State-of-the-Art for Entity-Centric Repository and Authoring Environment for Multimedia

Atta Badii, Oleksandr Kolomiyets, Chattun Lallah, Ali Khan, Meng Zhu, Ben Gardiner-Jones
IMSS, Intelligent Media Systems and Services Research Laboratory
University of Reading
UNITED KINGDOM
{atta.badii, o.kolomiyets, c.lallah, a.a.khan, meng.zhu, b.gardinerjones}@reading.ac.uk
http://www.imss.reading.ac.uk

Abstract: - Many projects, e.g. VIKEF [13] and KIM [7], present grounded approaches for the use of entities as a means of indexing and retrieval of multimedia resources from heterogeneous sources. In this paper, we discuss the state-of-the-art of entity-centric approaches for multimedia indexing and retrieval. A summary of projects employing entity-centric repositories are portrayed. This paper also looks at the current state-of-the-art authoring environment, Macromedia Authorware, and the possibility of potential extension of this environment for entity-based multimedia authoring.

Key-Words: - Semantic Annotation, Indexing and Retrieval, Entity-Centric Repository, Multimedia Authoring

1 Introduction

The enormous scale of the WWW sets a new challenge – handling this scale, so that the users do not get lost and can efficiently perform their intended tasks. However, the individual pieces of information do not co-exist with each other in order to infer knowledge to the information consumer, which may be computers or humans. This is because the information available does not contain enough metadata, which provides semantic information about them. To achieve the vision of the Semantic Web, we need a critical mass of metadata for web content, linked to formal knowledge about the world. This is the reason for fostering the creation of an entity-centric repository which indexed documents with respect to real-world entities.

In this paper, we present the global aspects revolving around the use of such kind of repository. We first highlight the technical aspects of multimedia information and its technical schematic representation in an Entity-Relation database. The paper then discusses two frameworks, KIM [7] and VIKEF [13], built upon the discussed technologies, and some other related projects and future work in this domain. An overview of a commercially available authoring environment, Macromedia Authorware, and its potential extension for development of an entity-centric user-intuitive multimedia authoring environment concludes this paper.

2 Multimedia Information Object Structure

In this section, we illustrate how a multimedia object is decomposed into different abstract layers. These layers provide different types of metadata, each having its own use in the data indexing and retrieval sphere.

2.1 Multimedia Objects

Considering a multimedia document from the point of view of its content the following properties can be identified: raw data, format, attributes, features and concepts. Based on these characteristics a multimedia information object can be decomposed into a series of layers [11] (Fig. 1):

- Data layer – a conjunction of raw data, format and attributes
- Feature layer – a domain-independent and format-dependent characterisation of the raw data
- Concept layer – a subjective domain-dependent and format-independent semantic interpretation of the data and features of an information object that has a degree of certainty

In the conceptual model of the content based representation of multimedia, the following relationships as presented on Fig. 2 can be identified.

2.2 MPEG7 Metadata

The major interest of the research community lies in bridging the gap between the feature and concept

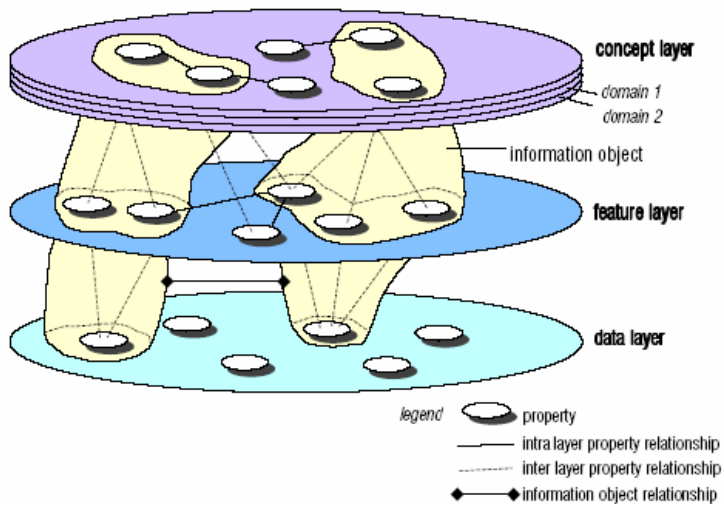

Fig.1: Three-layer multimedia representation model [11]

