Executive Summary of Deliverable 19.2

Application of X-Media Technology to the Public Use Case

Authors: Johannes Busse, Ontoprise, busse@ontoprise.de

Aba-Sah Dadzie, The University of Sheffield, a.dadzie@dcs.shef.ac.uk

Lei Xia, The University of Sheffield, l.xia@dcs.shef.ac.uk

Authors: Spiros Nikolopoulos, CERTH, nikolo@iti.gr

Thomas Franz, University of Koblenz-Landau, franz@uni-koblenz.de

Christoph Ringelstein, University of Koblenz-Landau, cringel@uni-koblenz.de

Authors: Alberto Lavelli, FBK-irst, lavelli@fbk.eu

Work-package 19: Public Use Case

Type: Public Deliverable

Distribution: Consortium

Status: Final

Date: 31.08.2009

Deliverable Coordinator: Alberto Lavelli

Reviewers: Steve Fullerton

Area Coordinator: Marina Giordanino, CRF

Project Coordinator: Fabio Ciravegna, University of Sheffield

EU Project Officer: Leonard Maqua

ABSTRACT

The deliverable reports the current status of the public use case and describes the X-Media technology applied in the use case.

As already reported in Deliverable D19.1, the definition of a public use case was needed to be able to publicly demonstrate the X-Media technologies. This is necessary because, due to 'Non-disclosure agreements' (NDAs) signed between the consortium and industrial partners, the research partners are unable to publish or demonstrate their research results or technologies with the datasets provided by industrial partners.

The aim of the public use case is to provide a data set and an ontology that are useable for public demonstration and publication purposes outside the consortium. The use case has to be able to accommodate the initial objectives specified by the original use cases from industrial partners. In order to test the technologies developed within the consortium it is necessary to provide a dataset that is large scale and able to accommodate a number of different information extraction tasks (namely, named entity extraction, relation extraction, image classification/clustering object recognition). In addition, the data set needs to be easily understandable without specific expertise.

These requirements were brought together to become the definition of the Bike Brake Fade use case, which is similar in spirit to the Rolls-Royce (R-R) Issue Resolution (IR) use case.

The deliverable describes the X-Media technologies suitable for integration in the public use case. First of all, Area 1 technologies are presented, more specifically Semantic Email with Process Support, which integrates the concept of semantic email developed in workpackage WP4 and methods on process support developed in workpackage WP1. The contributing components, namely COSIMail and DIALOG, as well as their technical integration are described and the user benefits by means of the bike brake scenario are briefly illustrated. The integration of COSIMail with DIALOG allows for tracking the history of the processing of files across multiple desktops.

Area 2 technologies are then described. First, those developed in workpackage WP5, i.e. the recognition of named entities and of relevant domain-specific terminology and extraction of relations between relevant entities. As for workpackage WP6, Content-Based Image Retrieval (CBIR) and Content-Based Image Clustering (CBIC) are described. The purpose of Content-Based Image Retrieval (CBIR) technology is to perform efficient content-based image retrieval on large-scale image datasets coming from highly heterogeneous environments, as in the case of industrial domains. The aim of Content-Based Image Clustering (CBIC) technology is to organise the image database according to visual content and, as much as possible, semantic content.

The adaptation of the R-R use case to the public use case includes extensive data collection (including both texts and images downloaded from the web) and the development of a new ontology. As for the former, we found a number of problematic issues: (i) lack of specific bike brake issue resolution scenarios; (ii) lack of expert knowledge/common sense of bike brake problem; (iii) although there is a significant amount of bike brake images, textual document content is very limited. Since X-Media considers tens thousands of documents as large-scale, manual creation of this amount of data is practically unfeasible. To overcome this difficulty, we have developed a template filling strategy that allowed us to simulate a noise process over the dataset.

The public use case is now not only defined in abstract but also it can be demonstrated practically using the XMediaBox GUI.