbc.edu/kqml/ for the current KQML standard). The communications are built on open Internet standards, TCP/IP, SMTP, and FTP. However, developers may easily build agent systems using other agent languages using JATLite.

5 COMMUNITY BUILDING

The Web has transformed communications and greatly mitigated the impact of geographic and time separation. Developments in rich media (e.g. video, live cameras, digitised photography, animation and others "new media") in conjunction with conversational technologies (e.g., instant messaging, e-mail) have enabled us to closely simulate conversation and face-to-face social exchanges. This has "humanized" the Web as a forum of people to interact and has provided an alternative to being together in space of time.

As result of that, new social structures and relationship styles are emerging as people separated by time and space are connecting and discovering common interests, shared spaces, and adopting online communities for communicating and exchanging ideas or knowledge.

These virtual communities can take many forms and different purposes. Some examples of that are:

- Newsgroups
- Discussions and technical forums, Message boards or Conferencing
- Product reviews and ratings
- Mailing lists
- Expert seminars
- Virtual meetings
- Newsletters
- Customized home pages and portals
- Chats

The benefits from these activities include the following:-

- Recognition. Approval and positive feedback from other community members.
- Learning. Increased knowledge.
- Best practices. Performance improvement.
- Time/effort saving. Development of new or better techniques and processes.
- Alignment along key objectives. Development of common goals and objectives among community members.
- Belonging. Meeting and interacting with other people interested in the same topic o issue.

The growth in demand for Online Communities has produced the corresponding increase in the number of companies offering community technologies. This competition has triggered more diverse offering and many improvements centred on three key themes: Focus, Integration and Interaction.

While some virtual communities represent actual physical communities, others are being developed that exist solely online.

Most successful models of engineered communities involve one of two starting approaches:

- Build a community to focus on generating revenue. For example, eBay built a business model based on peer-to-peer relationships, which enable community to form.
- Exploit an existing customer group "owned" by enterprise. For example, advertise within a customer group, such as the viewers of a television show or the fans of a specific music group.

Along these lines, models based on peer-to-peer relationship have experimented an enormous success in the last years. Good examples of theses communities are Gnutella or KaZaA, two applications that base their models in a free exchange of files between peers. Although both of them use the Internet as their network, their users also need to be install software, which allows delivery of advertisements .

One of the biggest challenges that many communities have is how to make contact with others having a specific expertise. The development of dedicated skill directories where individual could provide data regarding their level of expertise has demonstrated not to be always useful.

To help address this issue, a new generation of expertise technology has emerged, which incorporates "passive profiling". These systems use text-mining technologies to analyse the content in e-mail, instant messaging and other repositories to develop profiles of individual interests. Individual can then review and modify these profiles and allow the system to update the profile dynamically, based on additional content. These profiles allow individuals to query others in the community and identify those willing and able to provide insight on a range of topics.

Perhaps the biggest problem facing the development of a community in a virtual environment is the difficulty associated with building a common set of assumptions and understandings. In physical settings, the interactions around common artefacts, or tools of the trade, makes it easier to develop reference points that every one in a conversation can share. For example, when attempting to fix a broken piece of equipment, technicians working face-to-face all see the setting in which the equipment exists, sense environmental conditions that could potentially affect the operation of the machine, and point to potential tools that they could use to solve the problem.

In a virtual world, building the common context necessary for effective knowledge sharing is significantly difficult. The lack of environmental clues, compounded by the variety of assumptions that can be associated with cultural differences and language barriers, can significantly hinder the knowledge transfer process. A number of technologies can help overcome this barrier, however.

In this context, Videoconferencing offers enough "face-to-face" interaction and environmental information to help build a joint view of a given situation and allowed members of globally dispersed teams to understand and solve problems together.

Another similar technology is "whiteboarding", which allowed geographically separate individuals to jointly view and comment on a specific document. This ability to see changes

made to a report or a presentation simultaneously can help that everyone is literally “reading off the same page”.

In short, online communities are one of the most used media to share and increment knowledge. Even for those companies that want to promote their products and services on the Web, virtual communities have demonstrated be a powerful business tool.

5.1.1 COMMUNITY BUILDING TOOLS

Examples of Community Building Tools are: Dnews & DnewsWeb,Dr.B’s Virtual Message Board Tool,Emaze forums,Hypernews,Inchima,Place Ware,Site Scape Forums,UUB and yBulletin

6 INFORMATION/KNOWLEDGE MANAGEMENT AND INDEXING TECHNIQUES

This section will study how the system information is managed by way of indexation and ontology. How is information correlated so that when a user makes a request, information of interest can be retrieved for him.

6.1 ARCHITECTURAL REQUIREMENTS

The most relevant architectural requirements are listed below:

6.1.1 WEB-CENTRED ENVIRONMENT

Knowledge Management System is intended to run in a Web-centred environment. As such, all the Web-related matters have to be considered as part of the architectural requirements (distributed environment, security issues, thin clients, communication protocols, and so on).

6.1.2 ONTOLOGY-BASED

The Ontology is a key element in the Knowledge Management System architecture. It holds the definition of the valid *concepts* in a given domain. Besides that, intends to develop an Ontology Server capable of handle, initially, multiple views over one single ontology. In a second level, this component is expected to manage multiple ontologies and the required mapping among them.

6.1.3 PUSH MECHANISMS AND AUTONOMOUS PROCESSES

Knowledge Management System has to consider including some push mechanisms to handle them properly. For instance, when a new regulation is released in a Web site being monitored by a given company, the users interested in that document might be informed. This issue is related with the autonomy (at least partially) of the Knowledge Management System. Considering the previous example, Knowledge Management System might have an autonomous process running in background monitoring such a situation. Another very typical scenario is the one in which a *crawler* is searching for knowledge items in the Internet. Its work could be complemented by another process indexing the gathered knowledge items.

6.1.4 OPENNESS

An open (i.e., adequate, ease to use, flexible) mechanism to support the integration of such tools in the Knowledge Management System operation. Web-based standards, such as SOAP and XML, may play a crucial role in this topic.

6.1.5 CONFIGURABILITY

Each user may have a very particular way of managing knowledge based on cultural issues, strategic matters, and so on. Therefore, the Knowledge Management System has to be as configurable as possible.

6.2 COMPONENTS

The User Portal
A Web-based portal used to capture, browse and search knowledge across Project, Corporate, Domain, and User layers. Such a portal is intended to mainly support the users of the infrastructure.

The Kernel
This component plays the role of a "backbone", where the KM services, the Ontology-related services, and the Auxiliary Services are registered and deregistered. It also provides access to the Knowledge Base.

The Knowledge Base
This knowledge repository will be used to store the knowledge representation for the "knowledge items" managed within the infrastructure. In , a *Knowledge Item* (KI) is intended to represent an indivisible unit of knowledge. Indeed, a KI is an "abstract entity" represented by ontological concepts defined in the Ontology. Preliminarily, a KI may represent *documents*, *areas of expertise*, *roles*, *knowledge classifications*, and *communities of interest*.

The KM Services
These services have been structured in six categories of services, according to the knowledge life cycle, namely: *capture/extraction*, *cleansing/transformation*, *indexing*, *refreshing*, *discovery*, and *distribution/dissemination*. This list should not be seen as exhaustive and other services may be integrated into or removed from the Knowledge Management System according to the end user requirements.

The Ontology Server
This component is intended to handle all the ontology-related issues within the architecture, such as: ontology registration, processing of ontology-related queries, ontology maintenance, and so on.

