

# Analyzing European Research Competencies in IST: Results from a European SSA Project

Brigitte Jörg<sup>+</sup>, Jure Ferlež\*, Hans Uszkoreit<sup>+</sup>, Mitja Jermol\*

<sup>+</sup>German Research Center for Artificial Intelligence, Saarbrücken

\*Jožef Stefan Institute, Ljubljana

## Summary

With this paper we will present the approach of analyzing research competencies across European countries as performed within the EC funded SSA project IST World. We will therefore briefly introduce the format used for representing the involved entities and their relationships before providing a thorough insight into the repository by describing the process of data collection, data integration and data cleaning. Based on the knowledge about the data in the repository we will explain the IST World analytics by showing some interesting examples. Finally, related activities will be investigated and a conclusion will be drawn from the results and lessons learned. The IST World portal and tools are available online (<http://www.ist-world.org/>) and utilized by a growing user community.

## 1 Introduction and Background

Many European countries collect and store their research data in national repositories. The information is often spread across several regions or locations, and stored with proprietary encoding and structure. To find relevant information and get additional value out of multiple collections spread over several individual sources is difficult. A lack of information about European RTD competencies has been identified particularly in the New Member States (NMS) and Acceding and Candidate Countries (ACC) of the European Union (EU) where competencies are not systematically gathered or are not sufficiently known. As a consequence, consortia for research projects have mainly been built from partners that have already been active in previous projects while the innovative potential in new, small to medium-sized enterprises (SMEs) and even new branches of existing larger corporations was often neglected. By merging information from heterogeneous sources into one integrated repository, IST World provides a platform to discover and investigate the highly valuable knowledge about European RTD competencies, in IST. The heterogeneous sources include data bases filled by hand through knowledgeable innovation and RTD agencies in several NMS and ACC. They also include web-based information resources provided by the European Commission and by a commercial search engine. By offering tools to identify competencies on the one hand and means to add competencies on the other hand, the portal supplies a new and dynamic environment for collaboration, innovation and networking. The portal attracts its users with highly innovative analytics and visualizations, online and in real time. The IST World portal was built within the IST World project that started in April 2005 with a duration of 32 months. This paper will present available results at the end of the project in November 2007.

The public portal (<http://www.ist-world.org/>) integrates information about actors such as organisations and experts on a local, national and European level and shows the context of their co-operation in joint projects and publications.

## 2 CERIF Setup for IST World

The IST World project aimed at setting up an information portal with innovative functionalities to promote RTD competencies in the area of Information Society Technologies (IST), in NMSs and ACCs (Erbach et. al. 2005, Jörg et. al. 2006). The IST World portal architecture is based on the CERIF<sup>1</sup> 2004 format. The competencies according to IST World definition are basically represented by the core CERIF entities organisation, person, project, publication, their associated attributes, and the relationships between them. The list of the attributes employed is given in table 1. A presentation of the attributes and the relationships between the entities is considered useful for understanding the repository structure and thus the analytic tools built on top of it.

*Table 1: CERIF entities and employed attributes in IST World*

<b>Person</b>	<b>OrgUnit</b>	<b>Project</b>	<b>Publication</b>
ID	ID	ID	ID
URI	URI	URI	URI
Family Names	Acronym	Title*	Title*
First Names	Type	Abstract*	Type
Other Names	Name*	Keywords*	Publication Date
Sex	Research Activity*	Start Date	Reference
Academic Title	Expertise and Skills*	End Date	Reference Type
Research Interest*	Keywords*	Funding Programme	
Expertise and Skills*	Country		
Keywords*	Contact		
Language			
Nationality			
Contact			

The four entities are linked with each other in relationship entities by their IDs. Timestamps allow for a representation of multiple roles in any such composition within a relationship entity:

<sup>1</sup> CERIF: Common European Research Information Format (<http://www.eurocris.org/>).

\* CERIF allows for multilingual descriptions.

The CERIF datamodel originates back to the eighties (Asserson et. al. 2002). It supports a data-centric view on the entities involved in the scientific process, and the relationships between them (Jeffery et. al. 2002, Erbach 2006). The latest release of the CERIF model has been extended towards a full representation of publications and matured with capturing the semantics of relationships (Jörg et. al. 2008).

**Person Relationships:** Person\_Person, Person\_OrgUnit, Person\_Publication  
**OrgUnit Relationships:** OrgUnit\_OrgUnit, OrgUnit\_Publication  
**Project Relationships:** Project\_Project, Project\_OrgUnit, Project\_Person, Project\_Publication  
**Publication Relationship:** Publication\_Classification

### 3 IST World Repository

The IST World repository integrates research information from more than 15 European countries in English and many national languages. It contains FP5 and FP6 projects and involved organisations and experts from CORDIS (Thévignot 2000). Additionally, it captures information about national publications, projects, and involved experts and organisations in English and national languages from the following countries: Bulgaria, Czech Republic, Estonia, Cyprus, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Russia, Serbia, Slovakia, Slovenia, and Turkey. At the end of the project the repository contained about 96.000 organisation records, 41.800 project records, 60.000 (not cleaned) person records, 2.000.000 (not cleaned) publication records.

