



Knowledge Base for RTD Competencies in IST



Deliverable D1.2

Data Model for Knowledge Organisation

Author(s):	Brigitte Jörg, Jure Ferlež, Edward Grabczewski
Identifier:	D1.2
Work package:	WP1 Data Model for Knowledge Organisation
Lead Partner:	Deutsch. Forsch. für Künstliche Intelligenz (DFKI)
Partner(s):	Institute Jozef Stefan (JSI) Council for the Central Lab. of the R. Council (CCLRC) Ontotext Lab, Sirma Group (ONT)
State of document:	final
Version:	1.0
Dissemination Level:	Public
Date:	2005-09-15

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	MT
Technical University of Silesia	SUT	Poland
National Institute for Research and Development in Informatics	ICI	Romania
Silesian University of Technology	STUBA	Slovakia
TUBITAK	TUB	Turkey
CCLRC	CCLRC	United Kingdom

Abstract

The central data structure of the IST World repository builds on the CERIF model for Current Research Information Systems. In deliverable D1.1 we took the CERIF data model and extended it for the purposes of collecting data in the IST World project. With this document we consider the problem of classification of topic areas within the IST World repository and therefore propose further extensions to the already extended CERIF data model.

Since the IST World repository will be fed from a variety of sources, we need to consider how we intend to classify data from those heterogeneous sources. There are three main sources we will examine. They are as follows:

- **LT World:** Ontologically structured knowledge base representing the field of Language Technology
- **PI / CORDIS:** FP5 and FP6 projects organised within IST topics
- **Computed Classification Models:** Resulting from automated content and structure discovery

LT World entities co-operate within the range of Language Technologies, CORDIS projects appear in the context of IST, computed classification models from content and structure discovery are built automatically.

Compared with previous work in deliverable 1.1 where the pragmatic focus concerned a representation of entities in the IST World context, this deliverable deals with an integration of classification models and thus an identification of further CERIF extensions resulting from integration.

In this way, we will have demonstrated how the standard CERIF data model can be used and extended to support the IST World project. This is an important proof-of-concept and moreover a validation of the CERIF standard.

Content Table

Knowledge Base for RTD Competencies in IST	1
IST World Consortium	2
Abstract	3
Content Table	4
1. Introduction.....	5
2. LT World	6
2.1. Entities	6
2.2. Mappings	7
2.3. Classification	9
3. Project Intelligence / CORDIS.....	10
3.1. Entities	10
3.2. Mappings	10
3.3. CORDIS FP6 Keyword Hierarchy	12
4. Computed Classification Models	13
5. CERIF Classification System	14
5.1. CERIF Classification Model	14
5.2. CERIF Classification Scheme	14
5.3. CERIF Methodology for Classification Integration.....	14
6. IST World Model Extensions	15
6.1. CERIF Model Extensions	15
6.2. Classification Scheme	16
6.3. Methodology for Classification Integration.....	16
7. Conclusion.....	17
8. Bibliography	18

1. Introduction

The central data structure of the IST World repository builds on the CERIF model for Current Research Information Systems. In deliverable D1.1 we took the CERIF data model and extended it for the purposes of collecting data in the IST World project. With this document we consider the problem of classification of topic areas within the IST World repository and therefore propose further extensions to the already extended CERIF data model.

Since the IST World repository will be fed from a variety of sources, we need to consider how we intend to classify data from those heterogeneous sources. There are three main sources we will examine. They are as follows:

- **LT World:** Ontologically structured knowledge base representing the field of Language Technology
- **PI / CORDIS:** FP5 and FP6 projects organised within IST topics
- **Computed Classification Models:** Resulting from automated content and structure discovery

LT World entities co-operate within the range of Language Technologies, CORDIS projects appear in the context of IST, computed classification models from content and structure discovery are built automatically.