The Wrapper
This component is intended to support the interoperation among third party applications and the infrastructure. However, taking into account that this infrastructure will fundamentally operate in a Web-centred environment, the Wrapper is likely to be converted into an integration mechanism based on UDDI, WDSL, and SOAP. In this case, a formalised wrapper will only be required in the event that non web-enabled services require integration.

6.3 ACTIVITIES

6.3.1 KNOWLEDGE REPRESENTATION

The knowledge representation is a very strategic functionality in solution. It states the need of designing models and/or meta-models representing the *knowledge* being considered by . These models will be used by the search engines during the knowledge search/retrieval process.

The Ontology plays the central role in the knowledge representation. It contains *concepts* that represent meaningful elements in a given domain and the *signatures* representing a given relation involving a pair of concepts. A piece of knowledge, called Knowledge Item in , is indeed fully represented by ontological concepts.

6.3.2 KNOWLEDGE ACQUISITION

Considering that the KM process is about to begin, this represents the first operational step. Knowledge acquisition can be viewed as extraction and capture of knowledge. On one hand, Extraction, in , deals with the explicit knowledge and means the identification and the extraction of patterns within the existing data, which will represent the acquired knowledge. Broadly speaking, this can be considered an automatic (or semi-automatic) process that parses and extract knowledge from a given source of explicit knowledge. On the other hand, Capture in is related to the transformation of the implicit knowledge in explicit one. This is represented, for instance, by a process in which an expert explicitly register the knowledge he/she possesses.

6.3.3 KNOWLEDGE CLEANSING/TRANSFORMATION

After being captured / extracted, sometimes the knowledge needs to be filtered or transformed in order to be "enriched" especially when it has been captured from a external source (such as the Internet). This function ensures that only meaningful knowledge items are retained, and would be performed by a **person** responsible for the "quality" of the knowledge acquired.

6.3.4 KNOWLEDGE INDEXING

Before being used the knowledge can be classified according to a given "knowledge classification", which can be based on the ontological concepts. For instance one can think about having a knowledge classification grouping KIs by the phase in which they belong to during the development of a construction project. Later on, this classification can be used to provide a more accurate support to the dissemination of the knowledge, helping to target the right people.

6.3.5 KNOWLEDGE UPDATE

This functionality is related to both Ontology-related issues and the KIs acquired. The ontology can evolve and the KIs may change with time. This functionality represents the need to keep them updated considering a manual or an automatic updating process.

6.3.6 KNOWLEDGE REFRESHING

The update of a given KI or some aspect in the Ontology may have consequences to other elements within the "knowledge domain". This is where the knowledge refreshing is utilised and means that it takes care of the consequences of an update action. For instance, if a given KI is enriched with a list of ontological concepts, this might affect its knowledge classification.

6.3.7 KNOWLEDGE SEARCHING/DISCOVERY

Searching for knowledge represents the acquaintance of knowledge within certain sources of knowledge previously known. For instance, documents and databases, which belong to organisation's Intranets. Discovered knowledge is related to the acquaintance of knowledge from external sources of knowledge, mainly the Internet. Here, crawler agents can try to discover and bring to the organisation any type of relevant knowledge. For instance, the relevance of the knowledge can be measured by the "knowledge classification".

6.4 KM-RELATED TECHNOLOGIES

This section identifies some technologies for each one of the functionalities described in the previous section. It is important to bear in mind that sometimes the technologies can be applied to a number of functionalities.

6.4.1 KNOWLEDGE REPRESENTATION

Knowledge representation is one of the most strategic points in with the Ontology as the core topic. Ontology, by itself, does not represent a technology rather a more conceptual and philosophical area. However, after being worked by different research projects and international initiatives[3], there are currently some technologies that have be used to support the creation and maintenance of Ontologies. For instance, both semantic networks and relational databases can be used to hold ontological definitions. Therefore, before describing the technologies, this section starts by making a more conceptual presentation about Ontology in the context. This is followed by a potential list of technologies that can be used to support the representation of both Ontology and Kis, namely *Rule-Based Systems*, *Case-Based Reasoning*, *Semantic Networks*, *Databases*, and *Data Warehouse*.

6.4.1.1 Ontologies and Related Concepts

According to the European KM Forum (KM Forum 2001), an ontology is "an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them". Basically, an ontology allows the representation of useful notions and constraints expressing knowledge about a given domain. It is not a "knowledge repository" (as a database is) and it does not contain any knowledge, rather it can be viewed as a frame indicating how to represent, formulate knowledge about an activity, an organisation or a system in general. An ontology

[3] In US: DAML; in Europe: Onto-Knowledge, CoMMa, eConstruct, just to name a few.

helps to formulate correct assertions but it is not intended to check if these well formulated assertions are true or false.

The view about an ontology is depicted in Figure 1. The process of modelling an ontology starts with the enumeration of relevant concepts that are useful to describe it. Each concept will be labelled with a unique identifier. In order to facilitate the comprehension of the meaning of each concept by human beings, this identifier can be based on a combination of words used daily. This set of identifiers represents a **Vocabulary**. A definition (for instance, in natural language) is attached to each identifier and this produces a dictionary or a glossary.

Identifying and naming the relevant concepts in a given domain is a complex exercise. A good way to proceed is to classify these concepts into a hierarchical structure, creating a **Classification**. This hierarchy, which is actually a tree structure, must enable a multi-inheritance mechanism in order to allow the expression of a multi-dimension space in a 2D diagram.

If the way to classify is based on the use of the relation "**is a**" (for instance the concept "person" "**is a**" "human being"), the tree produced as result of such a classification is called a **Taxonomy** which is than a special way of classifying things.

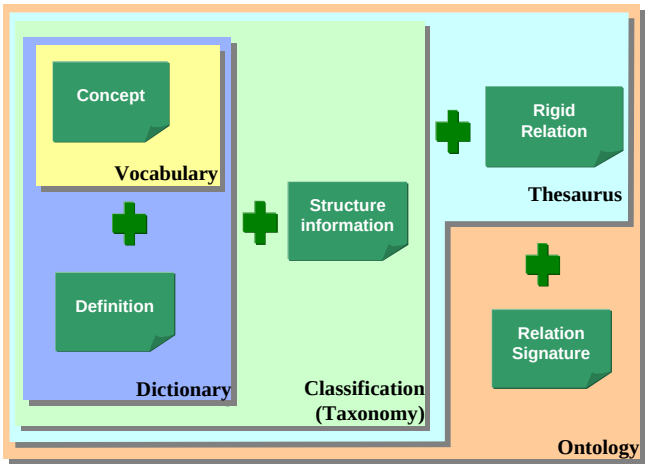


***Figure 1** – Illustration of the Ontology concept*

The use of the unique relation "**is a**" is not enough to model a complex system. Extra relations exist "de facto" between concepts even if these concepts are not closely defined in the taxonomy tree. This leads to the definition of a different structure (more complex than a tree) to express these semantic relations. One can consider that the "**is a**" relation is a semantic relation but to keep things simple, this particular relation will only be called a hierarchical relation. These semantic relations enable the expression / representation of a domain specific knowledge. A relation, called a **Signature,** may bind only two concepts. The notion of signature is very important. It allows the linking of each concept with any other existing concept within the ontology. The liaison of concepts can be done freely in order to really stick with the domain being represented by the ontology.

A **Thesaurus** can be viewed as a subset of an ontology, where the whole structure (hierarchical and semantic relations) is rigidly defined. The first consequence is that the semantic content in the thesaurus is not so rich because of this rigid structure of relations applicable to the concepts. Only high level relations such as the notions of close or far neighbourhood can be represented. The addition of a specific relation to link two given concepts is not allowed.

