

IST World: European RTD Information and Service Portal

Brigitte Jörg	Jure Ferlež	Edward Grabcewski	Mitja Jermol
Language Technology Lab, DFKI GmbH, Saarbrücken, Germany	Dept. of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia	E-Information, Business and Information Technology Dept., CCLRC Rutherford Appleton Lab, Oxfordshire, UK	Dept. of Knowledge Transfer, Jozef Stefan Institute, Ljubljana, Slovenia

Summary

The IST World Web portal (<http://www.ist-world.org/>) integrates information about RTD actors such as organizations and experts on a local, national and European level and shows the context of their co-operation in joint projects and publications. Although the portal is aimed at promoting RTD competencies in IST in the New Member States (NMS) and Associate Candidate Countries (ACC), the long-term goal is to analyze the competence map and collaboration diagrams of Europe. The portal is built on technology for Current Research Information Systems based on state-of-the-art knowledge technologies, techniques and tools that have been developed by DFKI, Jozef Stefan Institute, and Ontotext. The portal is structured into the set of functionalities that represent personal and organizational competencies, expertise and social network analysis. Moreover, the IST World portal will overcome the shortcomings of existing on-line services by offering advanced analytical and prediction services and facilitate and foster the networking among research actors and their involvement in joint RTD activities. In this paper the innovative IST World portal services are presented from a user perspective.

1 Introduction

Most European countries collect and store their research information¹ in national RTD² repositories. This information is often spread across several regional or local repositories that are realized with their proprietary encoding and structure. It is very difficult to get additional information value out of multiple collections of RTD information, spread over several individual sources. By integrating information from various sources into the integrated database that is based on the CERIF standard, IST World offers possibilities to discover existing and potential competency and collaboration networks. Furthermore, a lack of information about RTD competencies has been identified in Europe particularly in the NMS and ACC countries, where the competencies are not systematically gathered or are not known well enough. Therefore the consortia for research pro-

¹ An overview catalogue of available research information systems in European countries and worldwide is maintained by the Royal Netherlands Academy of Arts and Sciences (KNAW). The catalogue is available at: <http://www.onderzoekinformatie.nl/en/oi/dris/search/>

² Research and Technology Development (RTD)

jects are mostly built from the partners, that have been active in the previous projects while the new, innovative, small to medium-sized enterprises (SMEs) cannot be found easily. The IST World portal will provide access to currently hidden knowledge about European RTD competencies in IST by offering innovative functionalities to identify competence clusters and predict the development dynamics. IST World is a Specific Support Action project, funded within the Sixth Framework Programme of the Commission of the European Communities, which started in April 2005, with a duration of 30 month. The IST World portal is being realized with innovative techniques and tools developed in other previous and current research projects in which IST World partners are active (Erbach et. al. 2005).

2 IST World Portal Services

The following IST World portal services that were build on top of a CERIF³-based data repository are currently available: (1) complex search and navigation functionalities to retrieve relevant information from the data repository, (2) automated analytical methods and visualization techniques to present the results. The analytic processes depend on the pre-defined sets of information. First, in a *selection step* the entities (organizations, experts, publications, projects) or subsets of the entities that represent the target of interest are to be specified by making use of available search and navigation functionalities. Second, in an *analysis step* one of the analytical methods is applied upon the retrieved set of entities. The results are presented by advanced visualization techniques.

2.1 Search and Navigation

Full text search is currently employed for all of the IST World entities, simple and advanced query templates are available for organizations, projects, experts and publications. In the next step a topic-based navigation interface for projects, experts and publications will be provided, based on the Science part of the dmoz⁴ taxonomy.

2.1.1 Complex Full Text Search

A complex full text engine allows for simple or advanced search within the IST World repository. By default the simple search is enabled to check the available entity values for query words. A more advanced search allows for entity-based queries according to conjunctive constraints like organisations that are located in Germany or projects that are funded in the 5th Framework Programme. The search results are presented in alphabetical order as hyperlinks pointing to the details of entity instances (Ferlež et. al. 2005). At the instances level a browsing or navigation to related instances like in LT World (Jörg & Uszkoreit 2005) is implemented based on collaborative graphs.