Compared with previous work in deliverable 1.1 where the pragmatic focus concerned a representation of entities in the IST World context, this deliverable deals with an integration of classification models and thus an identification of further CERIF extensions resulting from integration. In this way, we will have demonstrated how the standard CERIF data model can be used and extended to support the IST World project. This is an important proof-of-concept and moreover a validation of the CERIF standard.

The document will be structured as follows:

- LT World Model
- PI / CORDIS Model
- Computed Classification Models
- CERIF Classification System
- IST World Model Extensions

We briefly introduce LT World and PI / CORDIS models and their mappings for IST World integration. LT World and PI / CORDIS models are pre-defined schemes that are quite similar to the current CERIF-based central data model of the IST World repository. Beyond pre-structured content, which will later be collected from all participating partners, the IST World project substantially relies on automated data collections from the Web. Automated content and structure discovery results in computed classification models and mappings. They have to be considered when thinking of model integration and model extensions. An overview of the CERIF classification system and its components will be given and finally, a description of necessary model extensions will be specified.

2. LT World

The Web-based information system LT World (<http://www.lt-world.org/>) is a comprehensive knowledge portal for the large field of language technology. LT World provides information on many aspects of research and technology such as methods, people, projects, products, events, IPR etc. Its overall architecture is ontology-based. Many fields of science and technology exhibit a multidimensional structure and ontologies provide a solid and meaningful connection between different types of content. Formal ontologies consist of classes (so called concepts) and their relations in a multiple inheritance order that goes beyond taxonomical "is-a" relations.

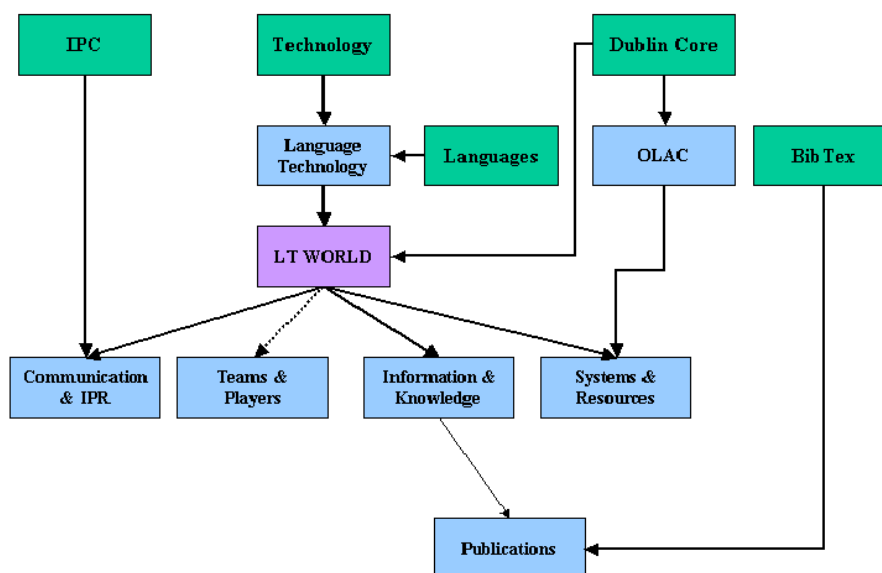


Fig 1: Conceptual LT World structure indicating multiple inheritance

The LT World ontology initially served as an underlying structure for conceptual design and for the classification of data. It was then extended for representing the complete system architecture and it is now in the process of being utilized for maintenance, presentation and interoperability [1].

2.1. Entities

LT World entities are grouped as follows (See also figure 1):

- Person, Project, Organisation (*Teams & Players*)
- Product, System, Repository (*Systems & Resources*)
- News, Event, Patent (*Communication & IPR*)
- Technology (*Information & Knowledge*)

D1.2: Data Model for Knowledge Organisation

When integrating the LT World model into the IST World central data structure, not all of the LT World entities are relevant. We only concentrate on those LT World entities that are comparable to IST World entities, Person, Project, Organisation. The concerned LT World entities are very similar to CERIF-based IST World entities as can be seen from the following mapping definitions.

