X-Media: Large Scale Knowledge Acquisition, Sharing and Reuse across-Media

Fabio Ciravegna and Steffen Staab

Abstract— **X-Media is an integrated Project funded by the European Commission, which addresses the issue of knowledge management in complex distributed environments. It will study, develop and implement large scale methodologies and techniques for knowledge management able to support sharing and reuse of knowledge that is distributed across different media (images, documents and data) and repositories (data bases, knowledge bases, document repositories, etc.). The project started in March 2006 and will last for 4 years.**

Index Terms—**Cross Media Knowledge Acquisition, Cross-media Knowledge Sharing, Architectures for Knowledge Management**

I. Introduction

While in the past, medium size, mainly textual, centralized archives used to be the only resources for knowledge management, nowadays large companies handle very large quantities of multimedia information in distributed archives. Their intranets connect thousands of computers and reach sizes of dozens of millions of documents. In addition, the increased use of the WWW as a source of information has made the boundary between intra- and inter-net very thin. This dramatically increases the size of the information space. Moreover, databases and archives are used to store huge amounts of information that is vital for the organization life, such as data on products, financial information, etc. Collecting and aggregating multimedia knowledge is of fundamental importance in order to gain competitiveness and to reduce costs. For example thousands of documents are produced during the design and manufacturing of a class of jet engines. During service, a single engine produces about 1Gbyte of vibration data per flight; if irregularities are found, part of the data is stored. Every time an engine is serviced, financial information is produced. If problems are found, pictures are taken, reports are written. Each individual engine has a potential "folder" of information describing the whole lifecycle of the engine that can easily sum up to several Gigabytes of information, potentially Terabytes, and contains highly interrelated information stored in different media.

The growing size and the multi-media nature of the archives have serious implication on the way knowledge management can be implemented. There are a number of dimensions along which the complexity arises:

- **Cross-Media**: evidence is often distributed in different media; it is possible that knowledge expressed in just one medium does not carry enough evidence. Connecting information in more than one medium is often required.
- **Knowledge integration**: large distributed archives require the ability to map the distribution of information, to weight every single source and to distribute searches carefully; this is very difficult and often search is performed just in some of the archives, disregarding others that can bring very useful information;
- **Focusing**: large amount of information implies that managing knowledge becomes more complex and needs powerful focusing methodologies. Focus of searching changes in time and from user to user, and requires a balanced mixture of exploration and searching;
- **Uncertainty and Dynamicity**: information is often ambiguous, incomplete, or referring to a specific context - therefore archives can contain noise and imprecision, as well as obsolete information; each piece of knowledge must therefore be judged based on provenance, evidence, etc.
- **Infrastructure**: different media cannot easily be shared. A folder of text documents may be sent via email, but a folder of images may not, and may instead require a shared image repository. For 10 GByte of data remote access to the underlying data base is to be considered.

Current knowledge management technologies and practises cannot cope with such new situation, as they mainly provide simple mechanisms (e.g. keyword searching) for supporting knowledge workers manually *pierce* together the information from different sources.

II. X-Media

X-Media addresses the issue of knowledge management in complex distributed environments. It studies, develops and implements large scale methodologies and techniques for knowledge management able to support sharing and reuse of knowledge that is distributed in different media (images, documents and data) and repositories (data bases, knowledge bases, document repositories, etc.).

X-Media studies, designs and develops:

1) Robust and scalable knowledge acquisition and data analysis tools operating across media boundaries (text, images and data) to automatically cross-relate and annotate text and images with metadata.

F. Ciravegna is with the Department of Computer Science of the University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP, Sheffield, UK. (e-mail: f.ciravegna@dcs.shef.ac.uk)

S. Staab is with the Department of Computer Science, University Koblenz-Landau, 56016 Koblenz, Germany. (e-mail: staab@uni-koblenz.de)

2) Novel and cutting-edge knowledge fusion methods to support knowledge workers in making decisions when confronted with – possibly contradicting – knowledge derived from different resources;
3) Effective and efficient new paradigms for knowledge retrieval, sharing and reuse working across media which enable users to define and parameterize views on the available knowledge according to their needs.
4) Techniques able to represent and manage (i) uncertainty, (ii) trust and provenance as well as (iii) dynamic aspects of knowledge;
5) A methodology and a technical infrastructure able to deliver knowledge from across media to the knowledge workers, taking into account the complexity of managing different media with different size of data.
6) A generic and flexible architecture allowing end users to easily customize it and integrate it with their KM practices or needs as well as a mainly open source reference implementation and libraries which technology providing companies can reuse.

