002068

ALVIS

Superpeer semantic Search Engine

STREP / IST

Milestone MS3.4

Semantic indexing support for the Alvis database engine framework

Title of contract	ALVIS - Superpeer Semantic Search Engine
Acronym	ALVIS
Contract number	IST-1-002068-STP
Start date of the project	1.1.2004
Duration	36 months, until 31.12.2006
Document name	Milestone MS3.4– Semantic indexing support for the Alvis database engine framework
Date of preparation	March 8, 2006
Author(s)	Mike Taylor
Coordinator of the deliverable	Index Data
	Phone : +45 3341 0100
	Fax : +45 3341 0101
	Email: mike@indexdata.dk
Document location	http://project.alvis.info/copies/2005pdf

Project funded by the European Community under the "Information Society Technology" Programme (2002-2006)

1 Task 3.4: semantic indexing support for the Alvis database engine framework

Task 3.4 in WP3 is semantic indexing support. The project proposal document describes it as follows:

> Within the database engine framework, provide richly structured indexing capabilities to support the complex results of semantic analyses such as sparse feature vectors.

The database framework adopted by Alvis is the Index Data's open source Zebra information management system, described and available for download at http://www.indexdata.com/zebra/

2 Milestone 3.4

Milestone MS3.4 is described in the proposal document as follows:

> Semantic indexing support complete in database engine framework.

This milestone was was fulfilled by the release to Alvis consortium partners of version 1.4 of Zebra on 14 June 2005, including the enhancements described below. Public releases of Zebra continue in the 1.3.x series for now, with the 1.4 functionality not yet publicised outside the consortium. Version 1.4 will be publicly released when the experimental features have been sufficiently tested and documented to be deployed in mission-critical applications.

In the mean time, daily snapshots of the 1.4 series Zebra can be downloaded from http://ftp.indexdata.dk/pub/snapshot/

(This report is not the milestone: the software release is the milestone. This report serves only to document the release and its satisfaction of the milestone criteria. Although the milestone was met on schedule, the need for a document marking it has only recently become apparent, hence the late date on this document.)

3 Zebra enhancements in version 1.4

Semantic indexing is supported through the provision of a a flexible general mechanism for extracting information from documents submitted for indexing. When the Alvis filter of Zebra is used, each XML document fed to it is transformed by an XSLT stylesheet to yield a set of instructions that explicitly indicate what content should be added to each of a set of arbitrarily many indexes. Therefore a suitable indexing stylesheet could, for example, extract the names of genes (as marked up by WP5's Linguistic Analysis module within the Alvis Pipeline architecture), and index them in a dedicated index reserved only for gene names.

Given the richness and depth of the Enriched Document format described in D3.1, this mechanism allows enormous flexibility in the indexing regimes which may be configured. Zebra's indexing engine allows new indexes to be created on the fly, so there is no limitation to the possible indexing schemes that can

be created by sufficiently complex documents and suitable indexing stylesheets. The challenge will be to create a UI that provides an intuitive approach to using this information.

Related to the XSLT-driven indexing facility is an analogous new record-display facility, whereby Zebra administrators may supply XSLT stylesheets that are applied to records at retrieval time in order to present various subsets and mappings of the records to the user.

The new XSLT-based record facilities of the Zebra 1.4 series are documented within the pre-release tarball available from http://ftp.indexdata.dk/pub/snapshot/