6.4.1.2 Rule Based System

Rule Based System aims to represent the knowledge as rules that describe pieces of a logical process. These rules are processed by an inference engine that can use combinations of rules to find some results. In other words, a rule-based system is composed by a sequence of IF-THEN *rules*, a group of *facts*, and an inference engine (*interpreter*). The facts in the system are represented in the *working memory*, which is continually updated. The *Rules* represent possible actions to be taken when predefined conditions are detected on the facts existing in the working memory (these rules are sometimes called *condition-action rules*). The conditions are usually *patterns* that must match facts in the working memory, while the actions usually involve *adding*/*deleting* facts to/from the working memory. The interpreter controls the application of the rules and updates the working memory.

There are two types of Rule Based Systems, namely *forward chaining* and *backward chaining* systems. In the former, the system starts with the initial facts and try to draw new conclusions applying the rules on the facts. In the latter, the system tries to prove some hypothesis using the rules in order to verify the initial facts.

6.4.1.3 Case Based Reasoning

Case Based Reasoning (CBR) is a methodology to model human reasoning and thinking in order to build intelligent computer systems that are able to find solutions for new problems, based on previous experiences. Simply stating, CBR systems solve new problems by adapting solutions that were used to solve old problems (Riesbeck & Schank 1989). The CBR systems' expertise is embodied in a library of past cases. Each case typically contains a description of the problem and the solution achieved and/or the outcome produced.

More information about this topic is available at http://www.aiai.ed.ac.uk/links/cbr.html.

6.4.2 INDEXING KNOWLEDGE

After being acquired, the knowledge needs to be indexed before being used. There are two techniques namely *Decision Trees* and *Inverted File Index*.

7 SECURITY

This will study how access to the system is controlled. Most systems use a system of user ID and password to control access. Many systems use Kerberos from MIT to encrypt data flowing between client and server

7.1 HTTP

The Hypertext Transfer Protocol (HTTP) is an application-level protocol for distributed, collaborative, hypermedia information systems. It is a generic, stateless protocol, which you can use for many tasks other than hypertext. For example, you can use it for name servers and distributed object management systems by extending its request methods, error codes and headers. A feature of HTTP is the typing and negotiation of data representation, allowing systems to be built independently of the data being transferred.

The HTTP protocol is a request/response protocol. A client sends a request to the server in the form of a request method, URI and protocol version, followed by a MIME-like message containing request modifiers, client information and possible body content over a connection with a server. The server responds with a status line, including the message's protocol version and a success or error code, followed by a MIME-like message containing server information, entity meta information and possible entity-body content

7.2 HTTPS

The Hypertext Transfer Protocol with SSL (HTTPS) is also known as Secure Hypertext Transfer Protocol, or S-HTTP. A secure socket layer is an encryption protocol invoked on a Web server that uses HTTPS. It provides secure communication mechanisms between a HTTP client-server pair to enable realtime commercial transactions for applications.
How does HTTPS work?

The protocol does not require asymmetric key pairs (public and private keys), although this mode is supported. Your application can implement symmetric key operations only. In this mode, individual users do not have to request a public-key certificate from a third party before making a transaction. However, the symmetric key, or session key, does need to be transmitted out of band before a transaction.

HTTPS supports end-to-end secure transactions, in contrast with the original HTTP authorization mechanisms, which require the client to attempt access and be denied before the security mechanism is employed. It provides full flexibility of cryptographic algorithms, modes and parameters.

7.3 PASSWORD

Generally access to a system is controlled by means of allocating a unique user ID and a password to each user.

The user has control of his password and can change it as many times as he wishes.

7.4 KERBEROS

Kerberos is a security system that helps prevent people from stealing information that gets sent across the wires from one computer to another. Usually, these people are after a password.
The name "Kerberos" comes from the mythological three-headed dog whose duty it was to guard the entrance to Hell. The Kerberos security system, on the other hand, guards electronic transmissions that get sent across the Internet. It does this by scrambling the information -- encrypting it -- so that only the computer that's supposed to receive the information can unscramble it. In addition, it makes sure that your password itself never gets sent across the wire: only a scrambled "key" to your password.

Kerberos is necessary because there are people who know how to tap the lines between computers and listen for your password. They do this with programs called "sniffers", and the only way to stop them would be to physically guard every inch of the Internet ... computers, cables and all.

7.5 FIREWALLS

A firewall is a gatekeeper computer or software that sits between the Internet and a private network or computer. It ptotects the private system by filtering traffic to and from the Internet based on policies defined by the user.You use the firewall to who can get onto your network and when.

A firewall typically provides two network interfaces- one connects to the internal protected system,and the other connects to the external unprotected network. Proxy firewalls – also known as application firewalls – are the most secure form of firewall.

7.6 INTRANET

Lans (Local Area Networks) are fast private extensions of the Internet. Geographically distributed Intranets can be connected via the Internet backbone.

7.7 EXTRANET

An extranet extends the corporate backbone to outsiders using standard Internet technology and ISPs; it is also a way for corporate intranets to speak to one another. The idea is to create a Virtual Public Network (VPN) on top of the public Internet to tie a corporation with its suppliers,employees and business partners anywhere in the world.

8 CONTENT STANDARDS

8.1 METADATA

8.1.1 METADATA FORMATS

Library services are managing the metadata in a variety of schemas and formats. By the term schema we understand *what* about a particular resource (e.g. book or paper) is managed. By the term *format* we mean how this data is presented on screen, in papers, in files or when exchanged between software. Standard or common metadata schemas and formats enable interoperability among various electronic library systems as well between those systems and the authoring environments where the metadata is used for citation proposes. There are hundreds of metadata formats that are an output of various bibliographic databases.

The SciX solution should support at least one of these.

8.1.1.1 The Dublin core

Dublin Core metadata is used to supplement existing methods for searching and indexing Web-based metadata, regardless of whether the corresponding resource is an electronic document or a "real" physical object.

The Dublin Core Metadata Element Set (DCMES) provides a semantic vocabulary for describing the "core" information properties, such as "Description" and "Creator" and "Date".
Dublin Core metadata provides card catalog-like definitions for defining the properties of objects for Web-based resource discovery systems. The DCMES is a set of 15 descriptive semantic definitions. It represents a core set of elements likely to be useful across a broad range of vertical industries and disciplines of study.

The Dublin Core Metadata Element Set was created to provide a core set of elements that could be shared across disciplines or within any type of organization needing to organize and classify information.

8.1.1.2 MARC

MARC is an acronym for *Ma*chine *R*eadable *C*atalogue or *C*ataloguing. MARC is a short and convenient term for assigning labels to each part of a catalogue record so that it can be handled by computers. While the MARC format was primarily designed to serve the needs of libraries, the concept has since been embraced by the wider information community as a convenient way of storing and exchanging bibliographic data.

The original MARC format was developed at the Library of Congress in 1965-6 leading to a pilot project, known as MARC I, which had the aim of investigating the feasibility of producing catalogue data in machine-readable form. Similar work was in progress in the United Kingdom where the Council of the British National Bibliography had set up the BNB MARC Project with

the remit of examining the use of machine-readable data in producing the printed *British National Bibliography (BNB)*. These parallel developments led to Anglo-American cooperation on the MARC II project which was initiated in 1968. MARC II was to prove instrumental in defining the concept of MARC as a communication format.

8.1.1.3 OCLC

The OCLC system uses eight MARC formats: Books (BKS), Serials (SER), Visual Materials (VIS), Mixed Materials (MIX), Maps (MAP), Scores (SCO), Sound Recordings (REC), and Computer Files (COM) and defines a syntax in which metadata is represented.