#### 3.1 Data Collection

The data have been collected from different sources: (i) crawled from public websites, (ii) collected manually without the availability of a national CRIS<sup>2</sup>, (iii) collected from structured or CERIF-based national CRISs, (iv) provided by the public (community). Except for the CERIF-based data, the collected sets had different labels for their attributes and for the relationships between their entities. The list of the involved data formats indicates the heterogeneous character of the datasets<sup>3</sup>:

- Data crawled from the Web; from CERIF-based CRISs; from public CRISs
- CERIF-based databases (MSSQL Server; MS Access) ; EPSRC database
- MSWord documents; MSEXcel documents
- Raw Text files; HTML files; XML files

Once collected, the available data have been extracted and mapped into the CERIF XML format to allow for a valid import. For validation, each XML file was run against pre-defined CERIF XML Schemas. Single schemas were available for the four entities organisation, person, project and publication. After successful validation, each dataset has been imported individually.

---

<sup>2</sup> CRIS: Current **R**esearch **I**nformation **S**ystem

<sup>3</sup> The whole range of data formats has been documented with deliverable D3.2 Base set of Data. (Grabczewski and Jörg 2005).

## 3.2 Data Integration

To enable the analysis of RTD competencies not only within single sets but at European level, all datasets in the IST World repository have been integrated into one. Within this integrated dataset all records kept their unique IDs and the reference to their origin. The merging of the datasets resulted in many duplicate records, due to an overlap of records across the datasets. It turned out, that the duplicates did not only result from the merging of datasets, but many duplicates had already existed within those single datasets before. The most obvious duplicates within the integrated IST World repository were identified for organisations and persons, a non significant number of duplicates were found for projects, the publication duplicates have been ignored.

The formation of the duplicate records in the IST World repository is understood. The occurrence of duplicate records in big data collections is a known problem (Winkler 2006). The same holds for duplicate records across data collections that have to be merged, and where a linkage between the duplicates is required. A formal description of the problem space at hand will allow for a better understanding of the problem and for the method that has been applied in reducing the duplicate entries within the IST World repository. An empirical evaluation of the method in one of the single datasets will allow for an estimation of the integration success.

### 3.2.1 Formal Problem Definition

Record linkage is the means of combining information from a variety of computerized files. It is also referred to as data cleaning or duplicate detection. The most basic application is identifying duplicates within a dataset or identifying duplicates across two datasets. If a single large dataset is considered, then the record linkage or matching procedures may be intended to identify duplicates (Winkler 2006).

*Formal Problem Description according to Winkler:*

PROBLEM: duplicate detection in record set A  
GIVEN: a set of records in A  
CLASSIFY: every pair  $(a,b) \in A \times A$   
into M set of true matches or U set of true non matches.

According to Winkler, search for redundant records is usually performed in two steps. In the first step (**called blocking**) different heuristics are used to identify potential redundant pairs. In the second step (**called matching**) the goal is to decide on every potential pair of records whether they represent the same real world entity. If so, the two records within the set A are matched. The idea is, to classify pairs of records  $(a,b)$  from the product space  $A \times A$  into M, the set of true matches, or U, the set of true non matches.

### 3.2.2 IST World Problem Definition

In order to get a basic idea about the integration difficulties and a possible solution in the scope of IST World, we investigated the typical duplicates and their inherent problem patterns within the individual IST World datasets by random samples. The samples showed that the most obvious duplicates were among the organisation records inside the CORDIS FP5 and FP6 dataset and

across the CORDIS FP5 and FP6 dataset. Not so many duplicate organisations were found within the national datasets. A lot of duplicate person records have been identified across all datasets, no duplicate records were found within the project datasets, and only some duplicate project records across the datasets, the publications have not been examined. As a consequence of the sample investigation, and by taking into account the scope of the IST World project, we decided to leave the project records, ignore the publication records, find a solution for the person records later, and first concentrate on cleaning the duplicate organisation records (Ferlez and Jörg 2007). The problem definition at hand therefore only refers to organisational records.

Most of the detected duplicate organisation records had slightly different names caused by additional special characters or character modifications. The most important patterns of differences between the duplicate organisation records were as follows (in descending order of importance):

- Capitalization, Lowercase Letters
- Blanks, extra Spaces
- Hyphens
- Quotes
- Coma in Different Places
- Article in Name
- Full stop in Name
- Incomplete Names
- English Translation
- Word Order
- Language Specific Characters (encoding Jorg instead of Jörg)
- Special Characters (wrong encoding of ‘&’ and ‘?’)
- Mixture of Organisation Names and Department Names
- Differences in Addresses

The knowledge about the patterns of differences for organisation records in the IST World repository was used for the adjustment of the integration mechanisms; first for the blocking and finally for the matching of the duplicate organisation records.

### 3.2.3 IST World Data Integration Approach

A five step approach has been setup to tackle the integration task within the IST World repository. We describe it by following the notation introduced in the Problem Definition, and as depicted in Figure 1.

**Blocking:** the magnitude of the problem space of considering all potential pairs  $(a,b) \in A \times A$  for duplicates is effectively reduced by full text indexing and querying (Zobel et. al. 1998). The applied technique was found most suitable for records with correct spelling in at least one of the common words of their names. The blocking procedure based on the full text search over names found those duplicate candidate records in the repository that were most probable towards all the records in the repository, based on their names. A query for *Jožef Stefan Institute* gave the following results (only the first ten results are presented here: (1) JOZEF STEFAN INSTITUTE, (2)

ENVIRONMENTAL SCIENCES / JOZEF STEFAN INSTITUTE, (3) TECHNOLOGY TRANSFER OFFICE/ JOZEF STEFAN INSTITUTE, (4) INSTITUT "JOZEF STEFAN", (5) STEFAN BATORY FOUNDATION, (6) STEFAN TISCHER LANDSCHAFTSARCHITEKT, (7) INSTITUT JOSEF STEFAN, (8) UNIVERSITY OF PAVOL JOZEF SAFARIK, (9) FORESTRY FACULTY, UNIVERSITY STEFAN CEL MARE SUCEAVA, (10) INSTITUTE OF MICROBIOLOGY OF THE GERMAN ARMED FORCES (BUNDESWEHR INSTITUTE OF MICROBIOLOGY)

