Knowledge Base for RTD Competencies in IST

Deliverable D5.2
Portal with enhanced functionality

Author(s): Jure Ferlež, Brigitte Jörg
Identifier: D2.3
Work package: WP2 Portal Design
Lead Partner: Institute Jozef Stefan (JSI)
Partner(s): Deutsch. Forsch. für Künstliche Intelligenz (DFKI)
Ontotext Lab, Sirma Group (ONT)
State of document: final
Version: 1.0
Dissemination Level: Public
Date: 2006-03-22

This document is part of a SSA project funded within the IST Programme of the Commission of the European Communities – Project No: **FP6-2004-IST-3 – 015823.**

IST World Consortium

Participant Name	**Participant Short Name**	**Country**
Deutsches Forschungszentrum für Künstliche Intelligenz (**Co-ordinator**)	DFKI	Germany
Institute Jozef Stefan	JSI	Slovenia
Ontotext Lab, Sirma AI EAD	ONT	Bulgaria
RTD Talos	Talos	Cyprus
Institute of Information Theory and Automation	UTIA	Czech Republic
Archimedes Foundation	AF	Estonia
Computer and Automation Research Institute, Hungarian Academy of Sciences	MTA SZTAKI	Hungary
Institute of Mathematics and Computer Science, University of Latvia	IMCS	Latvia
Lithuanian Innovation Centre	LIC	Lithuania
Projects in Motion	PiM	Malta
Technical University of Silesia	SUT	Poland
National Institute for Research and Development in Informatics	ICI	Romania
Slovak University of Technology	STUBA	Slovakia
TUBITAK	TUBITAK	Turkey
CCLRC	CCLRC	United Kingdom

Abstract

This document provides a description of the enhanced version of the IST World portal (http://www.ist-world.org/), which now brings the following functionalities:

1. Partner finding tool (beta)
2. Expertise Identification (beta)
3. Information Repository (beta)

The current release of the IST World portal provides a substantially improved user interface including additional innovative services and layout. For the information repository, administration and data provision interfaces have been implemented for a better maintenance.

The latest IST World portal offers a partner search or matching service as part of the full-text search. Search results are now based on rankings and allow for an identification of optimum consortia based on their competence and expertise. Expertise identification is available in the shape of a two-dimensional competence landscape based on the full-text search results. A topic-driven content classification is added for navigation. The collaboration diagram is now realized with latest Flash technology. A multilingual entry page has been published.

At the current state of the project we are still concerned with data provisions from our project partners, the maintenance tool for data providers is ready with this version. We are now able to focus on data provision and in parallel to prepare the tool for community building. User authentication and authorization is available with the maintenance tool for data providers. Extensions for community building are being prepared for the next release.

Content Table

1. Introduction

The current release of the IST World portal (http://www.ist-world.org/) provides a substantially improved user interface including additional innovative services and layout. For the information repository, administration and data provision interfaces have been implemented for better maintenance. Extensions for community building are being prepared for the next release:

- **User Interface Functionalities (Web Portal):**
    - Improved Layout
    - Partner Finding
    - Categorical Search
    - Competence Diagram
    - Enhanced Collaboration Diagram
    - Multilingual User Interface

- **Data Administration Site (Offline Processing of Data):**
    - Importing of CXML[1] files
    - Processing of IST World's textual content
    - Importing DMOZ classification schema
    - Automated classification of data into DMOZ schema

- **Data Provider Administration Site (Manual Manipulation of Data):**
    - Human manipulation of provided data
    - Data Provider Authentication and Authorization

[1] CXML = CERIF-based XML (Specified in deliverable D3.2 Base Set of Data).

2. Enhanced portal functionalities

Additional to the first version of the portal described in deliverable D5.1 *First version of the portal with basic functionalities*, the current document describes the 2nd, enhanced version of the latest portal release.

2.1. Improved Layout

When we started to implement the new IST World functionalities for the enhanced version of the portal, the structure of the previous user interface was not open for additional features. Therefore, we decided to re-organize the user interface structure for being scalable and open with respect to future extensions. At the same time we now concentrate on the introduction of the portal content first.