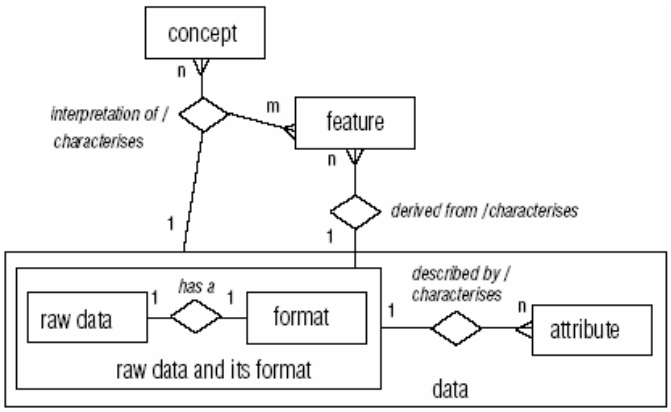

Fig.2: Entity-Relationship diagram for content-based information object annotation [11]

layer. Features represent meaningful pieces of information common to the type of the media, which can be easily extracted from raw data by feature extraction methods and encode into a feature-vector, which represents the original document in a compact numerical manner. Concepts, in turn, are semantic annotators of the document and provide the meaning of the multimedia content. Usually such information is not available at hand and can be derived by sophisticated methods of data mining making use of features or semi-automatically, with human intervention.

In order to standardise the semantic description of multimedia data, MPEG7 has been introduced as a comprehensive metadata standard for multimedia. MPEG-7, formally named "Multimedia Content Description Interface", is a standard for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code [9].

MPEG-7 offers a broad set of audiovisual Description Tools to create descriptions, which will form the basis for applications enabling the needed effective and efficient access (search, filtering and browsing) to multimedia content.

2.3 ER Model for Video File

Database and information retrieval communities employ entity-relation (ER) or object-oriented approaches to model the semantics of multimedia documents. These approaches can be classified as those which incorporate semantics into a primarily low-level feature-based framework and those which incorporate low-level features into a database-modelling approach.

An example of an ER model for describing a video event, as one of the major types of multimedia, is

presented (Fig. 3). This model is an extension of the ER model with object-oriented concepts [3]. The main entities in the model are events, objects that participate in these events, and "actors" entities that describe object roles in the events.

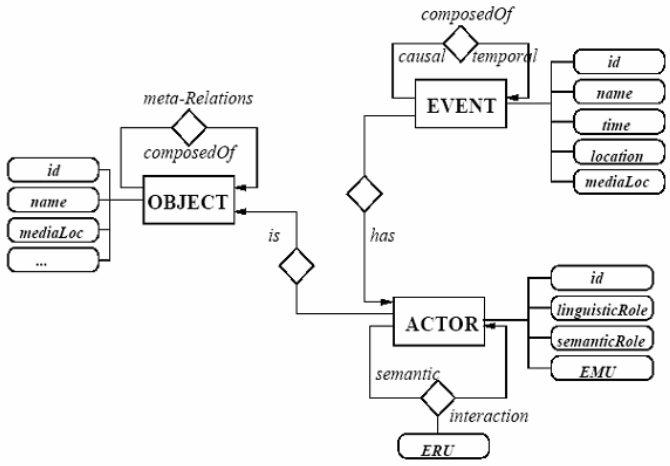

Fig. 3: Exemplar ER Model for representing a video event [3]

2.4 Named Entities (NE)

Named Entity Recognition is a process of identifying atomic elements in text (Name Entities, NE), such as names of persons, organisations, locations, expressions of times quantities, monetary values percentages etc. Many research groups use NE to link these to real entities in the world and create small ontologies based on the major classes of NE. A similar approach is also known for video, but an automatic approach for entity recognition in video is, as yet, far from fully accomplished. Current approaches make use of multi-modal analysis, such as textual, visual and auditory information to improve the performance of unimodal approaches [2].