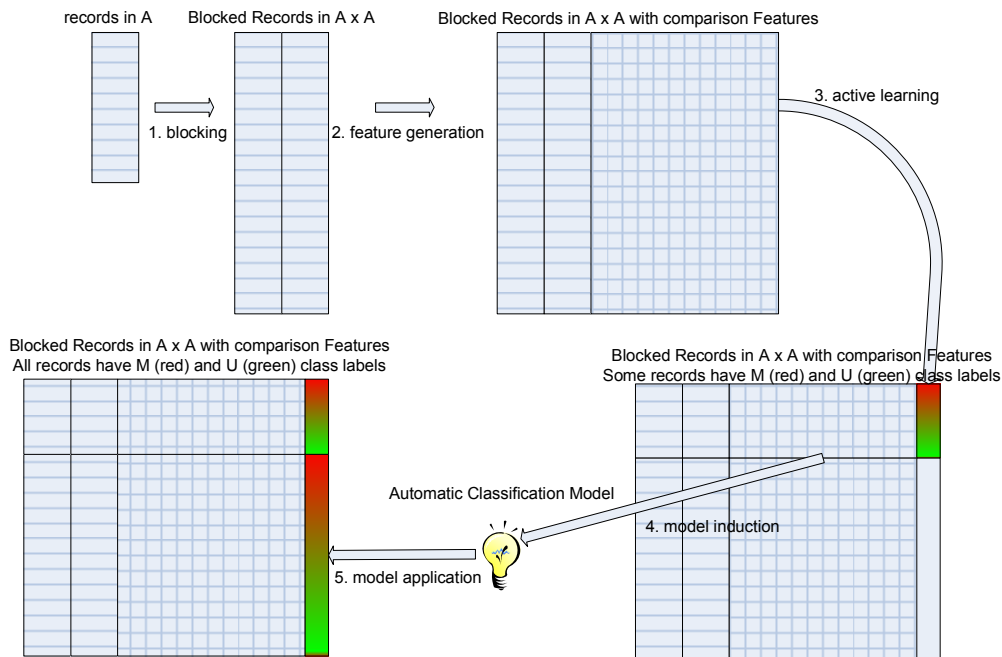


Figure 1: IST World Data Integration Approach

**Feature Generation:** A function to map a potential pair (a,b) into a useful pair comparison space vector was created, based on name and location information. To capture the different aspects of similarities between the name and location information of two records (a,b), four comparison approaches were developed.

First, we represented both, the name and the location strings with the vector of words using the well known technique called bag of words (Lewis 1998). This allowed us to capture the similarity of two records including words like LTD, GMBH or DOO. Next, we compared the order of words and the order of characters with a technique called string kernels (Lodhi et. al. 2002) to cope with the mixed word order or inclusion of special characters. To cope with spelling errors in words and with alternatives in word order, the edit distance measure method (Levenshtein 1966) was used. To combine all the mentioned approaches, the features of each approach were normalized and combined into one feature vector describing the similarity of a pair from the different comparison aspects (Ferlez 2007).

Table 2: Bag of Words comparison vector of a pair of names: "Jozef Stefan Institute" and "Institut Jozef Stefan"

<b>Feature Word</b>	Jozef	Stefan	Institute	Institut
<b>Feature Weight</b>	2	2	1	1

Table 3: String Kernel comparison vector of characters in a pair of names: "Jozef Stefan Institute" and "Institut Jozef Stefan".

<b>Kernel Parameters</b>	n=11	n=10	n=9	n=8	n=7	n=6	n=5	n=4	n=3	n=2
<b>Feature Weight</b>	0,005	0,009	0,014	0,023	0,036	0,053	0,077	0,109	0,148	0,194

**Active Learning:** A machine learning method called Active Learning (Tong et. al. 2002) was used to support the manual creation of the M and U sets of labelled pairs. To prepare the training set of potential pairs for an automated learning of the decision function, a set of potential pairs was evaluated and labelled manually into the sets M and U. The machine learning algorithm was used to decide, which potential pairs to label with M and U. For this approach we exploited the sampling behaviour of the active learning algorithm to effectively fill the sets M and U with labelled pairs based on the features generated during the Feature Generation. We selected the Support Vector Machine algorithm (Vapnik 1995) as the underlying learning algorithm to effectively fill the sets M and U. The Support Vector Machine algorithm as the underlying Active Learning algorithm results in an optimal sampling of the list of pairs for maximal information gain about the given classifier induction problem as the same algorithm will be used for the induction of the decision function.

**Classifier Induction:** A machine learning algorithm was applied to automatically construct a decision rule to answer the classification defined in section 3.2.1. The traditional record linkage problem can thus be solved by a standard supervised machine learning setting for learning a binary classification model. The Support Vector Machine (SVM) two-class classification model with linear kernel was used to generalize from the information in the labelled pairs to an automated classification mechanism into the sets M and U. The SVM was shown to work reasonably well when the attributes of the examples include bag of words features (Brank et al 2003) and other artificially constructed features like topological features, PageRank (Brin 1998), and others.

**Classifier Application:** Finally, the induced classifier was applied to the unlabelled candidate pairs to automatically classify them into sets M and U.