The entry page of the IST World portal (http://www.ist-world.org/) now contains images representing the four IST World content types with corresponding colors: Organizations, Projects, Experts, Publications (see figure 1).

Figure 1: *Entry page of the IST World portal*

The content types are also available from the horizontal navigation bar on top of each page and are equally supported by corresponding colors as depicted in figure 2 for organizations (blue).

***Figure 2**: Organizations view in the IST World portal*

The left menu (Browse/Search) is meant for query definitions and offers search and browse options. The toolbox on the right hand side offers the analytic tools for an analysis of the query results. This is a clear usage structure, indicating the content on top and directing the user from the query definition (left) over the results (middle) to finally the analytics tools (right).

2.2. Partner Finding

With the current version of the portal, the partner finding tool is in beta stage. With the new version of the portal the ranking of search results is improved. Search results are now ranked according to their relevance towards the search query. The new ranking was achieved by pre-processing the textual content in the IST World repository and by using the full text search capabilities of the MS SQL Server 2005 RDBMS. The most relevant result is now listed at the first position of the result list. This is a major improvement compared to the first version where the results were listed alphabetically.

The improved ranking of the search results allows for partner search or partner matching. This service allows for a search of people, projects or organizations, which are most likely to be interested in a field of study, submitted as a query. For example: Users looking for partners in the field of "machine learning" in Germany define a query within Organizations, and are provided with a list of potential partners. The partner finding tool is not explicitly a tool in the toolbox on the right hand side, but can be processed within each content type of the IST World portal.

2.3. Competence Diagram

The Competence Diagram in the current, enhanced version of the IST World portal is in beta stage. It is produced using a state-of-the-art approach, combining text analysis, statistical and machine learning methods (Fortuna 2005). The Competence Diagram is a two dimensional graph of keywords together with frequencies of entities related to the displayed keywords. The expertise or competence of the selected subsets of instances is represented by the displayed keywords. Similar entities (by competence) are grouped together and shown as hills (blue background) in the landscape. The two dimensional mapping is computed using offline processing of the data in the IST World data store for a production of aggregated text describing all of the main entities. This text is then analyzed with the TextGarden software developed at Institute Jozef Stefan (http://kt.ijs.si/Dunja/textgarden/) to produce a projection into a 2D space, which is rendered to the user by the Flash Player web browser plug in from Macromedia.

The competence diagram allows for a presentation of query results in a two-dimensional landscape, where the blue background-colors indicate competence clusters and the red crosses represent individual entities. Moving the mouse over the landscape with a blue focus circle along mouse movements, highlights the involved entities as depicted in figure 3 and 4. For example: A search for projects in the field of "natural language" results in a landscape of related keywords (Speech, Web, Regulation, …) indicating the frequency of involved entities based on this keywords as red crosses. Moving the mouse over the landscape highlights keywords in the near neighborhood (Speech, Dialogue, Learning, …) of individual entities (Project: Specificity in Graphics and Natural Language) as shown in figure 3.

***Figure 3**: Competence Diagram: Presenting a landscape of "natural language" related keywords and the frequency of related entities*