2.2. Mappings

In deliverable D3.1 we specified XML APIs for an import of IST World entities: Person, OrgUnit, Project, and ResultPublication. Ontology-based LT World entities are stored as XML objects. For the mappings we had to identify matching elements and the roles between the two models. The mapping definitions are therefore based on API specifications of D3.1.

Mappings for Entity Person	
LT World Person	IST World Person
rdf:ID	Person[id]
personFirstname	firstNames
personLastname	familyNames
homepageURL	URI
personTitle	academicTitle
subaffiliatedWith	orgUnitId[role=subaffiliatedWith]
affiliatedWith	orgUnitId[role=affiliatedWith]
personDiscipline	personResearchInterest
personEmail	email
participatedIn	projectId[role=participatedIn]
hasDeveloped	resultProductId
dc:keyword	keywords

Mappings for Entity Organisation	
LT World Organisation	IST World OrgUnit
rdf:ID	orgUnit[id]
organisationNameAbbreviation	acronym
organisationName	orgUnitName
homepageURL	URI
contact	email
location	cityTown
iso3166	countryCode
partOf	orgUnitId[role=partOf]

D1.2: Data Model for Knowledge Organisation

hasPart	orgUnitId[role=hasPart]
headedBy	personId[role=headedBy]
hasDeveloped	resultProductId
dc:keyword	Keywords

Mappings for Entity Project	
LT World Project	IST World Project
rdf:ID	project[id]
projectName	projectTitle
projectTheme	projectAbstract
homepageURL	URI
dateStart	startDate
dateEnd	endDate
contact	email
partOf	projectId[role=partOf]
hasPart	projectId[role=hasPart]
coordinatedBy	personId[role=coordinatedBy]
investigatedBy	personId[role=investigatedBy]
organizedBy	orgUnitId[role=partner]
fundedBy	fundingProgrammeId[role=programmeName]
hasParticipant	personId[role=hasParticipant]
resultedIn	resultProductId
dc:keyword	keywords
lt:technologicalMethod	classificationId

Most of the LT World elements can be mapped to IST World elements. Not only the elements of LT World entities overlap with IST World concepts, but also roles and relations. LT World entities will be imported into the CERIF-based IST World repository according to specification of deliverable D1.1 with support of the XML API definitions of deliverable 3.1.

Extensions that resulted from the mapping preparations of LT World entities to IST World entities are needed for an integration of classification schemes, what becomes clearer from the following sections.

2.3. Classification

A management of the LT World entities is supported by an ontology of language technologies and related methods. Each technology is represented within the ontology as a class. Classes are classified by subclasses, subclasses belong to several super classes as can be seen from the following ontology extract in figure 2:

- Language Technology
 - Information Extraction
 - Answer Extraction
 - Named Entity Recognition
 - Text Data Mining
 - Information Retrieval
 - Categorization
 - Clustering
 - Topic Detection
 - Knowledge Representation and Discovery
 - Automatic Hyperlinking
 - Ontologies
 - Semantic Web
 - ...
 - Authoring Tools
 - Automatic Hyperlinking
 - ...
 - ...
 - Evaluation

Fig 2: LT World ontology extract of class hierarchy

To make use of a modular LT World ontology within the IST World data repository, its hierarchical structure has to be transformed into a relational structure.

The following CERIF extensions resulting from LT World classification integration were therefore identified:

- There is a need for a general methodology of schema representation, as most classification schemata are represented as hierarchies and the base CERIF model is organized in relational form.

A more detailed definition of necessary CERIF extensions can be found in the section for IST World Model Extensions.

3. Project Intelligence / CORDIS

CORDIS is an information service devoted to European research and development (R&D) and innovation activities. It is accessible from the internet (<http://www.cordis.lu>). The main aims of CORDIS are: To facilitate participation in European research and innovation activities; to improve exploitation of research results with an emphasis on sectors crucial to Europe's competitiveness; and to promote the diffusion of knowledge fostering the innovation performance of enterprises and the societal acceptance of new technology.