### 3.2.4 Evaluation of Results

The evaluation of the IST World Integration Approach is performed by measuring the percentage of the incorrectly classified pairs within the randomly selected subset sample of automatically labelled candidate pairs within the CORDIS FP6 organisations. The classifier goodness will be further evaluated by observing the ROC curve (Spackman 1989) on the hand labelled pairs. The ROC curve is a graphical plot of the sensitivity vs. (1 - specificity) for a binary classifier system as its discrimination threshold is varied. The steeper the curve rises, the better the classifier. The goodness of a classifier can be summed by providing the AUC (Area under ROC curve) scores of the classifier on the sampled set of pairs. Thus, the AUC can be interpreted as the probability that the classifier will correctly order a random positive and negative example. The closer this probability is to one, the better the classifier.

We have evaluated the IST World integration process for the cleaning of CORDIS FP6 organisations. When the integration process finished we evaluated our model by randomly sam-

pling and by hand evaluating of 1000 organisation record pairs. The comparison of manual and SVM assigned labels yielded in 30 correctly classified pairs into the set M, 1 incorrectly classified pair into set M, 35 incorrectly classified pairs into set U and 934 correctly classified pairs into set U. Figure 2 displays the ROC curve of the classifier on the evaluation sample. The model achieved an AUC score of 0.96. The observed precision was 97%, the recall 46%.

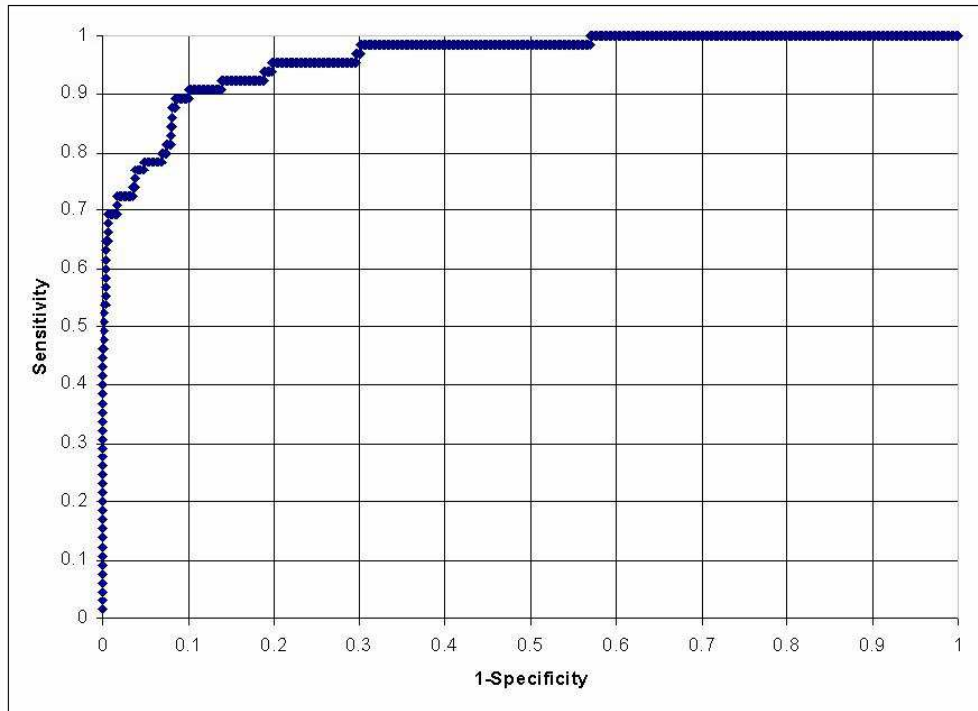


Figure 2: The ROC curve of the classifier on the CORDIS FP6 evaluation sample. The AUC of this curve is 0.96.

The evaluation showed that the applied IST World data integration approach worked well for solving the organisation duplicate detection in one of the datasets, namely the CORDIS FP6 dataset as contained in the IST World repository. The method can therefore be used for large scale data integration tasks. The supervised approach facilitated a semi automated identification of 4000 duplicate pairs, with a very high accuracy and a reasonable recall.

#### 4 IST World Analytic Tools and Examples

The analytic tools are the flagship component of the IST World portal. They perform in real time, and exhibit a high level of complexity in their usage and application. They are highly innovative and interact directly with the data in the IST World repository. Methods for spectral data analysis (Golub et. al. 1970), unsupervised clustering (Grobelnik and Mladenic 2005) and latent semantic indexing combined with multidimensional scaling (Fortuna et. al. 2005) were used for the visualizations. The presentation of real usage examples will demonstrate their power. All examples are

based on the data as of the end of the project. The IST World portal will continue beyond the project. The queries can be repeated publicly, and the diagrams can be created online by free registration with the portal at <http://www.ist-world.org>.

#### 4.1 Competence Diagram Analysis

IST World understands competence as the ability to perform a task. With reference to this definition, IST World competencies are represented by the four entities organisation, person, project, publication (in a broader sense), the relationships between them, and their abilities to perform tasks. The following usage scenario will be supportive for understanding.

**Usage Scenario (1):** We want to investigate the thematic range and goals of the SSA projects funded within the Sixth Framework Programme (FP6) in IST to identify overlaps with the IST World project.

A performed query for the IST SSA projects within FP6 returned 200 results. Those results have been visualized for further analysis as a Competence Diagram in figure 3.

The IST World competence diagram represents those thematic areas that are covered by the 200 projects from the result list (HEALTH, LEGAL, CHANGING, SEMANTIC, ROADMAP, ...) as blue clouds. The red dots are the 200 projects behind these themes. Their position is based on their thematic assignments. Many of the projects are positioned in the very centre and not properly assigned to topics, due to a rather large result space (figure 3).

We changed the result space by selecting only the top 40% and increased the topics to 30, as shown in figure 4.