8.1.1.4 BibTex

BibText is an ASCII format support by the Tex/Latex tools.

```
@ARTICLE{Kay:84,
 author = {Alan Kay},
 title = {Computer Software},
 journal = {Scientific American},
 year = 1984,
 volume = 251,
 number = 3,
 month = {September},
 pages = {41-47},
}
```

Figure 2: Example of some Metadata in BibTex format.

8.1.1.5 Web Syndication Metadata

The following are some of those commonly used in WEB Syndication

Information and Content Exchange (ICE)
(http://www.icestandard.org).

The Information and Content Exchange (ICE) is an XML-based protocol for content syndication and subscription between a syndicator and a subscriber. ICE is a message-based (request/response) protocol that assumes there exists a relationship and an agreement between the syndicator and the subscriber about the content, its vocabulary and format. The ICE group expects vertical industries to supply the necessary vocabularies (e.g. ontology.org, BizTalk, RosettaNet etc.), but the protocol itself is indifferent about the actual content being exchanged. Specifically “ICE manages and automates establishment of relationships, data transfer, and result analysis. When combined with industry specific vocabulary, Ice provides a complete solution for syndicating any type of information between information providers and their subscribers”.

ICS supports functionality for large scale negotiated content syndication and is targeted towards commercial businesses, which distribute or aggregate content in complex business relationships (e.g. product data exchanged in a manufacturer-wholesaler-retailer supply chain).

ICE is a "open layered protocol for establishing and managing controlled data exchange between business partners" that is supported by most high-end syndication platforms such as Vignette, Kinecta syndication server and ShiftKey CiClone with many others to follow. ICE is backed by heavy hitters like Vignette, Adobe and Oracle.

Rich Site Summary or RDF Site Summary (RSS)
RSS 1.0 is a lightweight multipurpose extensible metadata description and syndication format that conforms to the W3C's RDF Specification. XML-namespace and RDF allow modularisation of RSS and extensions to be built into the vocabulary. RSS 1.0 attempts to bridge the need for lightweight syndication format and the capability of representing rich content. Being modular provides capabilities for application specific functionality (e.g. vertical markets) without affecting the RSS core specification. Three core modules (namespaces) are included with the RSS distribution, the Dublin Core Meta-data module for describing basic web resources, the syndication module for controlling syndication and the content module for describing actual content, taxonomy, categorisation etc. Additionally RSS, being based on RDF, an extensible meta-framework for describing and interchanging meta-data, provides the capability to represent semantic information about resources and their relationships in a flexible and extensible manner.

Open Content Syndication (OCS) Directory format
(http://internetalchemy.org/ocs/directory.html)**.** The OCS Directory format is a development project by Internet Alchemy (http://internetalchemy.org). "The OCS Format is intended to provide a concise, machine readable-listing of a set of syndicated channels. The OCS is designed to enable an OCF Directory of channels to be syndicated for use by portal sites e.g. JetSpeed Apache Portal project (http://jacarta.apache.org/jetspeed), client based headline software e.g. Reptile client (http://reptile.openprivacy.org/) and other similar applications. The OCS Directory aggregates information about channels into a directory (channel listings). The directory may contain information about channels from multiple sites, each with multiple channels and each channel listing may be syndicated in multiple formats (e.g. RSS or ScriptingNews), languages and publishing schedules.

The OCS is an application of XML that consists of three namespaces the W3C RDF specification for expressing the relationship between OCS items, the Dublin Core (CD) for describing the origin and content of items (e.g. title, creator, description, subject and language) and the OCS namespace. The OCS namespace contains only six elements which of, only one is a mandatory element, the "format" element. The "format" element contains the link to the definition of the channel format being described by the item. The optional elements "image" and "contentType" contain a URL for a channel logo and the MIME type of the content respectively, while the remaining elements "updatePeriod", "updateFrequency" and "updateBase" handle scheduling of channel updates.

Channel Definition Format (CDF)
(http://www.w3.org/TR/NOTE-CDFsubmit.html) is a development of Microsoft. The Active Channel technology introduced by Microsoft in Internet Explorer 4.0 and the Channel Bar Windows desktop application was based on the CDF specification. CDF is an open specification, an XML vocabulary that enables smart-pull or true-push delivery of content for

offline browsing. CDF defines logical groupings to organize a set of related Web documents into hierarchies (channels) and scheduling delivery of channels. Using CDF, channel publishers and users can schedule automatic information delivery and content updates by webcasting (i.e. web publishers broadcast content to their clients on a regular schedule). CDF is suitable for delivering site news and bulletin information, dividing a site into more manageable subsets, notifying updates of web pages and providing site navigation.

8.1.2 OTHER FORMATS

8.1.2.1 MCF

The Meta Content Framework (MCF) provides a system for representing a wide range of information about content. The content targeted includes web pages, gopher and ftp files, desktop files, email and structured (i.e., relational and object oriented) databases, etc. MCF is not intended to be an extension of markup languages such as HTML which can be used to hold embedded metadata. Instead it provides a format for holding the metadata externally to the content described. It is possible that metadata embedded in content will be extracted automatically by robots that use the MCF to represent the results of their activities. MCF should be able to represent the metadata that proposals such as the Dublin Core aim to cover.

8.1.2.2 PICS

Platform for Internet Content Selection (PICS - http://www.w3.org/PICS/) is an infrastructure for associating labels with Internet content. It was originally designed to help parents and teachers control what children access on the Internet, but it also facilitates other uses for labels, including code signing, privacy, and intellectual property rights management.

8.1.2.3 HTML META TAGS

Meta tags in the <HEAD> section of HTML documents provide a way to add metadata information to Webpages. SciX generated pages should include META tags, preferably using the Dublin Core schema.

8.2 CITATIONS

8.2.1 CITATION AND REFERENCE MANAGEMENT SOFTWARE

Compatibility with this software is important for any digital library. Examples of this software include:

- Reference Manager (http://www.refman.com/)
- Endnote (http://www.endnote.com/) (see section 5.5)
- ProCite (http://www.procite.com/)
- BibTeX is a program and file format designed for the LaTeX document preparation system. (http://www.ecst.csuchico.edu/~jacobsd/bib/formats/bibtex.html)
- Groff (http://www.gnu.org/software/groff/groff.html)

- Refer is a program and a file format used by the troff document formatting package. (http://www.ecst.csuchico.edu/~jacobsd/bib/formats/refer.html)
- Biblioscape (http://www.biblioscape.com/)
- Bibliographix (http://www.bibliographix.com/)
- Powerref (http://www.cheminnovation.com/powerref.html)
- RefWorks (http://www.refworks.com/)

8.2.2 CITATION STYLES

The most common citation styles include:

- Vancouver (number citation system).
- Harvard (author-date system).
- References are compiled using these syntaxes:
- Cambridge
- Chicago
- APA (American Psychological Association)
- AGPS (Australian Government Publication Service)
- MLA (Modern Languages Association)

8.2.3 ENDNOTE

EndNote is a bibliographic management software program that allows one to maintain a database of references and citations and further reformat them according to different bibliographic styles. EndNote can also import, reformat, and store reference information retrieved from remote bibliographic databases, files, and catalogues. Additionally, EndNote allows the creation of bibliographies instantly and automatically from the references stored in your personal EndNote library database.

EndNote is tool that integrates the following tasks into one program:

- Store and organize references in a personal database library
- Search free bibliographic databases on the Internet
- Import reference information from database files, servers, and catalogues
- Insert citations from your EndNote library into word processing documents and create bibliographies of any format for those documents.