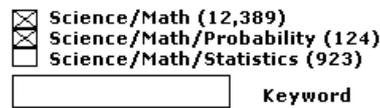
³ Current European Research Information Format (CERIF)

⁴ The Open Directory Project (dmoz): <http://www.dmoz.org/>

2.1.2 Topic-based Navigation

A topic-driven navigation interface will be provided based on the Science part of the dmoz taxonomy, which includes 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 & Mladenić 2005) and will be applied for the portal so that IST World instances will be aligned to corresponding dmoz classes.

A dmoz-driven navigation menu will provide access to topic-based instance selections, starting from the dmoz Science node. A simple example (Figure 1) shows the topic enabled search user interface.



Science/Math (12,389)
 Science/Math/Probability (124)
 Science/Math/Statistics (923)

Keyword

Figure 1: dmoz-based Navigation Menu

Navigation will start from the Science node to deeper levels like Math. At each topic level, the number of contained instances will be shown in dmoz way. A decision box to allow for inclusion or non-inclusion of the class instances will be available. The navigation interface will allow for additional keyword constraints.

2.2 Automated Analysis and Visualization Techniques

A variety of tools will be available to analyze and visualize search and navigation results. With automated methods we will provide insight into current, past and partly to the future research activities based on subset selections from the IST World information repository. The following functionalities are planned to be implemented: Community Identification, Expertise Identification, Partner / Consortia Finder, Trend Identification and Forecasting.

2.2.1 Community Identification

The current community identification tool shows the social relations between IST World instances using standard Social Network Analysis techniques (Mika 2005, Grobelnik & Mladenić 2002). The results are visualized in a Collaboration Diagram (See Figure 2 + 3).

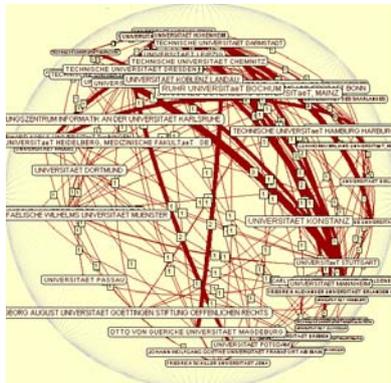


Figure 2: Collaborating Universities in Germany

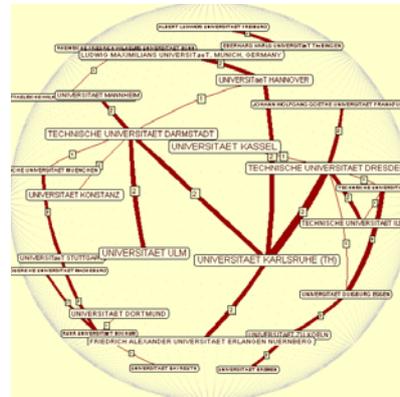


Figure 3: top 20% of Collaborating Universities in Germany

2.2.2 Expertise Identification

The expertise identification tool presents a description of selected entities. Complex expertise profiles provide users with an insight into the work, experience, and partly to the future ambitions of selected subsets of entities or instances. This functionality will be fully realized by implementing the Knowledge Map visualization and automated summarization techniques (see Figure 4).