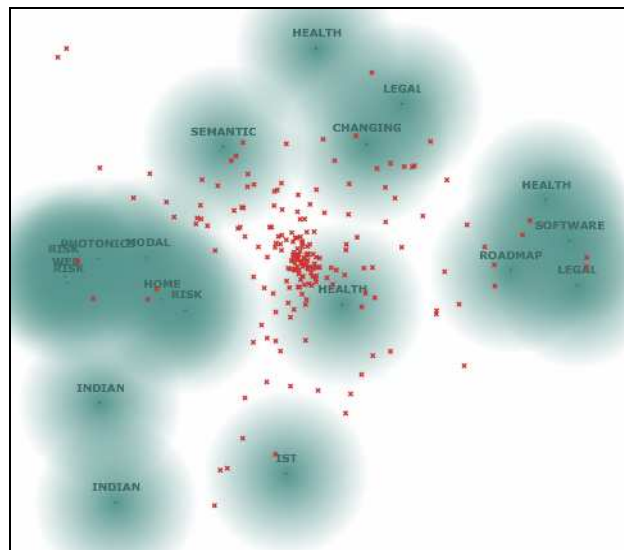


Figure 3: Competence diagram representing project themes and projects associated to them

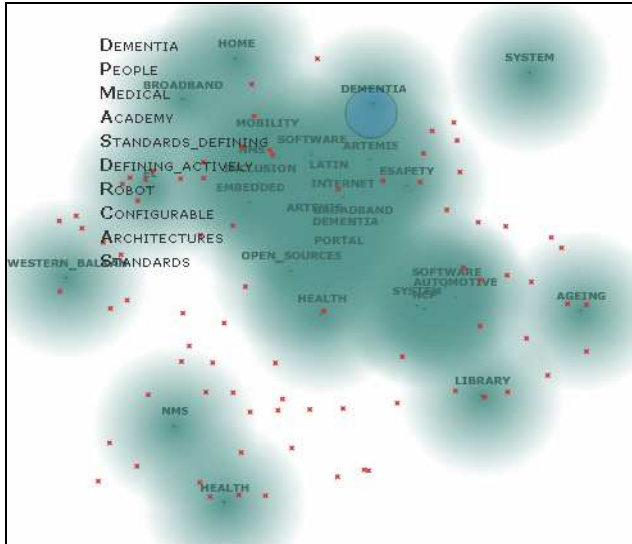


Figure 4: Allocation of the top 40% of the 200 projects within a diagram of 30 topics.

Moving the mouse over the diagram in IST World allows for a detection of goals behind the thematic areas (vertical lists: DEMENTIA, PEOPLE, MEDICAL, STANDARDS ...). The goals represent the space inside the small blue circle. The coverage of the blue circle is dynamic and can be enlarged or reduced to one point. Moving the mouse onto one particular red dot reveals the project behind, as in figure 5.

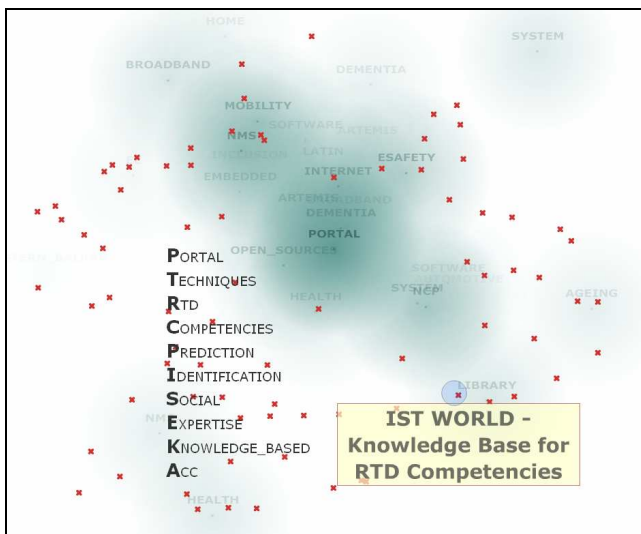


Figure 5: Competence diagram listing the IST World goals and highlighting the IST World thematic range

When pointing to a project in figure 5, the thematic clouds (representing the reduced result space in figure 4) disappear, and only the most relevant topics assigned to the chosen project in figure 5, show up (PORTAL, INTERNET, NMS, NCP). The vertical lists of keywords assigned to particular projects, like for the IST World project in figure 5, (PORTAL, TECHNIQUES, COMPETENCIES, PREDICTION, IDENTIFICATION ...) indicate the goals of that particular project.

Each project in the diagram is linked with the full record in the IST World repository and allows for further analysis. A thorough investigation of the thematic range and goals, and a comparison towards IST World as required in the usage scenario (1) can therefore be completed successfully.

## 4.2 Collaboration Diagram Analysis

Collaboration links in IST World are established between the four entities organisation, person, project, and publication. The IST World collaboration diagrams are based on techniques developed within a system called Project Intelligence (Grobelnik and Mladenec 2003). We continue from the query performed in the previous scenario.

**Usage Scenario (2):** We want to investigate the collaboration between the 200 FP6 IST SSA projects from the previous query, and see, how many projects have joint partners in their consortia, or (to put it differently) how many organisations participated in more than one of these projects.

A Collaboration Diagram has been generated in figure 5 and 6, for the investigation of collaborations as requested in usage scenario (2).