3.1. Entities

The Project Intelligence (PI) service is provided by Jozef Stefan Institute. It uses advanced data mining techniques for determining relationships among projects, organizations and countries in 5th and 6th Framework Program [2][3]. The database holds data about EU funded projects. The database was in part automatically built from the data made publicly available by the CORDIS information system and in part acquired internal data stores of the EC. The latest version of the Project Intelligence service provides data and analysis of the 6th Framework Program.

3.2. Mappings

The CORDIS and EC data in the PI data store was imported to the CERIF database of the IST World portal, to meet the project objective to quickly transfer the functionality of the PI service to the IST World portal in order to gain early interest in the new portal by the European research community. Therefore a mapping between different data models of Project Intelligence and IST World data stores had to be produced.

PI backend data store is a simple one. It employs object serialization to enable fast data loading and fast execution of predefined analysis queries. The underlying data model follows this scheme. It can be thought of as a collection of the three entities: Country, Organization and Project. Each entity is described with additional fields like: name, description, budget, etc. The relations between the entities are also simple and represent hierarchical structure: projects aggregate organizations and organizations aggregate countries, this makes projects and countries implicitly related.

The IST World Portal backend data store is a relational database management system based on the CERIF data model specialized for modeling the research information data. It presents three basic entities: Project, OrgUnit and Person; with additional entities like: Country, ResultPublication, etc. The model is carefully designed in order to provide solid ground for storing information on research activities.

As an example a straightforward mapping of the PI Project entity to the CERIF Project entity and related properties is given below:

PI Project	IST World
Acronym	<code>ProjectTitle.Title</code>
Subject	<code>ProjectKeywords.keywords</code>
Title	<code>ProjectTitle.Title</code>
Duration	<code>Project.StartDate, Project.EndDate</code>

D1.2: Data Model for Knowledge Organisation

Description	ProjectAbstract .Abstract
Reference	Project .ProjectId
Programme	Project_FundingProgramme .FundingProgrammeId, Project_FundingProgrammeRole .Role, FundingProgramme .FundingProgrammeId FundingProgrammeName .Name
Sub Programme	Project_FundingProgramme .FundingProgrammeId, Project_FundingProgrammeRole .Role, FundingProgramme .FundingProgrammeId FundingProgrammeName .Name
Instrument	ProjectNotes .Notes
Value	ProjectNotes .Notes
Funding	FundingProgramme .Budget
Prime Contractor	Project_OrgUnitRole .Role, Project_OrgUnit .OrgUnitId
Partners	Project_OrgUnitRole .Role, Project_OrgUnit .OrgUnitId

The simplicity of the PI data model and the specialization of the CERIF data model make the import of PI data into IST World data store a simple one. As both data models are relational, no XML API is needed. The PI Project entity is directly mapped to CERIF Project, and properties like projectName, projectDescription, etc. The PI Organization entity is mapped to CERIF OrgUnit entity, and properties like orgUnitName, orgUnit_Location, etc. The PI Country entity was not imported into the IST World data store. The instances of the CERIF Country entity are general to all research data sources and were therefore treated separately.

The mappings were realized by exporting the serialized PI data object into a XML format data file, which was then processed by the data base import program of IST World. This program was automatically generated by the ALTOVA MAPForce© application to enforce the defined simple mappings.

3.3. CORDIS FP6 Keyword Hierarchy

Project Intelligence (PI) data do not include any classification hierarchy. However CORDIS FP6 projects are organized according to a hierarchical list of keywords. A CORDIS classification scheme is publicly available for navigation at: <http://fp6.cordis.lu/fp6/fp6keywords.cfm>. Each of the classes has several subclasses. The following example extract demonstrates the range of FP6 projects:

- **Health Sciences**
 - Medical Sciences
 - Neurosciences
- **Humanities**
 - Arts
 - History
 - Information Science
 - Language Sciences
 - Literature
 - Philosophy
 - Religious Sciences
- **Natural Sciences**
 - Agricultural Sciences
 - Biological Sciences
 - ...
- **Physical Sciences**
- **Social Sciences**
 - ...
- **Technological Sciences**

To make use of the CORDIS class hierarchy within the IST World data repository, the modular structure has to be transformed into a relational structure in a similar way as with LT World or any classification schema as such.