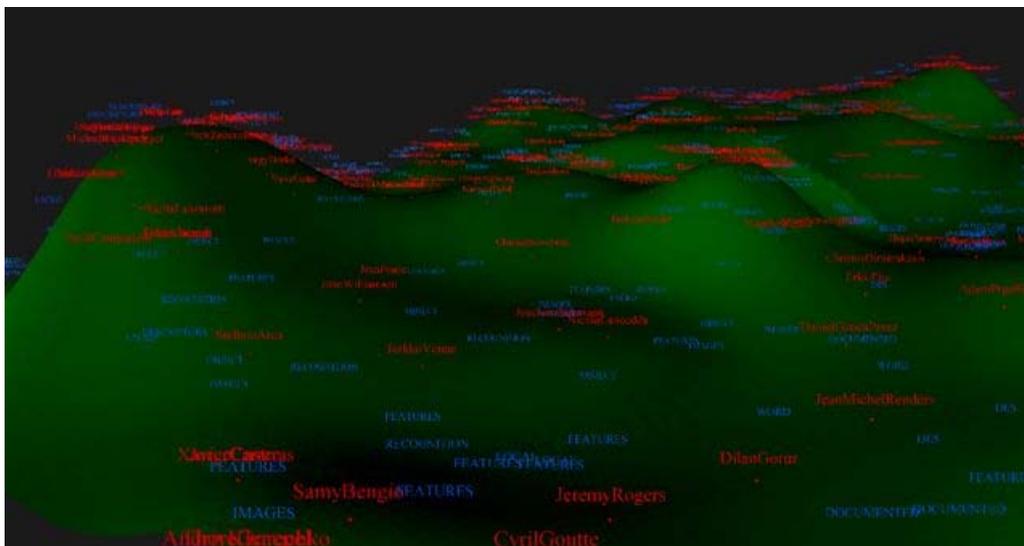


Figure 4: Knowledge map showing researchers (red) and their expertise (blue)

The Knowledge map is produced using a state-of-the-art approach, combining text analysis, statistical and machine learning methods (Fortuna 2005). A Knowledge map is a two dimensional graph of keywords together with frequencies of entities related to the displayed keywords. An

example of a knowledge map is presented in Figure 4. Knowledge or expertise of selected entity subsets or instances are represented by displayed keywords, the frequencies of the selected entities are shown as hills in a landscape defined by text content.

The automated summarization tool generates a short keyword description based on a selected set of text. Results are achieved using a simple word weighting scheme (Salton 1991) of all the words inside of the text documents and displaying the ones with highest weight.

2.2.3 Partner / Consortia Identification

The partner / consortia identification tool will provide recommendations on optimum subsets of IST World entities based on their expertise, past performance and/or trust (Figure 5).

Search for Thematic Consortium

Keywords:
knowledge technologies
semantic search

Proposed Consortium (based on project descriptions):

1. [0.846] **ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE**
◊ ALVIS, KNOWLEDGE WEB, DIP, AIM@SHAPE, DELIS, CASCOM, ENACTIVE, PRIME, BRICKS
2. [0.494] **UNIVERSITÄT INNSBRUCK**
◊ KNOWLEDGE WEB, SEKT, DIP, ASG, INFRAWEB
3. [0.456] **FRAUNHOFER GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V.**
◊ DIRECT-INFO, SATIRE, AIM@SHAPE, ACEMEDIA, PRIME, ASG, SEMANTIC HIPT, INTEROP, ORCHESTRA, BRICKS, USE-ME.GOV, K-WF GRID, BENTONWEB
4. [0.438] **UNIVERSITÄT KARLSRUHE (IHD)**
◊ KNOWLEDGE WEB, SEKT, DELIS, ACEMEDIA, ASG
5. [0.416] **JOZEF STEFAN INSTITUTE**
◊ ALVIS, SEKT
6. [0.398] **CONSIGLIO NAZIONALE DELLE RICERCHE**
◊ SEMANTICMINING, MUSCLE, METOCIS, AIM@SHAPE, INTEROP, BRICKS, ATHENA, CEC-MADE-SHOE, NEXTGRID
7. [0.396] **INTELLIGENT SOFTWARE COMPONENTS S.A.**
◊ SEKT, DIP, HOPS, ONTOGRID
8. [0.343] **THE UNIVERSITY OF SHEFFIELD**
◊ KNOWLEDGE WEB, SEKT, BRICKS, AMT
9. [0.343] **INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE**
◊ KNOWLEDGE WEB, MUSCLE, AIM@SHAPE, ACEMEDIA
10. [0.337] **THE VICTORIA UNIVERSITY OF MANCHESTER**
◊ KNOWLEDGE WEB, SEMANTICMINING, ONTOGRID, PRIME

Figure 5: Consortia Identification for projects on knowledge technologies and semantic search from Project Intelligence (Grobelnik & Mladenic 2002)

The criteria for best subset selection will be entered by the user as a list of keywords. Results will then be presented as an ordered list of entities, which best relate to the specified keywords. The first version of the partner/consortia identification tool will use simple frequency of search keywords to produce the correct order of potential partners. An example of partner identification analysis is displayed in Figure 5.