Examples of semantic types of named entities are:

- Dates: June 15, 2007
- Monetary expressions: $26 billion
- Locations: London, Stonehenge, etc.
- Person Names: President George W. Bush, Kofi Annan, Edward III of England
- Organisations: British Airways, NATO, etc.
- Events: World War II, Wimbledon 2007
- URLs and Emails

3 The KIM Platform

KIM [7] aims to provide a Knowledge & Information Management Platform for semantic annotation, indexing and retrieval and focuses on providing the necessary infrastructure and scalability for extraction of Name Entities (NE) and some relations including their identification and indexing with respect to the entity references and key-phrases.

The semantic annotation offered in KIM is a specific metadata generation and usage schema targeted to enable new information access methods and extend existing ones. The annotation scheme based on the understanding of the Name Entities, mentioned in the documents constitutes an important part of their semantics. Using different sorts of redundancy and external knowledge, the entities can be used along with formal descriptions and provide more semantics and connectivity for the web. This aims at drastically improving information retrieval on the web.

The KIM Server automatically extracts semantic annotations from unstructured information and associates these with the processed document. Each annotation keeps URI references for the extracted entity. The annotated documents with the entity occurrence information are stored in a RDBMS, while the semantic description of entities is stored in a semantic repository. The usage of two heterogeneous systems inherently causes duplication of some pieces of information. To keep the two pieces of information about the same entity in synchrony, a correspondence between the entity description in the semantic repository and the database is kept. Three major types of information are stored in the database: entity information, document information and the relation between them (information for the occurrence of entities in the documents).

The entity information is the mapping between its identity in the database and semantic repository. Since the database indices are optimised to work with integers while the semantic repository is based on URIs, a unique integer ID for each entity URI is generated. This is the most important part of the entity information since it links its representation in the database to the full entity description in the semantic repository. In some kinds of queries, users are interested only in entities of a given class or with a given alias. The entity class and aliases are stored in the semantic repository but for performance reasons, KIM also keeps them in the database. This is not essential but is the only way to make the queries efficient without the need to combine query results from different sources [10].

However, the evaluation work done until now in KIM does not provide enough evidence regarding this approach. The major obstacle is that there are neither test data nor well-developed metrics for semantic annotation and retrieval. Although naïve in

some aspects, the KIM platform provides a test bed and proves the following hypothesis and design decisions:

- It is worth using massive Entity Knowledge, and even without comprehensive named entity disambiguation, the retrieval precision still seems acceptable.
- It is possible to store and query tens of thousands of entities together with their descriptions.
- Few light-weight front-end tools can deliver the results of semantic annotation, indexing and retrieval in an intuitive way.

One of the future challenges in KIM is to experiment different approaches towards disambiguation of named-entity references: adaptation of a Hidden Markov Model Learner successfully used for non-semantic disambiguation is one of the first ideas; technically similar to those used for word-sense disambiguation (namely, lexical-chaining); techniques for "symbolic" context management.

4 The VIKEF Framework

VIKEF (Virtual Information and Knowledge Environment Framework) is an application oriented Integrated Project within the IST Sixth Framework Programme of the European Community, which started in April 2004 [6]. Its main aim is to bridge the gap between the (partly) implicit knowledge and information conveyed in scientific and business Information, Content and Knowledge (ICK) resources and the explicit representation of knowledge required for a targeted and effective access, dissemination, sharing, and use of such ICK resources by scientific and business communities [13].

The RTD effort within VIKEF has built on and significantly extended the current Semantic Web efforts by addressing crucial operational and application challenges in developing real-world semantically enriched virtual information and knowledge environments.

In the VIKEF project, an entity-centric repository is implemented. Entities are collected from multiple sources, assigned a unique ID (a URI) and stored together with any available descriptions [14]. This is a component of the VIKEF knowledge infrastructure, but also a contribution to the conceptual Semantic Web Architecture. The Systematic management of entities (individuals) is an important step towards integration and interoperability in the Semantic Web, where retrieval is expected to be done not on a keyword-based basis but also by entity-based one.