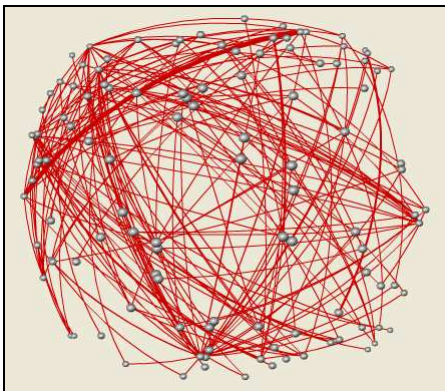


Figure 6: Collaboration Diagram of all projects

The diagram in figure 6 represents the collaboration between the 200 projects, based on participation of organisations in their consortia. The diagram shows that some projects have a stronger linkage.

For accessibility reasons, the complexity of the diagram presented in figure 6 has been reduced, by showing only the top 20% of the collaboration links between the projects in figure 7. The stronger a link, the more joint partners are involved in the linked projects. The NESSI-GRID and NESSI-SOFT project in figure 7 have a strength of 10. This means that they have 10 partner organisations in common.

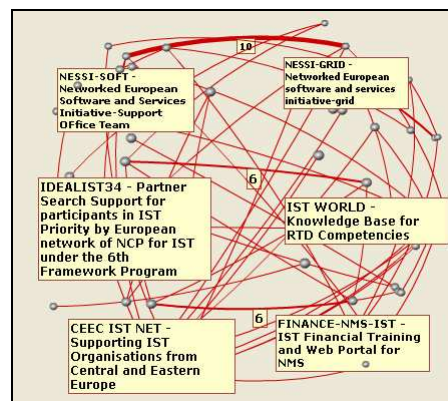


Figure 7: Top 20% of collaborating projects

The request from usage scenario (2) can therefore successfully be completed.

IST World allows for the creation of analytic diagrams over single records. Figure 8 represents the collaboration diagram of the IST World consortium members.

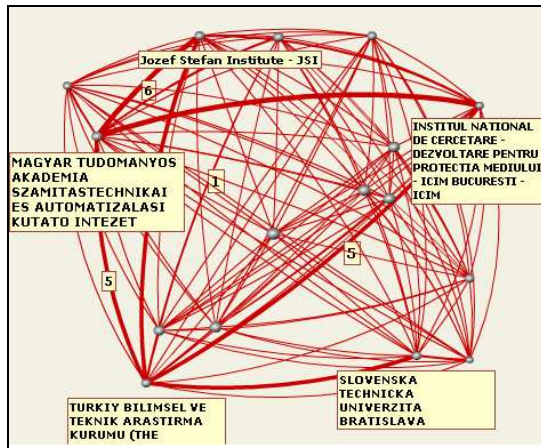


Figure 8: IST World project collaboration diagram representing the IST World consortium members and their collaboration strength

The diagram reveals that several partners within the IST World consortium had collaborated in other projects. A very strong connection can be identified between the Hungarian partner and the Romanian and Slovenian partners, but also between the Turkish and the Hungarian or Slovakian and Romanian partners. For a better readability only some organisation name labels have been activated in figure 7.

### 4.3 Evaluation of Analytic Tools

The functionality of the IST World tools towards the usage scenarios proved very well. The competence diagram provided a concise summary of the thematic areas and the goals of the queried projects can easily be investigated by moving the mouse over the diagram. The collaboration diagram showed the strength of collaboration across projects from participating organisations and revealed information about multiple participations. The given results refer to the data records and their storage within the IST World repository. A quality (content wise) of the results can only be measured by real world knowledge or by comparison with the sources from where the data originated. An extensive analysis of the IST World portal content is far beyond the scope of this paper. To provide insight, we examined some queries, and we compared the results of the competence diagrams of usage scenario (1) in IST World<sup>4</sup> with the results from a human investigation at CORDIS<sup>5</sup> based on project titles and abstracts. The summary of the CORDIS investigation is presented separately in the paragraph IST World Related Activities.

#### 4.3.1 Query Evaluation

Queries were executed in March 2008 in the two datasets IST World and CORDIS. Whereas the public CORDIS dataset is constantly updated, the version of the CORDIS data as collected in the IST World repository dates back to January 2007. From 208 projects (query 1 in table 4) in CORDIS only 176 have been found in the IST World database by acronym. We therefore assume that at least 32 missing projects had not been recorded at CORDIS by the time the data were crawled, in January 2007. Table 4 shows the experiments with different queries (1-5) in the two databases.

During queries we discovered many inconsistencies in the CORDIS dataset with strings like FP6 (out of 80 relevant records 30 did not contain the string), SSA (out of 208 relevant re-

<sup>4</sup> IST World Portal: <http://www.ist-world.org/>

<sup>5</sup> CORDIS Search: <http://cordis.europa.eu/en/home.html>

cords 15 did not contain the string), Specific Support Action (out of 208 relevant records 15 did not contain the string), or Dates (the year of the call to which the projects belong were not consistently recorded). From the 208 given results (query 1 at CORDIS) 22 were classified as Coordination, Coordination Action, Specific Targeted Action, Integrated Project, and others.

Table 4: Query results from IST World and CORDIS

	IST World (crawled Data from CORDIS)	CORDIS
Investigation Date	March 2008	March 2008
Last Updated	January 2007	constantly updated
Query 1: Specific Support Action IST FP6	64	208
Query 2: Specific Support Action IST	377	1178
Query 3: Specific Support Action FP6	185	1507
Query 4: Specific Support Action	1554	2012
Query 5: Project Keywords: Specific Support Action Programme: IST, StartDate: After 01/01/2002	200	not checked

#### 4.3.2 Content Evaluation

Based on the results from table 4 we decided that query 5 in IST World represents most appropriately the knowledge base for both, usage scenario (1), and evaluation and validation of its content towards the results of query 1 in CORDIS. A 1:1 comparison of the results between the two datasets IST World and CORDIS is very difficult first, due to inconsistencies within records, second, due to a difference of the timestamps, and third, due to the different query mechanisms: the 200 results from query 5 in IST World miss about 80 records when compared with the 208 CORDIS results from query 1, based on their acronyms. The results from the human investigation of IST World related activities at CORDIS (see section 5), namely 40 projects in the wider range, are only a part of the result space in the IST World portal.