2.2.4 Trends Identification and Forecasting

The tool for trends identification and forecasting analysis will aim at finding relevant trends in research and forecast future RTD activities based on the monitoring of the past and current research initiatives, project domains and achievements. The object of analysis will be specified in the selection step by search and navigation of IST World entities. The identified trends will be visually presented using time dependant graphs. An example of visualization is a diagram presenting theme evolution through time shown on Figure 6. The forecasting information in the IST World portal will be induced and presented in the same way. Trends and forecasting functionality will use advanced methods and techniques in the field of dynamic graph analysis (Leskovec et. al. 2005).

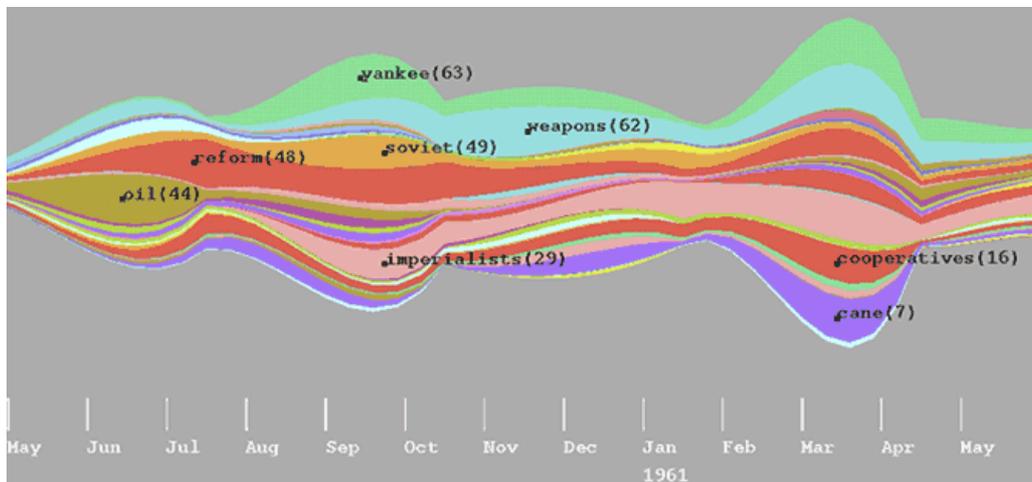


Figure 6: Visualization of trends of news topics from 1961 (Nowell et. al. 2001). [[ThemeRiver©](#)]

2.3 Portal Enhancements

Some of the services presented will only be available for registered users or members. In order to organize and manage the different views and access rights, the IST World portal will employ techniques for social trust networking. As IST World intends to take advantage of a warm welcome, a multilingual user interface in 14 languages is available.

2.3.1 Social Trust Network

To help users to build up their individual Social Trust Network the IST World portal will provide web forms for user registration and profile updates. Moreover additional functionalities, similar to existing services like LinkedIn⁵ or OpenBC⁶ will be implemented: (1) an invitation form for new users (i.e. sending an email), (2) a form for issuing a request for linking with another already registered user (sending an email and linking two users if request confirmed), (3) a form for requesting to share information across the trustful set of links (e.g. asking for information about the user which is not directly linked to me).

2.3.2 Multilingual User Interfaces

The user interface of the IST World portal will be extended with the following languages: Bulgarian, Czech, Estonian, German, Greek, Hungarian, Latvian, Lithuanian, Polish, Romanian, Slovak, Slovenian, Maltese, and Turkish. A translation of all contents is not considered, since people who work in transnational environments need to have a good working knowledge of English.

⁵ LinkedIn: <http://www.linkedin.com/>

⁶ OpenBC: <https://www.openbc.com/>

3 IST World Repository

The services of the IST World portal will be offered on top of the IST World repository and thus will depend on data input. Data will be provided by the consortium members in specified formats (Jörg 2005), by the community via Web forms and from the Web by automated crawling. We expect that the main source of data will be public databases.

3.1 Conceptual Baseline

The IST World portal needs a pragmatic view of the data as well as a powerful conceptual view. We needed a basic version of the portal and an operational data store very early in the project. At the same time we wanted to ensure that a deeper semantic analysis (Popov 2004) will be possible at a later stage. Therefore, we decided to start with a combination of two data models:

- *Relational model* – a detailed RDBMS schema and extension of the CERIF 2004 Full Data Model.
- *Conceptual model* – an ontology, allowing for proper conceptualization of the domain and deeper analysis of the RDBMS data.