5 Other Related Projects

5.1 EMM NewsExplorer

A relevant work has been performed at the Joint Research Centre (JRC) of the European Commission in Italy by the Ralf Steinberger and Clive Best group. One of their products is called the EMM NewsExplorer [4], which analyses news articles in several languages. This tool extracts key-phrases and name entities (NE) from articles, and calculates clusters of news articles, where each cluster represents a single event covered in the news. This analysis is cross-lingual and each cluster is formed of articles in different languages referring to the same event. The identification of a single event out of many news articles that speak about the same thing brings the usability of the NewsExplorer to an entirely new level. Using the events, a user can read only the most relevant article and get all the information instead of going through tens or hundreds of news.

EMM NewsExplorer also provides timelines of the major news on the basis of the sizes of the event clusters. Another interesting side is the front-end provided to the users to help them navigate through the news. For each article there are people and organisations that are either related to or associated with the content. Similarly for each entity one can see the articles that refer to it. Generally their ideas, efforts and results are remarkable in several aspects: multi-lingual, IE, clustering of news to find out events, and presentation of the news to the end-users.

5.2 BRICKS

Building Resources for Integrated Cultural Knowledge Services, BRICKS [1], aims at integrating the existing digital resources into a common and shared Digital Library, a comprehensive term covering "Digital Museums", "Digital Archives" and other kinds of digital memory systems. Two highlights of the project should be noted:

JCR-empowered Multimedia Content Repository: The JCR API, the first content repository API to be standardised by the Java Community Process (JSR-170), provides an excellent example of content repository specification, along with a simple yet powerful API. JCR API is based on a very simple hierarchical model, which can be used to define different information models and manage content instances. And BRICKS covers the common

features of content repository i.e. object management, dynamic extensibility, relationship management, notification, version management and configuration management.

Metadata Modelling: BRICKS follows the Open Archives Initiative (OAI) methodology for metadata management and organises descriptive metadata as metadata records. A metadata record always describes a content item that is unambiguously identified by a URI. A content item can have one or more metadata records, each corresponding to exactly one metadata schema. Within the Metadata Manager, metadata records are grouped into repositories.

5.3 The OpenDLib Project

OpenDLib [5] is a software toolkit that can be used to create a digital library easily, according to the requirements of a given user community, by instantiating the software appropriately and then either loading or harvesting the content to be managed. A major advantage, when adopting the OpenDLib Digital Library Management System is the possibility to manage a great number of different document structures and types, and metadata formats in the same Digital Library (DL) using the same tool. This goal is achieved with the introduction of the OpenDLib Document Model. OpenDLib does not manage document files directly; it manages "entities" (named digital objects) that conform to the Document Model for Digital Library (DoMDL). This model can represent a large variety of document structures.

6 Entity-Centric Multimedia Authoring in User Intuitive Environments

This section discusses an innovative environment that can make use of information in an entity-based repository for creating and editing new content. The environment allows querying the entity repository, authoring, manipulation, packaging, combination and presentation of retrieved content and metadata enrichment to create rich content objects. Such an environment can include features that allow preparation of task specific content objects and effective exploration of available information space from various dimensions using task specific templates, which affect the ranking and presentation of information. User Profiling and Behavioural User Analysis information can be used by such an environment for introduction of personalisation into the authoring tool by way of building a local and temporal knowledge base of advanced user information and context information. This information is used to offer customised variations of templates and contextual suggestions of existing entities from the repository that may or may not be selected by the user during the authoring process of the new content.