The following CERIF extensions resulting from CORDIS classification integration were therefore identified:

- If more than one classification model will be applied for content organization, there is a need for the definition of an identifier indicating the name of the classification scheme.

A more detailed definition of necessary CERIF extensions can be found in the section for IST World Model Extensions.

4. Computed Classification Models

Beyond pre-structured contents (like LT World and CORDIS), which will be collected from all participating partners, the IST World project substantially relies on automated data collections. Automatic content and structure discovery results in computed classification schemes. The following tasks have to be considered when thinking of model integration and model extensions:

- *Classification schema generation*; a critical task performed during the information integration process. Its goal is to handle the variety of data semantics by constructing a hierarchy of the covered data.
- *Classification schema mapping*; a process of translating semantics of one schema to another. Automated classification schema mappings deal with computer aided mapping generation employing advanced text processing procedures.

In the IST World project we plan to reuse the SEKT approach in particular in the field of knowledge discovery techniques where the idea is to reuse and adapt Text-Garden software environment for automatic handling large textual data corpora and large social networks. The key issue is to prepare the data in the form (conceptually and technically), which will enable efficient analytic work in the second phase of the IST World project - current developments on IST worlds (first 6 months deliverables) go all in this direction.

A general methodology for IST World schema integration will be defined as part of the Portal architecture in deliverable 2.3. The degree of automated processing involved in mapping generation also has to be discussed. It is constrained by the quality of automated mappings. A best trade-off between manual and automated mapping construction therefore has to be discussed, defined and realized.

5. CERIF Classification System

The central structuring concept of traditional scientific libraries is a classification scheme reflecting the system of the represented disciplines. Usually one of the few widespread library classification systems such as the Dewey Decimal Classification or the Classification of the Library of Congress are employed if a library covers several disciplines or all academic fields. In libraries dedicated to a special subject, often specialized classification schemes are applied for structuring the relevant discipline. Such systems and therefore also the classification schemes mostly are hierarchies that can be represented as trees. Orthogonal dimensions can be reflected in the keyword index.

5.1. CERIF Classification Model

The current CERIF model contains a basic means for data classification. It consists of the following three tables for classification:

- *Classification* – used to classify the entities in the entity representation part of CERIF.
- *ClassificationDescription* – used to describe the language-dependent description of the classification identifiers
- *ClassificationScheme* – used to save a pointer to the external classification schema definition

5.2. CERIF Classification Scheme

The current CERIF model does not make use of an explicit classification scheme as such. The CERIF documentation refers to an external classification scheme¹. The referred scheme however is not anymore maintained and out of date for the IST World context. The CERIF task-group is aware of this problem and works on this topic.

5.3. CERIF Methodology for Classification Integration

The current CERIF model does not foresee a methodology to maintain, integrate, and map classification schemes as such. The IST World data repository however will have to integrate data from 15 different countries based on very different classification schemes and moreover has to manage computed classification models. Therefore a methodology for integrating multiple classification schemes is needed. The CERIF task-group is also aware of a missing methodology and works on a general solution.

Extensions concerned with classifications schemes as such and moreover extensions for a general methodology of classification integration into the CERIF model are specified in the IST World Model Extensions part of this document.

¹ The CERIF documentation refers to the Beat Sottas classification:
<http://www.ub.uib.no/avdeling/fdok/cris/taskgroups/lists.htm>

6. IST World Model Extensions

The LT World and PI/CORDIS data models have been analyzed. Based upon the results of their analysis and with the aim of identifying further extensions for the IST World central data structure, we conceptually separated the central data model into:

- Entity Representation
- Classification Representation

An integration of LT World and CORDIS entities can be realized without major model extensions as demonstrated with LT World and PI / CORDIS mapping definitions. For schema or classification integration and related mappings however, the following extensions are of need:

- CERIF model extensions for classification
- A general methodology for integrating multiple classification schemes

6.1. CERIF Model Extensions

In order to provide means for storing mappings between different schemes the CERIF database model has to be further extended. Proposed extension tables are:

- *Classification_Classification*
- *Classification_ClassificationRole*.