During these initial phase of the project the portal uses only the relational model, implemented in a conventional RDBMS. Currently, the conceptual model (or ontology) is not operationally involved in the portal and is used only as a design guide and a base for developing a proper expertise modeling schema. Later, the ontology will be used in addition to a *semantic repository* for properly integrating the RDBMS data (Kiryakov et. al. 2005).

3.2 Technological Baseline

The Common European Research Information Format (CERIF) was formerly developed under the co-ordination of the European Commission to harmonize national Current Research Information Systems (CRISs) within Europe and is now in the responsibility of euroCRIS⁷. CERIF is an open set of guidelines prepared to deal with research information systems. More information on the CERIF data model, history and relevant *architectures* can be found in (Asserson et. al 2002) and (Jeffery 2004). A description of the main types of metadata and their application in CERIF-based information systems is provided in (Jeffery 1999).

The IST World implementation uses a relevant subset of CERIF entities and their relationships and follows the current CERIF practices in extending the data model. Database creation was facilitated by the use of SQL scripts⁸ provided by the CERIF task group and extended with additional entities and relationships.

⁷ euroCRIS: <http://www.eurocris.org/>

⁸ CERIF SQL: <http://www.eurocris.org/en/taskgroups/cerif/cerif2004/>

CERIF Extensions:

To meet the necessary storage requirements for the repository and for the offered portal functionalities, extensions to the CERIF model were necessary to cover the following requirements:

- to support the display of information on trends and the prediction of the state of European and national research activities
- to support computer aided social networking.
- to allow the user authentication
- to allow the source identification
- to store the content of publication documents and not only their metadata

The trends detection and prediction functionalities requires that the extended CERIF data model stores additional data and meta-data on scientific publications and information. Moreover, the functionality to support the provision of computer aided social networking enables users to search and collaborate with existing social networks and requires that we address the issues of privacy, trust and the interests of network members. All mentioned issues were input to the CERIF extensions for IST World CERIF based data model that is specified and documented in (Kiryakov et. al. 2005, Ferlez 2005, Jörg et. al 2005)

The IST World Web portal is implemented using Microsoft's Internet Information Server (IIS) technology based on the latest Microsoft .NET v2.0 implementation framework. The Web page design follows the ASP.NET programming framework recommendations. The IST World Repository is built using a MS SQL Server 2005 database running on a MS Windows 2003 Server using a fast 64-bit computer hardware.

4 Conclusion and Future Work

We presented the innovative services of the IST World portal for RTD competences in IST, which will be built on top of the IST World RTD competencies repository that is based on the CERIF standard. Some functionalities are already implemented and available at <http://www.ist-world.org/>. The full version of the portal will offer improvements in graph visualization so that hyperlinked graph navigation to individual entities will be possible. We plan to enrich search results with text snippets to indicate the relevance of individual results. An enhanced portal version is being prepared for February 2006. A full functionality of presented portal services is planned for the summer 2006.

5 Acknowledgements

This work is kindly supported by the Commission of the European Communities within the Sixth Framework Programme in IST - Contract no.: FP6-2004-IST-3 – 015823.