An evaluation of the IST World content by comparing the results from usage scenario (1) in IST World with the results from titles and acronyms at CORDIS (table 5), allowed for the conclusion (as expected) that the competence diagram in figure 3 represents the real world topics from table 5 only partly. A review of the 80 missing records showed, that their topics in areas like INDIA, RUSSIA, CHINA, SME overlap with those topics that were missing in the diagram of figure 3, and that their missing is therefore correct.

A further investigation of the query results confirmed that the diagram in figure 3 represents the 200 results behind the query. However, a higher number of topics (currently only 30 for 200 projects) could be of help for a more detailed investigation. The configuration of the competence diagram is limited to 30 topics per diagram. Some inconsistencies with the position and with the highlighting of topics for individual records (IST World in the vicinity of LIBRARY in figure 3) need more investigation. The comparison of the project goals in the diagram of figure 5 (vertical list of keywords in IST World) with the planned results, services and goals in table 5, shows substantial overlap and is therefore of high value.

## 5 IST World Related Activities

IST World was a Specific Support Action (SSA) project, funded within the 6<sup>th</sup> Framework Programme of the European Commission, in the thematic area of IST. According to the public CORDIS database as of March 2008, a total of 1178 projects have been funded within the Sixth Framework Programme (2002 – 2006) in IST, and about 208 of those projects were Specific Support Action projects. A human investigation of those 208 projects in the CORDIS database<sup>6</sup> by titles and abstracts revealed that about 40 of those projects are related to IST World in the wider range of their goals. A rough summary of the thematic range, planned results, and the goals of those 40 projects – based on their titles and abstracts – related to IST World is provided in table 5.

*Table 5: Comparison of IST World with related FP6 IST SSA projects*

Coverage	IST World	Other FP6-IST SSA projects
<b>Geographic Area</b>	NMSs, ACCs	New Member States (NMS), Associate Candidate Countries (ACC), Central Eastern European Countries (CEEC), South Eastern Europe (SEE), Western Balkan (WB), Africa, Asia, Balkan, Canada, China, India, Latin America, Mediterranean, Russia
<b>Thematic Area</b>	IST in general	eInclusion, eHealth, Finance, Mobile Services, eBusiness / eWork, ICT, GRID
<b>Target Group</b>	RTD actors in IST	RTD actors in IST
<b>(Planned) Results</b>	Knowledge Base for RTD competencies with innovative functionalities.	Raising Awareness, Increasing Participation, Concrete Proposal Submissions / Involvement, Improved International Cooperation, Extension of Networks / Collaboration, Monitoring and Addressing Synergies, Identification of Players / Excellence Building (Extend) the Knowledge Space / Units
<b>Services</b>	Public Web Portal, Advanced Analytic Tools, Community Building	Partner Search , Databases of Experts / Directories of Organisations, Proposal Submission Assistance,

<sup>6</sup> CORDIS Search: <http://cordis.europa.eu/en/home.html> as queried in March 2008.

		Counseling, Organizing (Local) Events, News Networking / Community Building
<b>Goals</b>	Knowledge Base Setup of Analytic Tools, European Research Infrastructure (based on CERIF), European Research Community, Participation, Inclusion, Innovation	Participation, Quality Control, International Cooperation, Inclusion, Research Infrastructure / Network, Development of Standards, Innovation, European Knowledge Space

## 6 Summary and Conclusion

The IST World project applied methods and tools from previous projects (Erbach et. al. 2005) and set up the IST World portal (<http://www.ist-world.org/>). Data from many sources have been integrated. The described data integration methodology could be further developed by employing a transductive instead of an inductive learning function. As all test data were already known in advance, they could be used to generate an even better classification model. The transductive SVM machine introduced by (Vapnik 1999) could be used for these purposes. The feature generation step could be improved using the ontological information (e.g. collaboration of organisations) to compare two records. As the feature generation step needs to be redesigned for every particular data integration problem at hand, the transfer learning methods (West et. al. 2007) might enable the reuse of one integration model for several instances of integration problems. The recall of the automated classification methods could be substantially improved by manually evaluating those record pairs that failed to meet the classification threshold to be classified into the M set. Regular data updates require the repetition of the data cleaning effort and are extremely difficult in large datasets. The semi-automated data cleaning approach was applied for organisation records only. Means to maintain person records and to merge them with existing records have been provided via the community portal. The power of the IST World tools for large datasets has been successfully demonstrated. The evaluation of the analytic tools by some examples showed how much analytics depend on the data behind, and that inconsistencies in data make a substantial impact on the results. For a quality evaluation of large databases domain knowledge is needed, which is not easy at hand. The maintenance of large datasets is extremely difficult, especially when collected from many sources with heterogeneous structure. The employment of a proper data model can substantially reduce the number of duplicate records and the maintenance efforts. The CERIF datamodel supported an efficient setup of the IST World repository and enabled the smooth data exchange. Without a datamodel and quality data management it is hard to perform reliable analysis.

## 7 Acknowledgements

The work reported here was supported by the European Commission through the project IST World (contract FP6-2004-IST-3 - 015823).