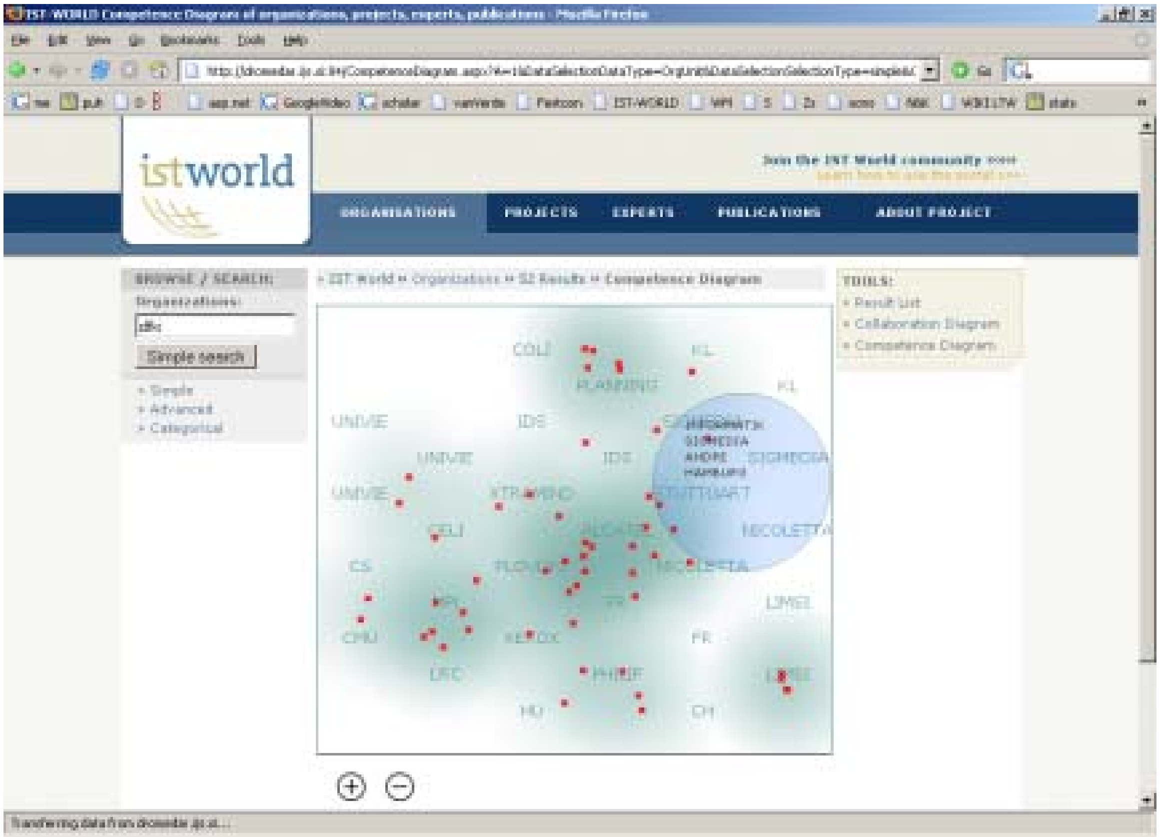

Figure 4: Competence Diagram: Presenting a landscape of "DFKI" related keywords and the frequency of related entities.

2.4. Enhanced Collaboration Diagram

With the current version of the IST World portal we provide an updated version of the Collaboration Diagram. The new diagram is implemented in Flash technology and allows for a more interactive usage than the previous version (see figure 5). It allows for:

1. Clicking on the nodes of the graph to get details of a selected instance node.
2. Hiding the edges of the graph according to weight for only to display the most important collaborations between entities to be displayed.
3. Hiding the labels of the edges and nodes of the graph to view the topology of the collaboration diagram more clearly.