The EndNote database software was designed for the specific purpose of storing bibliographic information. EndNote can access and search many Z39.50-compliant databases around the world.

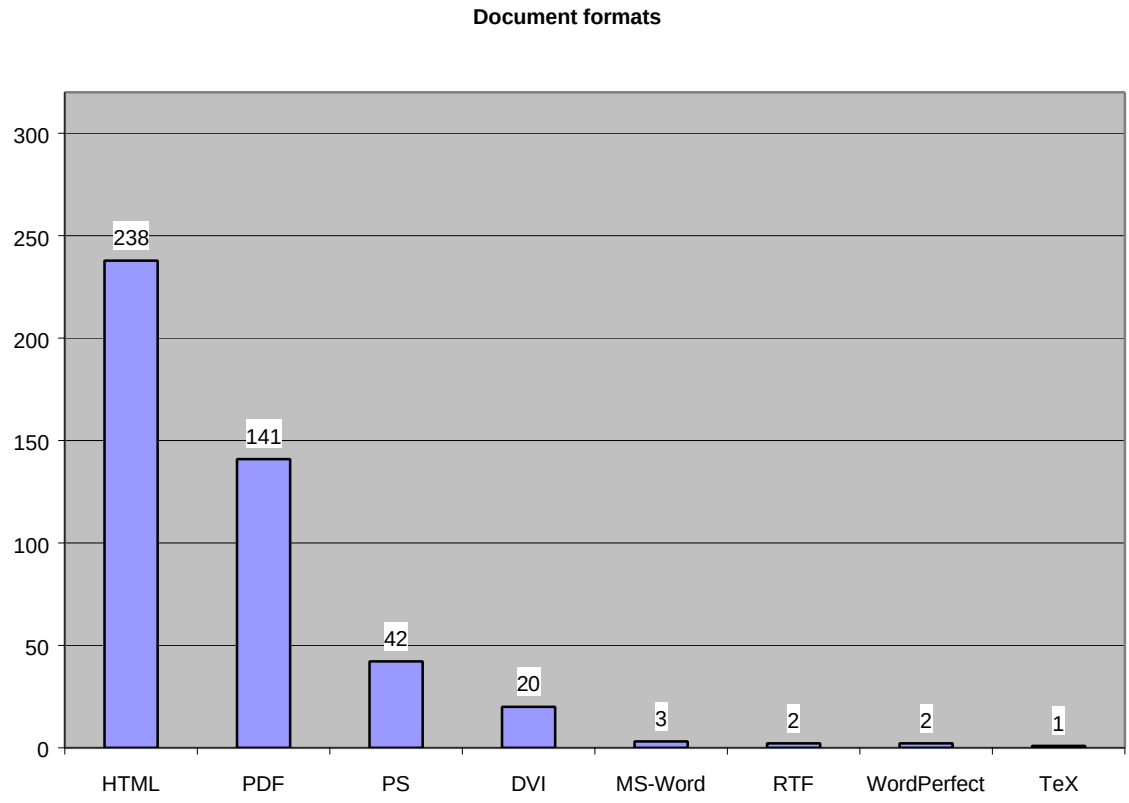
EndNote is compatible with many word processing document types, which include

- Microsoft Word
- WordPerfect
- Rich Text Format (RTF)
- Ami Pro 3.0
- ANSI Text
- HTML

8.3 FULL TEXT

In the study on the availability of free electronic refereed journals, described in "Scientific Publishing: As-is Business and Information Model" (Bjork, Hedlund & Gustafsson), different document formats was one of the data collected from the sample. In below is presented the number of times the different formats appeared in the sample consisting of 318 journals.

Table:



8.3.1 BRIEF GLOSSARY FOR DIFFERENT DOCUMENT FORMATS

HTML:(HyperText Markup Launguage)	A common language used to create documents for the world wide web, interpreted by browsers. Besides an Internet browser, no other interpreting programs are needed.
PDF: (Portable Document Format)	Format that displays on any computer regardless of the software the original was created in. A PDF reader created and freely distributed by Adobe, is required in order access the document.

PS: (PostScript)	A format created by Adobe that will precisely read graphics and fonts. Therefore suitable in creating, and reading, mathematical formulas etc.
RTF: (Rich Text Format)	A document format that allows the exchange of text files between different word processors on different operating systems.
TeX**:**	A typesetting system developed especially in order to create and read text, containing mathematics and other formulas.
DVI: (DeViceIndependent)	A format developed from the TeX
MS-Word â :	Format used by the “Microsoft Office Word” word processor.
WordPerfect â :	Format used by word processor developed by Corel.

9 INITIAL REQUIREMENTS ANALYSIS

This section describes the initial requirements identified for the Core Repository from the point of view of various classes of user.

9.1 AUTHORS

9.1.1 UPLOAD A PAPER/SUBMITTING FOR PUBLICATION

Authors need to be able to upload papers, including the input of metadata, with a minimum of effort. Consideration needs to be given to automation of some aspects of the metadata definition. Metadata needs to include information on the editorial status of the article (e.g. draft, version, etc.) Cross-referencing between papers needs to be carried out automatically from citation lists, and appropriate user interaction will need to be provided to resolve ambiguities and uncertainties in cross-referencing. Formats to be permitted should include: PDF, PS, MSWord, Plain text, and TEX

9.1.2 REMOVE OWN PAPER

Authors need to be able to remove their own articles from the repository after authentication and authorisation. Suitable audit trails will need to be kept in relation to the addition and removal of articles, and the legal aspects of this need to be explored further.

9.1.3 VERSIONING

Support needs to be provided for uploading different versions of articles and maintaining links between the versions.

9.1.4 CREATING REFERENCES FOR INSERTION INTO OWN PAPERS

A convenient means should be provided for authors automatically to generate suitable citations/references for inclusion in their own articles. This will improve the reliability of cross-referencing between the articles in the database.

9.1.5 TRACKING/NOTIFICATION

Facilities need to be provided to notify authors of: citations of their own work in other authors' articles, and comments/discussions by readers. Information needs to be provided to authors regarding the level of readership of their articles.

9.1.6 SOME METHOD TO VERIFY/INDEMNIFY AGAINST COPYRIGHT ISSUES.

A means needs to be provided to ensure that the owners of a SciX server are indemnified against IPR and copyright issues arising from the electronic publication of articles. This may

take the form of a copyright declaration certified by the author, but legal aspects of this need to be investigated further.

9.2 READERS

9.2.1 RETRIEVE AN ARTICLE

The most basic functionality is to retrieve an article based on its unique identifier.

9.2.2 SEARCH FOR PAPERS

Readers need to be able to carry out keyword searches for articles based on metadata and on free text searching.

9.2.3 BROWSING

Readers need to be able to "browse" the contents of the repository by following citations (in both directions) and/or by searching for articles similar to ones that they have already identified. Similarity will be determined by means of suitable clustering algorithms.

9.2.4 PROFILING AND NOTIFICATION

The system needs to be able to maintain a profile of interests for each registered user and provide notification when a new paper is uploaded that matches their profile.

9.2.5 COMMENTS/REVIEWS AND DISCUSSION

Provision needs to be made for readers to comment on and discuss articles publish in the repository. This could be provided in a form similar to the Usenet-news discussion forums. They may need facilities for referencing sections of articles or copying and pasting sections of articles in comments.

9.2.6 SOCIALIZATION

Provision may be made for registered users to locate other registered users with similar interests, based on their profile and searching habits. (It is not proposed to implement this functionality in the first iteration of the SciX system.)