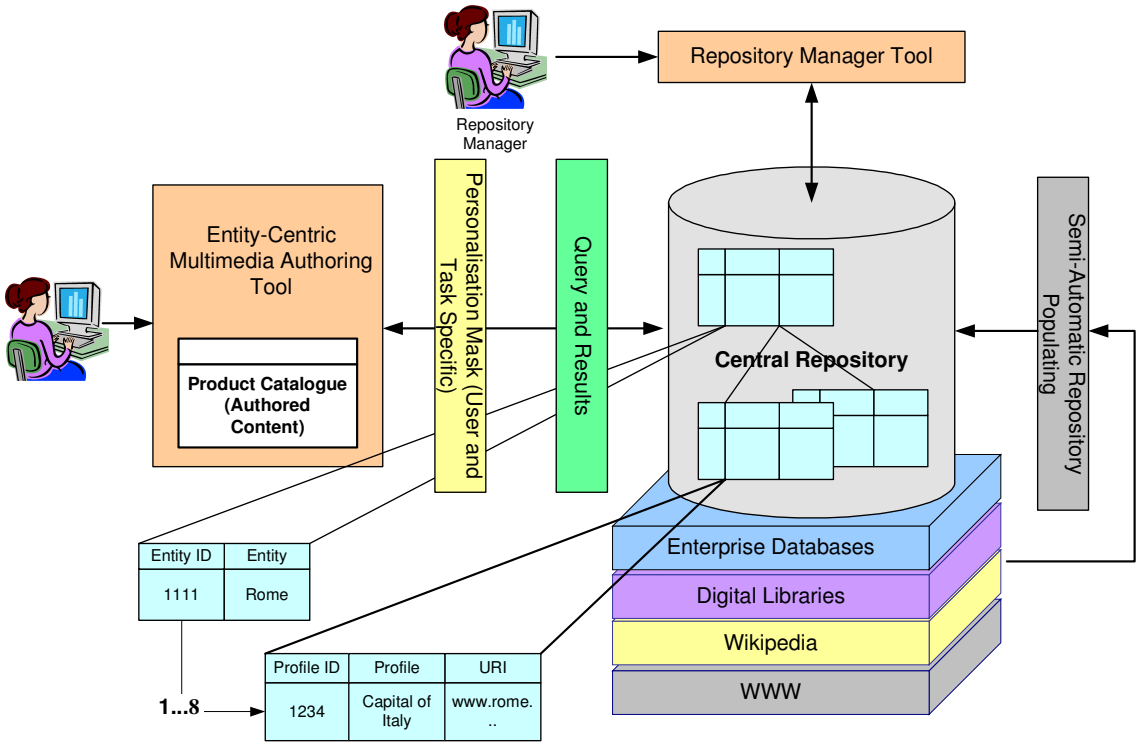

Fig. 4: Entity Centric Repository and Authoring Environment Architecture

6.1 Entity Centric Repository and Authoring Environment Architecture

Fig. 4 shows the architecture of an Entity-Centric empowered Multimedia Authoring Tool. In the illustrated context, the user's goal is to create new subject-specific document. A number of information sources are available for the user at hand. However, such huge collection of information requires efficient tools for information access and retrieval. These foster new solutions for indexing, information connection, semantic categorisation, etc.

A unified entity-centric platform will merge multiple information channels, like databases, libraries, and the Web into a central information resource via extracting vital entities from data. The Repository Population is done in a semi-automatic manner employing the state of the art data mining approaches. A Repository Manager will be responsible for the maintenance of the entity repository, e.g. continual URI verification, alteration of trust levels, deletion of un-trusted sources, etc.

Knowing his/her topic of interest, the user can search for relevant information by entities in the entity-centric repository. He/she can also browse through entity profiles, which presents related information resources and links, for any selected entities, e.g. the entity "Rome" will have profiles "Capital of Italy", "population of 2.7 million", etc.

This approach makes authoring new documents more easy by identifying existing relevant content through the entity centric repository. For example, a researcher who needs to write a new scientific paper, which refers to other existing papers, will discover the situational landscape of the available materials like PowerPoint presentations, Video and Audio, referenced papers, etc., indexed in the central repository.