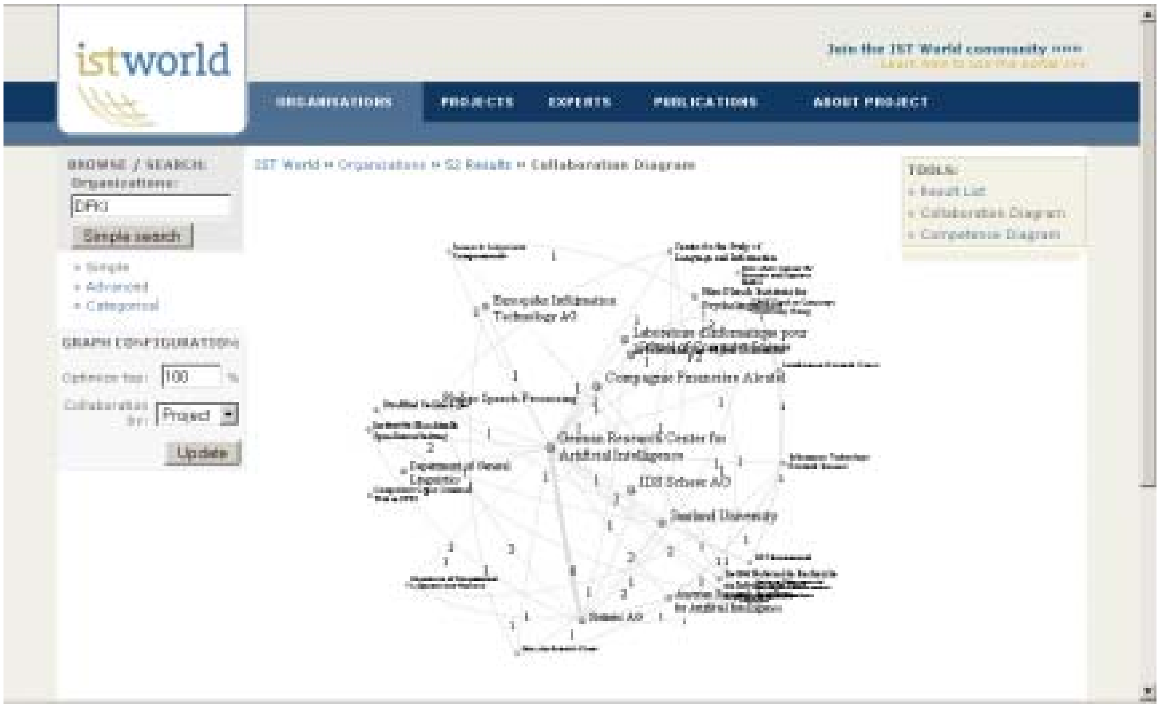

***Figure 5**: Collaboration Diagram: Based on a query for "DFKI" within Organizations.*

2.5. Multilingual User Interface

The layout of the portal has been improved and consequently the translations are effected. We now wait for feedback from the IST World Consortium and prepare the next call for the left translation components afterwards. As described in deliverable D4.1 *Multilingual User Interfaces*, the implementation of multilingual interfaces is a highly iterative process. Each change in the portal results in a change of involved translations. With this release we concentrated on a final layout and on the functionalities of the portal and now wait for feedback from the IST World consortium, only then we will start another process for collecting translations, due to efficiency reasons. As described in deliverable 4.1 the translations for the basic functionalities are available, however with the new functionalities some components are still missing due to the latest changes. For now we decided to release translations for the entry pages and left the website itself all in the English language.

2.6. Categorical Search

The need for a topic-driven content classification was identified with deliverable D1.2 *Data Model for Knowledge Organisation*, and a general methodology for "schema integration" has subsequently been introduced as part of deliverable D2.3 *Portal Architecture Specification*. Most of the IST World data will be automatically crawled from the Web, resulting in huge amounts of data. A classification of the crawled sources according to specific topics allows for navigation to the topics of interest. The need for providing a means for topic-based user access has been discussed at the 1st Consortium Meeting in Vilnius and was agreed to be necessary.

The categorical search was not planned in the original work, however this extension was considered to be supportive for the user. A topic-driven navigation interface is now available based on the Science part of the dmoz[2] taxonomy, including most of the relevant scientific areas for IST World. Automatic procedures for learning the classification of documents into the dmoz taxonomy were developed in previous projects (Grobelnik & Mladenic 2005) and will be applied for the portal so that IST World instances will be aligned to corresponding classes. The navigation menu allows for topic-driven queries as demonstrated in figure 6, where the Organizations are listed that are classified according to Science/Technology. The navigation can be started from the Science node to deeper levels. At each topic level, the number of contained instances is shown in dmoz way (see figure 6). A decision box to allow for inclusion or non-inclusion of sub-categories is available.