9.2.7 ANONYMITY IF REQUIRED

Readers can have anonymous access for the basic functionality. If they wish to use higher functionality, such as commenting, etc., then they would have to register.

9.3 INDUSTRY READERS/DIGEST WRITERS

9.3.1 SUPPORT FOR DIGEST WRITERS

Support needs to be provided for people to produce digests, from academic articles published in the SciX repository, for the consumption of industry readers. This should include facilities for copying and pasting, and tracking which sources have been used to produce a digest.

9.3.2 INDUSTRY RELEVANCE RATING

A mechanism needs to be provided to indicate the relevance of an article to an industry (non-research) reader.

9.4 JOURNALS/JOURNAL EDITORS

Much of the functionality required here relates to the submission and review process.

9.4.1 MANAGING EDITORIAL BOARD MEMBERS

Editors need to be able to record and manage membership of the editorial board of an online journal, with appropriate profile information.

9.4.2 SELECTION OF REVIEWERS

Editors need to be able to select reviewers for a particular article based on a match with their profile. Some degree of automated support could be provided for this.

9.4.3 SUBMISSION OF DRAFTS TO REVIEWERS

Drafts need to be submitted to reviewers either by sending them, or by notification of the availability of a draft at a specified location on the internet. Some mechanism needs to be in place to allow the reviewer to accept or decline a request to review and article.

9.4.4 STATUS TRACKING OF REVIEWING AND NOTIFICATION OF EVENTS

Editors need to be notified of events occurring in the submission/review/publication process, including: receipt of a submitted paper, receipt of a review,

9.4.5 STATISTICS ABOUT REVIEW PROCESS

Editors need to see statistical information regarding the process of review and acceptance. This includes information about the responsiveness and reliability of particular reviewers as well as general information about average times to publish, etc.

9.4.6 SUBMISSION/STORAGE OF REVIEWS

Reviews and information about them need to be recorded in a manner that ensures auditability of the review process. Confidence needs to be maintained regarding the confidentiality of stored reviews.

9.4.7 RETURNING ARTICLES FOR CORRECTION/EDITING

Editors need to be able to return articles, with comments for correction and modification prior to publication, and to receive revise articles back for publication.

9.4.8 RELEASING ARTICLES FOR PUBLIC VIEWING

Once an article has been accepted for publication, it needs to be released for public viewing.

9.4.9 SETTING UP A JOURNAL

A facility needs to be provided for setting up a new journal on a server and ensuring that all necessary elements are completed.

9.5 REVIEWERS

9.5.1 UPDATING OF PROFILE

Reviewers need to gain access to their profile and the ability to update it to reflect changes in their expertise, contact details, etc.

9.5.2 NOTIFICATION

Reviewers need to receive requests to review articles and a mechanism to respond to those requests (accept or decline). They also may need to be sent reminders regarding reviews that they have agreed to perform.

9.5.3 ACCESS TO ARTICLES FOR REVIEW

Articles could be sent to reviewers, or reviewers could be told where to go to gain access to articles for review.

9.5.4 CREATION OF REVIEWS

Reviewers need facilities to create reviews and may need facilities for referencing sections of articles or copying and pasting sections of articles in a review.

9.5.5 SUBMISSION OF REVIEWS

Reviewers need to be able to submit completed reviews to the editor. Confidence needs to be maintained regarding the confidentiality of reviews.

9.6 LIBRARIANS

9.6.1 REPOSITORY META-DATA

Librarians need to be able to access meta-data for cataloguing, statistical analysis, bibliometric work, etc. This includes compatibility with OAI harvesting protocols.

9.7 SERVER ADMINISTRATOR

Server administrators need to have the usual facilities of a database administrator with regard to such things as problem notification and identification, and Backup/Disaster recovery.

9.8 GENERAL REQUIREMENTS

Many of these requirements are a logical consequence of requirements that are listed above in relation to specific actor types.

9.8.1 CHARACTER SETS

At least ISO-8895-1 need to be supported in the initial system.

9.8.2 CHECKING DUPLICATION

Some checking needs to be performed by the repository to ensure that articles are not duplicated. This will be fairly simple in the initial system, but the API will take account of the possibility of more sophisticated, heuristics-based approaches later.

9.8.3 UNIQUE REFERENCES.

Each publication needs to have a unique reference identifier, which must be defined in such a manner that it may be allocated uniquely over a distributed system of repositories in the future.

9.8.4 CHECKING ACCEPTABILITY OF SUBMISSIONS.

9.8.5 USER MANAGEMENT

Facilities need to be provided to allow for user registration and management, including: registration; data recording (e.g. email address); and authentication, access control

9.8.6 DISTRIBUTED/FEDERATED/NETWORKED REPOSITORIES

Initially, the system should be based on a single, central repository, but will designed to facilitate the use of a distributed repository in the future.

9.8.7 ACCESSIBILITY FOR INTERNET SEARCH ENGINES

Suitable provision needs to be made for indexing by internet search engines, in a manner that does not interfere with the normal working of the repository.

9.8.8 COUNTING/LOGGING/MONITORING OF USAGE

Statistical data needs to be made available to support monitoring of the usage of the system.

9.8.9 BATCH SUBMISSION.

A mechanism needs to be provided for the bulk submission of large numbers of articles and/or metadata.

9.8.10 SUBMISSION OF PRINTED MATERIAL (DIGITISATION, SCANNING, ETC.)?

9.9 OTHER ACTORS

Other potential actors have been identified, but their requirements are not addressed in this iteration of the system development

- Evaluators
- Professional Associations
- Commercial Publishers
- Policy-makers/analysts (research into research)
- General Public
- Publishers

10 CONCLUDING REMARKS

It has been generally accepted that the implemented solution should conform to the Open Archives Initiative and specifically to the Dublin Core.

As part of the published API there should be a mechanism to enable data to be extracted in ENDNOTE format. This will enable many institutions around the world to have access to the information held within SciX.

The final decisions as to design and implementation will be contained in Deliverable 9 and the reader is referred to that document.

Appendix 1 - References

[RFC1777] "Lightweight Directory Access Protocol", RFC 1777, ftp://ftp.isi.edu/in-notes/rfc1777.txt

[RFC1778] "The String Representation of Standard Attribute Syntaxes", RFC 1778, ftp://ftp.isi.edu/in-notes/rfc1778.txt

Aouad, G., P.S. Brandon, T.M. Child, G.S. Cooper, S. Ford, Kirkham, J.A., Sarshar, M.(1995). ICON Final Report, University of Salford. http://www.salford.ac.uk/docs/depts/survey/staff/GAouad/pubs.html

Bernstein, P. (1996). Middleware: A Model for Distributed Services. Communications of the ACM. 39 (2) (February 1996) 86-98.

Bernstein, P., Hadzilacos, V. & Goodman, N. (1997) Concurrency Control and Recovery in Database Systems. Addison Wesley.

Birngruber , D., Kurschl, W. & Sametinger, J. (1999). Comparison of JavaBeans and ActiveX –A Case Study, STJA 99, Smalltalk und Java in Industrie und Ausbildung, Erfurt, Germany, September 28-30, 1999.

Björk, B-C., 1994. RATAS Project - Developing an Infrastructure for Computer-Integrated Construction, Journal of Computing in Civil Engineering, Vol. 8, No. 4, 400-419. http://www.vtt.fi/cic/ratas/index.html

Bohms, M., F. Tolman, & G. Storer, 1994. ATLAS, a STEP Towards Computer Integrated Large Scale Engineering, Revue internationale de CFAO, 9 (3): 325-337. http://www-uk.research.ec.org/esp-syn/text/7280.html

Bray, M. (1997). Middleware. 25 June 1997. Cited 18 January 2000. Available on the World Wide Web at http://www.sei.cmu.edu/activities/str/descriptions/middleware.html.