These tables will provide means to store mappings from an identifier in one schema to an identifier in another schema. The detailed definition of the extension tables is given below:

- *Classification_Classification* has columns:
 - ClassificationId1 NCHAR(32)
 - ClassSchemeId1 NCHAR(32)
 - ClassificationId2 NCHAR(32)
 - ClassSchemeId2 NCHAR(32)
 - Role NCHAR(32)
 - StartDate TIMESTAMP
 - EndDate TIMESTAMP
- *Classification_ClassificationRole* has columns:
 - Role NCHAR(32)
 - RoleFull NCHAR(64)

The more following requirements are to be met:

- Provide support for creation of mappings between different schemata; either computer aided support for manual creation of mappings or provide automatically computed mappings.
- Provide means for a transparent use of these mappings in the context of the functionality of the IST World information service

6.2. Classification Scheme

An idea is to start with the classification scheme of the Open Directory Project dmoz (<http://dmoz.org/Science>), and then refine it to the degree, which serves the IST World purpose. Since there is already software available at Jozef Stefan Institute to classify into dmoz, we could have a working system pretty fast.

6.3. Methodology for Classification Integration

The current CERIF model does not foresee an explicit methodology to maintain, integrate, and map classification schemes as such. The IST World data repository however will have to integrate data from 15 different countries based on very different classification schemes and moreover has to manage Web-extracted data that are organized according to automatically computed classification models.

Classification schemes have to be maintained individually, ongoing changes have to be implemented and multiple schemes have to be integrated. In order to meet this goal, we will develop a methodology for classification integration into the IST World central data structure as part of the portal architecture in deliverable 2.3.

The following requirements are to be specified and discussed by then:

1. Define the one classification scheme for IST World.
2. Define the process of producing mappings from any classification scheme to the one general scheme. Either computer-aided support for manual creation of mappings or provide automatically computed mappings.
3. Define the means of a transparent use of these mappings in the context of an IST World functionality.

7. Conclusion

The current CERIF model contains a basic means for data classification but it does not contain any explicit classification scheme as such. The CERIF documentation refers to external schemes, referred schemes however are not anymore maintained and out of date for the IST World context. We have to find an appropriate classification scheme that can be used in the IST World context and be further extended according to IST World needs. An idea is to start with the classification from the Open Directory Project dmoz, and then refine it to the degree, which serves the IST World purpose.

The current CERIF model does not foresee an explicit methodology to maintain, integrate, and map classification schemes as such. The IST World data repository however will have to integrate data from 15 different countries based on very different classification schemes and moreover has to manage Web-extracted data that are organized according to automatically computed classification models. A general methodology for multiple schemes integration is needed, that is on the one hand related to the CERIF model, but on the other hand also independent itself and as such part of the IST World portal architecture.

We will develop a general methodology for multiple schema integration in the IST World context in co-operation with the CERIF task-group as part of deliverable 2.3 Portal Architecture Specification.

Next steps will be discussed in a next meeting among the technical partners.

8. Bibliography

- [1] 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 der DGI, Frankfurt, 2005. http://www.dfki.de/lt/publications_show.php?id=715
- [2] Grobelnik, M., Mladenic, D., (2002) Analysis of IT projects funded by the European Commission in 5FP. Technical Report IJS-DP 8678, J Stefan Institute, Ljubljana, Slovenija, Nov. 2002.
- [3] Grobelnik, M., Mladenic, D., (2002) Approaching Analysis of EU IST Projects Database, International Conference on Information and Intelligent Systems. IIS-2002, Varazdin, Croatia, Sep 2002.