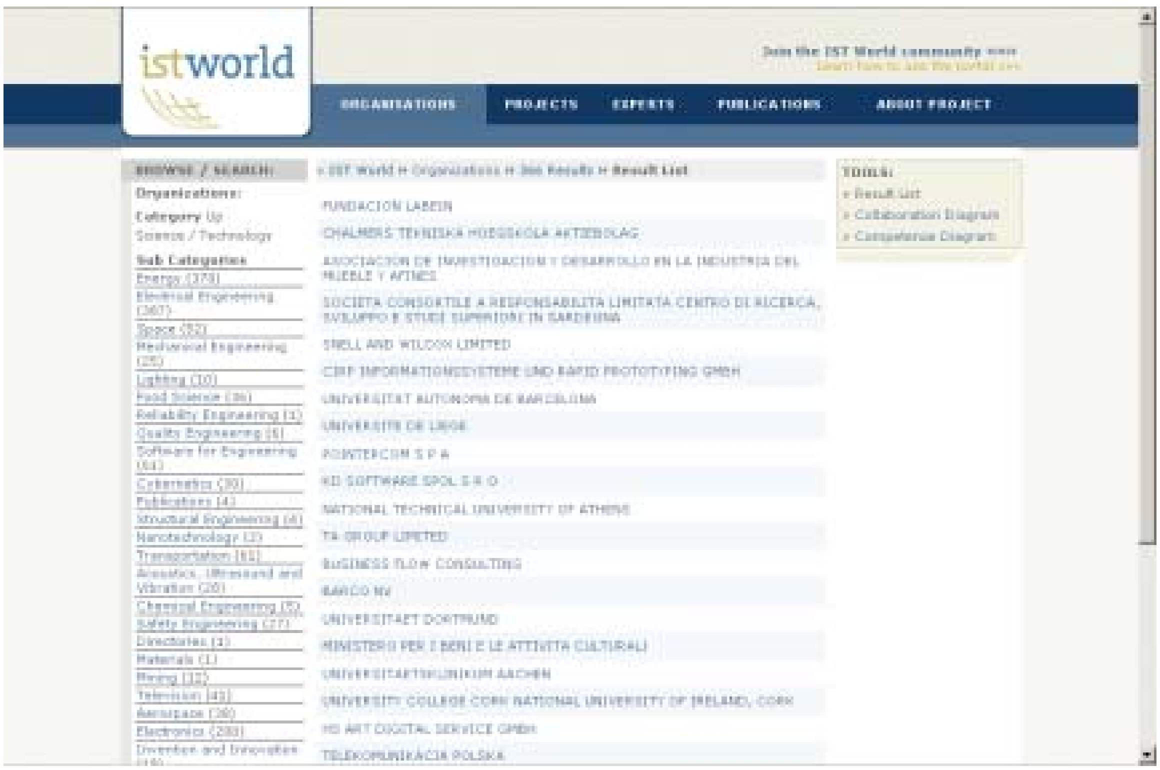

***Figure 6**: Topic-driven queries*

The automatic classification into the DMOZ hierarchy was achieved by offline processing of the data in the IST World repository. A previously developed automatic classifier into a scientific part of DMOZ hierarchy was used on the aggregated text describing individual entities to produce the classification of the entities. We have used a manually setup list of allowed and denied DMOZ subcategories which is still open for manipulation in order to constrain the classification to categories which semantically correspond to research and development topics in the IST domain. Using this technique we classified each of the entities in the IST World repository into the most probable DMOZ category.

[2] The Open Directory Project (dmoz): http://www.dmoz.org/

3. Data Administration Site

The IST World Data Administration site is setup (http://www.dromedar.ijs.si:85/). It is used for the offline processing of the IST World data. According to deliverable 2.3 *Portal Architecture Specification*, a web site is setup, which enables authenticated users to manipulate the IST World data. This includes:

1. Importing data in the form of CXML files
2. Processing of the portal's textual content
3. Importing of the DMOZ classification schema
4. Automated classification of data into the DMOZ schema

3.1. Importing CXML files

The Data administration site allows authenticated users to upload validated XML files, based on the CERIF XML schema specification defined in deliverable D3.1 *Data import/export specification as XML Schema*. The CXML data are then processed and imported into the IST World data repository. The IST World Repository allows for data insertion, deletion or updates.

3.2. Textual Content Processing

Most of the IST World's analytic techniques are based on the analysis of textual content. Therefore the affiliated text of individual entities had to be processed and stored in a correct way, to enable (1) efficient entity text description extraction and (2) efficient entity full text indexing.

Efficient text description extraction (1) is being used by the text analysis tools in the IST World portal, like the Competence Diagram, the DMOZ classification and the Partner Finding Tool (Improved Search Result Ranking). These tools make use of all the available textual description of entities as an input for a proper analysis process. Therefore the implementation of the text description extraction had to be implemented efficiently.