6 References

- (Asserson et. al 2002)** Asserson, A., Jeffery, K. and Lopatenko, A. *CERIF: Past, Present and Future: An Overview*. CRIS2002 Conference, Kassel, Germany.
<http://www.eurocris.org/en/taskgroups/cerif/articles/>.
- (Brank 2004)** Brank, J. *Drawing graphs using simulated annealing and gradient descent*. V: TRČEK, Denis (ur.), LIKAR, Borut (ur.), GROBELNIK, Marko (ur.), MLADENIĆ, Dunja (ur.), GAMS, Matjaž (ur.), BOHANEK, Marko (ur.). *Zbornik C 7. mednarodne multi-konference Informacijska družba IS 2004, 9. do 15. oktober 2004*, (Informacijska družba). Ljubljana: Institut "Jožef Stefan", 2004, str. 67-70. [COBISS.SI-ID 18596135].
<http://eprints.pascal-network.org/archive/00000744/01/JanezBrank-GraphDrawing.pdf>
- (Erbach et. al. 2005)** *Network Approaches to Current Research Information Systems*. e-2005: eChallenges Conference, October 19-21, 2005. Ljubljana, Slovenia.
- (Ferlez et. al. 2005)** Ferlez J., Jörg B., Jermol M. Public IST World Deliverable 5.1 – *First Version of the Portal with Basic Functionality*.
http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D5.1_FirstVersionOfThePortal.pdf
- (Ferlez 2005)** Ferlez J. Public IST World Deliverable 1.3 – *Data Model for Representation of Expertise*. http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D1.3_DataModelForRepresentationOfExpertise.pdf
- (Fortuna 2005)** Fortuna B., Mladenec 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 & Mladenec 2002)** Grobelnik M., and Mladenec 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/DunjaMladenec/papers/SolEuNet/EUProjectsIISSep02.pdf>
- (Grobelnik & Mladenec 2005)** Grobelnik M., Mladenec D. *Simple classification into large topic ontology of Web documents*. In Proceedings: 27th International Conference on Information Technology Interfaces (ITI 2005), 20-24 June, Cavtat, Croatia.
http://eprints.pascal-network.org/archive/00000844/01/GrobelnikITI_20April2005.pdf
- (Jeffery 1999)** Jeffery, K.: *Metadata*. euroCRIS CERIF TG Web page:
<http://www.eurocris.org/en/taskgroups/cerif/articles/>.
- (Jeffery 2004)** Jeffery, K.: *CRIS Architectures and CERIF*. euroCRIS CERIF TG Web page:
<http://www.eurocris.org/en/taskgroups/cerif/articles/>.
- (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>
- (Jörg 2005)** Jörg B. Public IST World Deliverable – 3.1 *Data import/export specification as XML Schemata*. http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D3.1FormalImportExportSpecification.pdf

- (Jörg et. al. 2005)** Jörg B., Ferlež J., Grabczewski, E. Public IST World Deliverable 1.2 – *Data Model for Knowledge Organisation*. http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D1.2DataModelForKnowledgeOrganisation.pdf
- (Kiryakov et. al. 2005)** Kiryakov A., Grabczewski E., Ferlež J., Uszkoreit H., Jörg B. Public IST World Deliverable 1.1 – *Definition of the Central Data Structure*. http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D1.1CentralDataStructure.pdf
- (Leskovec et. al. 2005)** Leskovec J., Kleinberg J. Faloutsos C. *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations*. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), Chicago, IL, USA, 2005. <http://www.cs.cmu.edu/~jure/pubs/powergrowth-kdd05.pdf>
- (Mika 2005)** Mika, P. *Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks*. Journal of Web Semantics. Vol. 3, Issue 2, 20 pages. <http://scholar.google.com/url?sa=U&q=http://www.cs.vu.nl/~pmika/research/papers/JWS-Flink.pdf>
- (Nowell et. al. 2001)** Nowell L., Havre S., Hetzler B. and Whitney P. *Themeriver: Visualizing thematic changes in large document collections*. IEEE Transactions on Visualization & Computer Graphics, 2001.
- (Popov 2004)** Popov B., Kiryakov A. , Ognjano D., Manov D., Kirlolov A. *KIM – a semantic platform for information extraction and retrieval*. Nat. Lang. Eng., 10(3-4): 375–392, 2004. http://www.ontotext.com/publications/KIM_SAP_ISWC168.pdf
- (Salton 1991)** Salton, G. *Developments in Automatic Text Retrieval*. Science, Vol 253, pages 974-979, 1991.

7 Contact Information

Brigitte Jörg (brigitte.joerg@dfki.de)
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH),
Saarbrücken, Germany

Jure Ferlez (Jure.Ferlez@ijs.si)
Dept. of Knowledge Technologies, Jozef Stefan Institute,
Ljubljana, Slovenia

Edward Grabczewski (E.Grabczewski@rl.ac.uk)
E-Information, Business and Information Technology Dept. CCLRC Rutherford Appleton Lab,
Oxfordshire, UK

Mitja Jermol (mitja.jermol@ijs.si)
Center for Knowledge Transfer, Jozef Stefan Institute,
Ljubljana, Slovenia