Bullet 2001 Diamond Bullet, An introduction to Groupware available at http://www.usabilityfirst.com/groupware/index.txl.

CBR 1997 Overview by Ian Harrison. available at http://www.aiai.ed.ac.uk/links/cbr.html.

Cocoon 2001 Cocoon Project, http://xml.apache.org/cocoon/.

Cooper, G.S. & Rezgui, Y.R. (2000). Objects and Integration in the "Wired Society". To appear in: Proceedings of the UK National Conference on Objects and Integration for Architecture, Engineering, and Construction, London, 13-14 March 2000.

Curtin, M. & Ranum, M. (1999) The Firewalls FAQ. 25 November 1999. Cited 9 February 2000. Available on the World Wide Web at http://www.faqs.org/faqs/firewalls-faq/.

Drucker 1999 P. F. Drucker (1999). Knowledge-Worker productivity: The biggest challenge. California management review, 41, 79-94.

Dubois, A.M., J. Flynn, , M.H.G Verhoef & F. Augenbroe, 1995. Conceptual Modelling Approaches in the COMBINE Project, presented in the COMBINE final meeting, Dublin. http://erg.ucd.ie/combine/papers.html

Durfee and Lesser 1989 E. Durfee, and V. Lesser (1989). Negotiating Task Decomposition and Allocation Using Partial Global Planning. In DistributedArtificial Intelligence, Volume 2, eds. L. Gasser and M. Huhns, 229–244. San Francisco, CA, Morgan Kaufmann.

Eastman, C & Augenbroe, F (1998). Product Modeling Strategies for Today and The Future. Proceedings of the CIB Working Commission W78 Information Technology in Construction Conference, Stockholm.

Emmerich, W. & Ellmer, E. (1998). A Survey of Object-Orientated Middleware. Accepted for the (cancelled) IDPT98 Conference, Turkey but obtained direct from the author.

Emmerich, W. (2000). Software Engineering and Middleware: A Roadmap. To appear in: A. Finkelstein (ed): Future of Software Engineering - State of the Art Reports given at the 22nd Int. Conf. on Software Engineering, Limerick, June 2000. ACM Press. 2000. Also available on the Internet at http://www.cs.ucl.ac.uk/staff/W.Emmerich/publications/ICSE2000/SOTAR/index.html.

Erzberger, M. & Altherr, M. (1999). Every DAD Needs A MOM: Message Oriented Middleware. Cited 24 February 2000. Available on the World Wide Web at http://www.softwired.com/pubs/momdad_en.pdf.

F.A.Q on Document and Content Management by Document Management Avenue available at http://www.documentmanagement.org.uk/pages/faq.htm.

FAQ Lutz Prechelt, http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/ai-repository/ai/html/faqs/ai/neural/faq.html.

Frankel, D.S. (1999) CORBA Components –alive and wll. Java Report. October 1999. 70-77.

Franklin & Graesser 1997 S. Franklin, A. Graesser (1997) "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents," Intelligent Agents III, Berlin: Springer Verlag, pp. 21-35.

Frawley et al 1992 W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An overview. AI Magazine, Fall 1992, pgs 213-228.

Fröhlich 1997 Jochen Fröhlich, An overview of Neural Networks available at http://rfhs8012.fh-regensburg.de/~saj39122/jfroehl/diplom/e-1.html.

Gallaugher, J. & Ramanathan, S. (1996). The Critical Choice of Client Server Architecture: A Comparison of Two and Three Tier Systems. Information Systems Management. 13 (2) 7-13.

Hunter et al. 2001 J. Hunter and W.Crawford (2001). JAVA Servlet Programming, Second Edition, ISBN 0-596-00040-5.

ICE 2000 Overview of ICE by IDEAlliance at http://www.icestandard.org/. (1998)Description of ICE bon W3C web site at http://www.w3.org/TR/NOTE-ice.

ISO 1988. ISO 9735:1988 Electronic data interchange for administration, commerce and transport (EDIFACT). International Standards Organization. TC 154 (See: http://www.iso.ch/cate/d17592.html)

ISO 1994. ISO 10303-1:1994 Industrial automation systems and integration -- Product data representation and exchange -- Part 1: Overview and fundamental principles. International Standards Organization. TC 184/SC 4 (See: http://www.iso.ch/cate/d20579.html)

JavaSoft (2000). The JavaBeans FAQ: General. Author Unknown. Date Unknown. Cited 7 April 2000. Available on the World Wide Web at http://www.javasoft.com/beans/faq/faq.general.html#Q15.

KM Forum 2001 J. Kemp, M. Pudlatz; P. Perez; A. Ortega, Deliverable D1.1 - KM Technologies and Tools, available at http://www.knowledgeboard.com.

KM http://searchwebmanagement.com ; http://searchebusiness.com ; http://searchebusiness.techtarget.com/sDefinition/0,,sid19_gci498329,00.html ; http://www.knowledgeboard.com/.

Margherio et al. 1998 L. Margherio, D. Henry, S. Cooke, S. Montes, The Emerging Digital Economy, Report from the U.S. Department of Commerce, April, 1998. Available at http://www.ecommerce.gov/emerging.htm.

Marir, F and Watson, I A, 1995. CBRefurb: case-based cost estimation, Colloquium on case-based reasoning: Prospects for application organised by Professional Group C4 (Artificial Intelligence), 7, March.

Moore et al. 2001T. Moore, C. Jesse, R. Kittler, An overview and Evaluation of Decision Tree Methodology, ASA Quality and Productivity Conference, Amy 23-25, 2001. Available at http://www.ydyn.com/pubs/2001/asa.pdf.

Morris, E. & Litvak, E. (1997). Component Object Model (COM), DCOM and Related Capabilities. 23 June 1997. Cited 17 January 2000. Available on the World Wide Web at http://www.sei.cmu.edu/str/descriptions/com.html.

Nokia WAP Development Zone: http://www.forum.nokia.com/main/1,,1_1,00.html

OMG (1999). The Complete formal/99-10-07 CORBA/IIOP 2.3.1 Specification. Available on the World Wide Web at http://www.omg.org/corba/corbaiiop.html.

ONTOLOGY http://www.ontoknowledge.org; http://www.swi.psy.uva.nl/projects/ibrow/home.html; http://www.si.fr.atosorigin.com/sophia/comma/Htm/HomePage.htm; http://www.daml.org/.

Orfali, R, et al. (1996). The Essential Client/Server Survival Guide. Second Edition. New York: Wiley.

Orfali, R, et al. (1999). The Essential Client/Server Survival Guide. Third Edition. New York: Wiley.

Orfali, R., Harkey, D., and Edwards, J. (1997). Instant Corba. ISBN 0-471-18333-4 . John Wiley & Sons.

Raj (1998). DCOM, CORBA, Java-RMI - A Step by Step Comparison. 28 September 1998. (Cited 3 February 2000). Available on the World Wide Web at http://www.execpc.com/~gopalan/misc/compare.html.

Sadoski, D, (1997). Client/Server Software Architectures -- An Overview. 2 August 1997. (Cited 17 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/str/descriptions/clientserver_body.html.

Sadoski, D, (1997b). Two Tier Software Architectures. 10 January 1997. (Cited 17 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/str/descriptions/twotier.html.