Full text indexing (2) is used in the IST World repository to enable a fast search within the database including the ranking of the results according their relevance towards the search query. The constraints of the MS SQL Server Out-Of-the-Box Full Text Indexing requires that the text to be indexed is stored inside one relational table.

For the above reasons all the textual content in the IST World repository is duplicated and stored inside one relational table called TextContent. This table is then full text indexed and allows (1) efficient text retrieval and (2) efficient full text indexing.

This process of duplicating all the text of the IST World repository into a single table can be started from the Data Administration Site by authenticated users. Users can select to re-process all entities or just new ones.

3.3. Importing the DMOZ schema

In order to enable the Categorical Search functionality, first the categorization schema has to be imported into the IST World repository, which can be done by an authenticated user of the data administration portal.

3.4. Automated classification of data into DMOZ schema

In order to enable the Categorical Search functionality, all of the IST World entities had to be classified to one of the DMOZ categories. The process of classifying can be started by authenticated users. Users can select to classify unclassified entities or to re-classify all entities. The automated classification process is controlled by specifying the DMOZ category branches into which the processing instances can or cannot be classified. This information is stored inside an XML specification file.

4. Data Provider Administration Portal

The Data Provider Administration Portal is setup (http://dromedar.ijs.si:87/). It is prepared for a human manipulation of the data stored in the IST World repository. This allows IST World data providers to manually change the data they provided. The site uses authentication and authorization of the users. It allows:

1. Controlled data reporting according to the roles of the authenticated and authorized user. E.g. the data providers are only allowed to see the data they provide.

2. Controlled data inserting, updating and deletion according to the roles of the authenticated and authorized user. E.g. the data providers are only allowed to see the data they provide.

5. Conclusion

The IST World portal with Enhanced Functionality was successfully released. It now provides users with tools for efficient searching, browsing and analysis of the data contained in the IST World repository and thus to produce new information and knowledge about activities and competences of IST experts, organizations and consortia.

Web interfaces for data provision and data manipulation are ready for usage with this version and have been applied already for the Estonian datasets based on CXML. Now we will be able to process data imports efficiently.

Community building components and associated Web forms are being prepared for the next version, which allows us to concentrate on data provision by our partners. Moreover we are now able to focus on usability aspects for the forthcoming deliverable D7.1 *User Satisfaction Evaluation* and ask for feedback from the User Steering Board.

6. Bibliography

(Fortuna 2005) Fortuna B., Mladenic D., Grobelnik M. *Visualization of text document corpus*. Slovenian KDD Conference (SiKDD 2005). In Proceedings: International multi-conference Information Society IS-2005, Ljubljana, Slovenia.

(Grobelnik & Mladenic 2002) Grobelnik M., and Mladenic D. Approaching Analysis of EU IST Projects Database. In Proceedings: The International Conference on Information and Intelligent Systems (IIS-2002), 2002.
http://www-ai.ijs.si/DunjaMladenic/papers/SolEuNet/EUProjectsIISSep02.pdf

(Grobelnik & Mladenic 2005) Grobelnik M., Mladenik D. Simple classification into large topic ontology of Web documents. In Proceedings: 27th International Conference on Information Tech-nology Interfaces (ITI 2005), 20-24 June, Cavtat, Croatia.
http://eprints.pascal-network.org/archive/00000844/01/GrobelnikITI_20April2005.pdf

(Jörg & Uszkoreit 2005) Jörg B., Uszkoreit H. The Ontology-based Architecture of LT World, a Comprehensive Web Information System for a Science and Technology Discipline. In: Leitbild Informationskompetenz: Positionen - Praxis - Perspektiven im europäischen Wissensmarkt. 27. Online Tagung (zugleich 57. Jahrestagung) der DGI. Frankfurt am Main, 23.-25. Mai, 2005.
http://www.dfki.de/~brigitte/publications/dgiOnline2005.pdf

(Salton 1991) Salton, G. *Developments in Automatic Text Retrieval*. Science, Vol 253, pages 974-979, 1991.