## 8 References

- Asserson, A.; Jeffery, K.G.; Lopatenko, A. (2002): CERIF: Past, Present and Future: An Overview. In Proceedings: Gaining Insight from Research Information, 6th International Conference on Current Research Information Systems, Kassel, Germany.
- Brank, J. and Leskovec, J. (2003): Download Estimation on KDD cup 2003, in KDD Cup 2003, eds., Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg, 2003.
- Erbach, G.; Jermol, M.; Jörg, B.; Uszkoreit, H.; Grobelnik, M. (2005): Network Approaches to Current Research Information Systems. In: Innovation and the Knowledge Economy: Issues, Applications, Case Studies, Volume 2., Paul Cunningham and Miriam Cunningham (Eds), 2005. IOS Press Amsterdam, ISBN: 1-58603-563-0.
- Erbach, G. (2006): Data-centric view in e-Science information systems. In: Data Science Journal (Vol. 5 (2006) pp.219-222), ONLINE ISSN: 1683-1470.
- Ferlez, J. (2007): Metode analize podatkov o raziskovalni dejavnosti na primeru aplikacije IST World. Masters Thesis at the University of Ljubljana, Faculty of Computer and Information Science, October 2007.
- Ferlez, J.; Jörg, B. (2007): D8.1 Component for automated content acquisition and database integration. IST World deliverable within WP8 – Portal Maintenance and Sustainability. April 2007.
- Fortuna, B.; Mladenič, D.; Grobelnik, M. (2005): Visualization of text document corpus. Informatica Journal, 29(4):497–502, 2005.
- Grabczewski E. and Jörg, B. (2005): Base Set of Data. IST World deliverable within WP3 – Data Acquisition. October 2005.
- Golub, G. H.; Reinsch, C. (1970): Singular value decomposition and least squares solutions, Numerische Mathematic, Volume 14, Number 5, April 1970.
- Grobelnik, M. and Mladenic, D (2003): Analysis of a database of research projects using text mining and link analysis. In Mladenic, D., Lavrac, N., Bohanec, M. and Moyle, S. (eds), Data Mining and Decision Support Integration and Collaboration, Kluwer, Dordrecht, 2003, pp. 157-166.
- Grobelnik, M. and Mladenič, D. (2005). Contexter - a system for visualization of large collections of novel stories. V: 29th Annual Conference of the German Classification Society, March 9-11, 2005, Magdeburg. From data and information analysis to knowledge engineering. Magdeburg: Otto-von-Guericke-University, page 269, 2005
- Jeffery, K.G., Asserson, A. & Lopatenko, A. S. (2002) Comparative Study of Metadata for Scientific Information: The place of CERIF in CRISs and Scientific Repositories. Gaining Insight from Research Information, 6th International Conference on Current Research Information Systems, Kassel, Germany.
- Jörg B.; Jeffery, K.G.; Asserson, A.; van Grootel, G.; Price, A.; Rasmussen, H.; Vestam, T. (2008): *CERIF2008 1.1 Full Data Model (FDM) – Model Introduction and Specification*. euroCRIS, October 2008. *To appear*.

- Jörg, B; Jermol, M.; Uszkoreit, H.; Grobelnik, M.; Ferlež, J.; Kirkyakov, A. (2006): Analytic Information Services for the European Research Area. In: Exploiting the Knowledge Economy: Issues, Applications and Case Studies, PT.1-2, Paul Cunningham and Miriam Cunninghamf (Eds.), 2006. IOS Press, Amsterdam. IOS Press Amsterdam, ISBN 1-58603-682-3.
- Levenshtein, V. I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, pp. 707–710, 1966.
- Lewis, D. (1998): Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. Proceedings of ECML-98, 10th European Conference on Machine Learning: 4-15, Chemnitz, DE: Springer Verlag, Heidelberg, DE.
- Lodhi, C.; Sunders, J.; Shawe-Taylor, N. Cristianini, C. Watkins, (2002): Text classification using string kernels, The Journal of Machine Learning Research, Volume 2, pp 419-444, March 2002.
- Spackman, K. A. (1989): Signal detection theory: Valuable tools for evaluating inductive learning". Proceedings of the Sixth International Workshop on Machine Learning: 160–163, San Mateo, CA: Morgan Kaufman.
- Thévignot, C. (2000): The redesigned CORDIS web service contributes to the Commission's eEurope Initiative. Conference on European Research Information Systems CRIS2000, Helsinki 25-27 May 2000.
- Tong, S.; Koller, D. (2002): Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research, Volume 2, pp. 45-66. March, 2002.
- Winkler, W. E. (2006): Overview of record linkage and current research directions, Technical Report RRS2006/02, US Bureau of the Census.
- Vapnik, V. (1995): The Nature of Statistical Learning Theory. Springer-Verlag, 1995.
- Vapnik, V. (1999): The Nature of Statistical Learning Theory. Springer-Verlag, 1999. ISBN 0-387-98780-0.
- West, J, Ventury D., Warnick S. (2007): Spring Research Presentation: A Theoretical Foundation for Inductive Transfer (Abstract Only). Brigham Young University, College of Physical and Mathematical Sciences. 2007. Retrieved on 2007-08-05.
- Zobel, J.; Moffat, A.; Ramamohanarao (1998): Inverted files versus signature files for text indexing. ACM Transactions on Database Systems (TODS), Volume 23, No. 4, pp. 453 - 490, 1998.

## 9 Contact Information

Jörg Brigitte  
 Stuhlsatzenhausweg 3  
 66123 Saarbrücken  
 e-mail: [brigitte.joerg@dfki.de](mailto:brigitte.joerg@dfki.de)  
<http://www.dfki.de/~brigitte>