Sadoski, D, (1997c). Three Tier Software Architectures. 10 January 1997. (Cited 17 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/str/descriptions/threetier.html.

SAP, (2000). SAP Tools & Components. http://saplabs.com/usa/devarea/auto.htm
Schussel, G. (1996). Client/Server Past, Present and Future. Date unknown. (Cited 17 January 2000). Available on the World Wide Web at http://news.dci.com/geos/dbsejava.htm.
SciX (1999). Proposal Part B; Open System for Inter-enterprise Information Management in Dynamic Virtual Environments, IST-1999-10491.
SOAP 1999 Specification of SOAP on W3C web site available at http://www.w3.org/2000/xp/.
SOAP on MSDN: http://www.msdn.microsoft.com/xml/general/soaptemplate.asp
SOAP Specification: http://msdn.microsoft.com/workshop/c-frame.htm#/xml/index.asp
SOAP:
Socolofsky, T, and Kale, C, (1991) A TCP/IP Tutorial RFC 1180. Available on the Internet at ftp://ftp.isi.edu/in-notes/rfc1180.txt.
Sun (1997). RMI and IIOP in Java. Author unknown. 26 June 1997. (Cited 3 February 2000). Available on the World Wide Web at http://java.sun.com/pr/1997/june/statement970626-01.faq.html.
Sun (1999). What is the Java™ Platform? Author unknown. 19 October 1999. (Cited 3 February 2000). Available on the World Wide Web at http://java.sun.com/nav/whatis/.
Sun Microsystems, (2000). http://java.sun.com/
TechWeb (1999). Intranet. Author unknown. Date unknown. (Cited 26 January 2000). Available on the World Wide Web at http://www.techweb.com/encyclopedia/defineterm?term=intranet.
UDDI http://www.uddi.org/.
Vanhelsuwé (1997). Mastering JavaBeans. Alameda: Sybex. Also available on the World Wide Web at http://www.lv.clara.co.uk/masbeans.html.
Vinoski, S. (1998). New Features for CORBA 3.0. Communications of the ACM. 41 (10) 44-52.
Vivek et al. 1997 R.Vivek, R.Gupta, Senior Consultant, System Services corporation,Chicago, Illinois. An Introduction to Datawarehouse. available at http://www.system-services.com/dwintro.asp.
Vondrak, C. (1997). Message-Orientated Middleware. 10 January 1997. (Cited 26 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/activities/str/descriptions/momt.html.
Vondrak, C. (1997b). Remote Procedure Call. 10 January 1997. (Cited 26 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/activities/str/descriptions/rpc_body.html.
W3C 1999 REC-xml-names-19990114, Document available at http://www.w3.org/TR/REC-xml-names/.
Wallnau, K. (1997). Common Object Request Broker Architecture. 10 January 1997. (Cited 28 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/str/descriptions/corba.html.
Wallnau, K. and Foreman, J. (1997). Object Request Broker. 25 June 1997. (Cited 28 January 2000). Available on the World Wide Web at http://www.sei.cmu.edu/str/descriptions/orb.html.
WAP Forum, “Wireless Application Protocol: White Paper”, Wireless Internet Today, June 2000
WAP Forum: http://www.wapforum.com
WAP Specifications: http://www.wapforum.com/what/technical.htm
WAP Tutorials: http://www.waplinks.com/
WAP:
Watson, I and Marir, F, 1994. Case-based reasoning: A review. The Knowledge Engineering Review. 9. (4).
Webopedia (1998). Local Area Network. Author unknown. 16 May 1998 (Cited 26 January 2000). Available on the World Wide Web at http://webopedia.internet.com/TERM/l/local_area_network_LAN.html.
Whatis.com (1999). What Is … the Internet (a definition). Author unknown. 25 October 1999. (Cited 25 January 2000). Available on the World Wide Web at http://www.whatis.com/internet.htm.
Wooldridge 1999 M. Wooldridge. Intelligent Agents, In G. Weiss, editor, The MIT Press, April 1999, ISBN 0-262-23203-0.
Workflow http://www.e-workflow.org/ ; http://www.wfmc.org/.
Written et al. 1999 I. Written, A. Moffat, T. Bell, Compressing and Indexing documents and images, Morgan Kaufmann Publishers, Second Edition, 1999, ISBN 1-55860-570-3.
WSDL specification: http://msdn.microsoft.com/xml/general/wsdl.asp
XML XML Schema by W3C at http://www.w3.org/XML/Schema.
Yee, A. (1999). Making Sense of The COM vs. CORBA Debate. Performance Computing. June 1999. Also available on the World Wide Web at http://www.performancecomputing.com/features/9906dev.shtml.
Zarli, A. & Richaud, O. (2000). Requirements and technology integration for IT-based Business-oriented frameworks in Building and Construction. Submission to the Electronic *Journal of Information Technology in Construction.*

Appendix 2 Table of Abbreviations

ADL	Advanced Distributed Learning
AICC	Aviation Industry CBT Committee
CMI	Computer Managed Instruction
AEC	Architecture, Engineering and Construction
API	Application Programming Interface
BP	Business Process
CBR	Case Based Reasoning
CCM	CORBA Component Model
CSS	Cascade Style Sheet
CIDL	CORBA Interface Definition Language
COM	Component Object Model
CoMMA	Corporate Memory Management through Agents. IST project IST-1999-12217
CORBA	Common Object Request Broker Architecture
DTD	Document Type Definition
DW	Data Warehouse
Knowledge Management System	Knowledge Management Infrastructure Methodology, tools and architectures
EJB	Enterprise Java Bean
FM	Facilities Management
HTML	Hyper Text Markup Language
HTTP	HyperText Transfer Protocol
IC	Information and Communication
ICE	Information Content Exchange
ICT	Information and Communication Technologies
IDL	Interface Definition Language
IEEE/LTSC	Learning Technology Standards Committee
IMS	Global Learning Consortium
ISP	Internet Service Provider
ISO	International Organisation for Standardisation
IST	Information Society Technologies
IT	Information Technologies
J2EE	Java 2 Enterprise Edition
JSP	Java Server Pages
JVM	Java Virtual Machine
KB	Knowledge Base
KI	Knowledge Item
KM	Knowledge Management
LAN	Local Area Network
LEAP	Lightweight Extensible Agent Platform
MOM	Message-Orientated Middleware
OIL	Ontology Inference Language
OPS	Open Profiling System
OS	Operating System

ORB	Object Request Broker
OMG	Object Management Group
PC	Personal Computer
PSDL	Persistent Sate Definition Language
RDF	Resource Description Framework
RDFS	RDF Schema
RMI	Remote Method Invocation
RPC	Remote Procedure call
R&D	Research and development
RMI	Java Remote Method Invocation
SciX	Open,self organising repository for scientific information exchange
SQL	Structured Query Language
SOAP	A publishing methodology for Web-based services
SGML	Standard Generalised Markup Language
SN	Semantic Networks
SME	Small and Medium size Enterprise
SMTP	Simple Mail Transfer Protocol
TCP/IP	Transport Control Protocol / Internet Protocol
UBR	UDDI Business Registry
UDDI	Universal Discovery Description and Integration
URL	Uniform Resource Locator
VE	Virtual Enterprise
WAN	Wide Area Network
WfMC	Workflow Management Coalition group
WP	Work Package
WSDL	Web Services Description Language
WWW	World Wide Web
W3C	World Wide Web Consortium
XML	eXtensible Mark-up Language
XSD	XML Schema Definition
XSL	Extensible Style sheet Language