Technologies will be able to support knowledge workers in an effective way, (i) hiding the complexity of the underlying search/retrieval process, (ii) resulting in a natural access to knowledge, (iii) allowing interoperability between heterogeneous information resources and (iv) including heterogeneity of data type (data, image, texts). The expected impact on organizations is to dramatically improve access to, sharing of and use of information by humans and between machines. Expected benefits are a dramatic reduction of management costs and increasing feasibility of complex knowledge management tasks. The project plan is structured along the four areas described below.

Area 1: knowledge sharing and reuse

X-Media studies and implements technologies and methodologies for easy and intelligent access to and reuse of formalized and non formalized knowledge. The reuse takes into consideration the user context to help focus searches and reuse. Reuse and sharing is enabled via cross-media ontology supported automatic indexing. The technology works in a largely automated way, but it is centered on supporting users' work, rather than replacing them. This is because the activity of a knowledge worker is complex and humans are irreplaceable agents in this process.

In this context, we are studying, designing and developing:
(1) Effective and efficient new paradigms for knowledge retrieval, sharing and reuse which enable users to define and parameterize views on the available knowledge according to their needs.
(2) Novel and cutting-edge knowledge fusion methods to support knowledge workers in making decisions when confronted with – possibly contradicting – knowledge derived from different resources.
(3) techniques able to represent and manage (i) uncertainty, (ii) trust and provenance as well as (iii) dynamic aspects of knowledge.

Area 2: automated knowledge acquisition from documents, images and raw data

Functional to the methodologies for knowledge sharing investigated in Area 1, is the ability to acquire knowledge across media in a rich, semantically-oriented way. X-Media develops a set of tools able to support sharing methodologies in a seamless and automatic way. Media addressed are raw data, texts and images (e.g. results or parameters in experiments, raw images, textual documents, etc.). The outcome of the acquisition technologies will be a semantic representation of the content (conceptualization) to be used for knowledge management purposes. Enrichment of multimedia documents with additional layers of automatically generated annotation is the main medium of associating conceptualizations to resources. Current technology focuses on single medium technologies to acquire knowledge in multi media environments; this means that retrieval methods use mainly one medium (e.g. text) even in multimedia environments. X-Media designs and develops technologies for information extraction that work truly cross media and that can be used in cases where information in one medium is necessary to understand the information in the other.

Area 3: Infrastructure

A knowledge acquisition, integration and sharing environment is defined. Since X-Media is an application-oriented integrated project, integration is required on the implementation as well as on the conceptual level. The main outcome of this area of activity will be a methodology and a technical infrastructure able to deliver knowledge from across media to the knowledge workers, taking into account the complexity of managing media with different size of data.

Area 4: Application and Testing

The technology above is used to define showcases and prototype applications. Two main testbeds are defined by the two industrial users (Rolls Royce and Fiat). They concern competitor analysis in the car industry and product lifecycle monitoring in aerospace. System trials with final users will showcase the technology and pave the way to further exploitation.

III. Consortium

Partners: University of Sheffield (coordinator, UK), University of Koblenz (D), ITC-Irst (I), University of Ljubljana (Slovenia) University of Freiburg (D), CERTH (G), Labri (F), University of Karlsruhe (D) and the Open University (UK). Quinary (I), Ontoprise (D), Solcara (UK), CognIT (N), Rolls Royce (UK) and Centro Ricerche Fiat (I).

Acknowledgment

X-Media is funded by the European Commission as part of Framework 6 of IST, contract no FP6-26978. Project web page: http://www.x-media-project.org.

For information: Prof. Fabio Ciravegna, email: xmedia-coordinator@dcs.shef.ac.uk