6.2 Standard Authoring Tool/Environment

The User-intuitive interface for an Entity-centric Authoring environment/tool would make use of an extended version of a Standard Authoring Tool to produce new content by manipulating, and packaging entities retrieved from the entity repository with complementary information. An example of possible candidate to be used and extended as Entity-centric Authoring Environment can be Macromedia Authorware [8]. The role of the Entity-Centric Authoring Tool consists of:

- Acquisition of resources from a repository
- Manipulation, formatting, editing, presentation and authoring of retrieved content
- Metadata enrichment
- Packaging of content

6.3 Potential/Scope for Extension

As depicted in Fig. 5, a Standard Authoring Tool such as Macromedia Authorware allows packaging of content which can be viewed or played by a user in the Standard Authoring Tool Player. An extended version of such an Authoring Tool with support for entity-centric multimedia authoring would allow the inclusion and packaging of entities within new content (hence these objects would comprise of one or more entities along with complementary information). The extensions required to the standard authoring tool for the achievement of objectives mentioned under section 6.2 would be:

- Introduction of entity support using scripting features of Authorware
- Acquisition of multimedia entities from repository
- Metadata enrichment (possible use of MPEG-7) – semantic infusion of multimedia content with entity
- Creation of new content involving one or more multimedia entities and complementary information

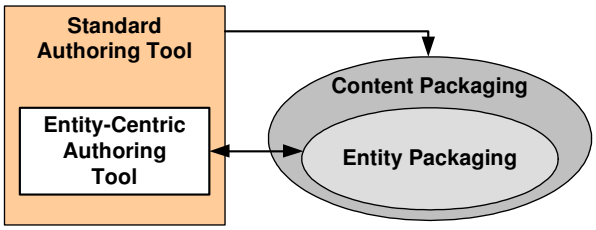

Fig.5: Possible extension to current state-of-the-art authoring tool

7 Conclusion

The analysis of the state-of-the-art for entity-centric repository confirms the need for a comprehensive repository of semantic metadata which can be used to fulfil the needs of the Semantic Web. These needs have been stressed in two projects, KIM and VIKEF. Both projects adopt an Entity-Centric Approach for indexing, retrieval and management of information resources. Other projects, like NewsExplorer also confirm the importance of using such an approach for clustering information from heterogeneous sources and presenting it to the user through one central access point. In addition, the development of a user-intuitive entity-centric multimedia authoring environment as a front-end for such a repository can act as proof of concept and a showcase.

References:

[1] BRICKS Project, http://www.brickscommunity.org, Accessed: Apr 2007

[2] Cees G.M: Snoek and Marcel Worring, "Multimodal Video indexing: A review of the state-of-the-art", 2005

[3] Ekin, A., Tekalp, A. M., "Integrated semantic-syntactic video event modelling for search and browsing," *IEEE Trans. on Multimedia*, vol. 6, no. 6 , Dec. 2004, pp. 839-851

[4] EMM NewsExplorer, http://press.jrc.it/NewsExplorer, Accessed: Apr 2007

[5] IST project OpenDlib, http://www.opendlib.com, Accessed: Apr 2007

[6] ISTWeb – Content – Knowledge and Content Technologies – Projects – VIKEF, http://cordis.europa.eu/ist/kct/vikef_synopsis.htm, Accessed: Apr 2007

[7] KIM Platform: Knowledge & Information Management, http://www.ontotext.com/kim/index.html, Accessed: Apr 2007

[8] Macromedia – Authorware 7, http://www.adobe.com/products/authorware, Accessed: Apr 2007

[9] MPEG7 Overview, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm, Accessed: Apr 2007

[10] Popov, B., Kitchukov, I., Angelov, K., Kiryakov, A., "Co-occurrence and Ranking of Entities", *Ontotext Technology White Paper*, May 2006.

[11] Velthausz, D.D, "Cost-Effective Network-Based Multimedia Information Retrieval", *PhD thesis, University of Twente*, 1998

[12] VIKEF, "R&D Highlight - Entity & Identity Management", http://www.vikef.net/news/upload/uploadFiles/VIKEF_Poster2c_A3.pdf, Accessed: Apr 2007

[13] VIKEF, Virtual Information and Knowledge Environment Framework, http://www.vikef.net Accessed: Apr 2007

[14] What is Okkam – Okkam Workspace Portal, http://okkam.dit.unitn.it:8088/okkam, Accessed: Apr 2007.