



Knowledge Base for RTD Competencies in IST



Deliverable D1.1

Definition of Central Data Structure

Author(s):	Atanas Kiryakov, Edward Grabczewski, Jure Ferlež, Hans Uszkoreit, Brigitte Jörg
Identifier:	D1.1
Work package:	WP1 Data Model for Knowledge Organization
Lead Partner:	Ontotext Lab, Sirma Group (ONT)
Partner(s):	Deutsch. Forsch. für Künstliche Intelligenz (DFKI) Institut Jozef Stefan (JSI) Council for the Central Lab. of the R. Council (CCLRC)
State of document:	final
Version:	1.0
Dissemination Level:	Public
Date:	2005-07-26

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 goal of this deliverable is to outline the data model determining the basic structure of the data repository of IST World, its central database hosting the majority of the data and allowing for their analysis, augmentation, querying and presentation. The major objective at this level is the modeling of the agents of interest (People and Organizations), which are interconnected in a sort of social network. It shall also be possible to model the contexts in which the agents cooperate with each other, e.g. projects, publications, events – these co-occurrences are used as evidence for the structure of the social network. Finally, various classifications of the entities of interest will be possible (by location, skills, research areas, etc.) The network will allow for access to the data at different levels of generality, e.g. countries, cities, organizations, departments, and individuals.

This data model will be extended in D1.2 to cover all sorts of information relevant to the so-called current research information systems (CRIS) in the necessary level of detail. It will be extended further, in deliverable D1.3, to allow for the comprehensive representation of expertise necessary for proper clustering and trend analysis, as well as for matching in the process of partner search.

The IST World conceptual model needs to combine a pragmatic data-oriented view with a more far reaching conceptual view. We need a basic version of the portal and thus an operational datastore very early in the project. At the same time we shall assure that deeper semantic analysis is possible at a later phase. Therefore, we had decided to start with a combination of two data models:

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

During the initial phase of the project, the portal will directly use only the relational model; the basic datastore will be a conventional RDBMS. The conceptual model will not be formally involved in the portal operations, i.e. there will be no component such as a semantic repository or reasoner in the architecture. The conceptual model (the ontology) will be used as a design guidance and ground for development of a proper expertise modelling schema (deliverable D1.3). The ontology will be involved later on together with a semantic repository which properly integrates with the RDBMS.

The main part of this deliverable presents: (i) an overview of CERIF; (ii) analysis of the necessary extensions; and (iii) the conceptual model RENO – Research Networks Ontology.

Content Table

IST World Consortium	2
Abstract	3
Content Table	4
1. Introduction.....	5
1.1. Related Work.....	5
1.2. Requirements	6
1.3. How IST World is Different from a Typical CRIS	6
1.4. Modelling Approach	7
2. Relational Data Model.....	9
2.1. A Brief History of CERIF.....	9
2.1.1. CERIF Features.....	9
2.1.2. CERIF 2004 Full Data Model Release 1.1	9
2.1.3. CERIF and IST World	11
2.2. CERIF Extensions	12
3. Conceptual Model: Research Networks Ontology (RENO)	16
3.1. Knowledge Representation Approach.....	16
3.2. Design Principles	16
3.3. Conceptual Grounds.....	17
3.3.1. Entity	18
3.3.2. Object.....	18
3.3.3. Agents	19
3.3.4. Topic	19
3.3.5. Location	19
3.3.6. Context	20
3.3.7. Person, Organization, Project, Publication, Event.....	20
3.3.8. Clusters.....	20
3.4. Specification of the Ontology.....	21
4. Conclusion.....	22
5. Bibliography	23

1. Introduction

The goal of this deliverable is to outline the data model, which determines the basic structure of the data repository of IST World, its central database, which hosts the majority of the data and allows for their analysis, augmentation, querying and presentation. The major objective at this level is modelling of the agents of interest (People and Organizations), which are related in a sort of social network between each other. It shall also be possible to model the contexts in which the agents co-operate with each other, e.g. Projects, Publications, Events – these co-occurrences are used as evidence for the structure of the social network. Finally, various classifications of the entities of interest will be possible (by location, skills, research areas, etc.) The network will allow for access to the data at different levels of generality, e.g.:

- countries, cities, organizations, departments, and individuals;
- sciences, research and applications areas, specific fields and techniques.

All levels are necessary since they can offer required level of data abstraction for further analysis of the data.

This data model will be extended in D1.2 to cover in the necessary details all sorts of information relevant to the so-called current research information systems (CRIS). It will be extended further, in deliverable D1.3, to allow for the comprehensive expertise representation, which is necessary for proper clustering and trend analysis, as well as for matching in the process of partner search.

1.1. Related Work

Analysis was performed over the following schemata (i) the CERIF model and the OWL ontology derived from it, (ii) the LT World model, (iii) the ProjectIntelligence data model.

The European CERIF standard has been formerly funded by European Institutions and is now in the responsibility of euroCRIS (<http://www.eurocris.org/>). CERIF stands for "Common European Research Information Format". It is a set of guidelines meant for everyone dealing with research information systems. CERIF is relevant to IST World, because it is a mature CRIS model developed in Europe and there are considerable amounts of data available under this model. Compliance with CERIF also ensures that the IST World data will be easily interchangeable with other systems. CERIF is presented in further details in section 1.

LT World (<http://www.lt-world.org/>) is an ontology-based Web portal on the wide spectrum of technologies for dealing with human languages, developed by DFKI and funded by the German Federal Ministry of Education and Research (BMBF). The so-called virtual information center LT World was first released to the public in October 2001. Since then, the numbers of worldwide visitors considerably increased. In May 2004 the system was completely re-launched and built upon ontological specifications. The LT World knowledge portal offers broad information on people, projects and organizations that deal with LT, collects research systems, tools, products and patents, lists events and news from research, development and market. In addition, LT World provides access to overview and background knowledge on 110 different language technologies. To cover the complexity of this knowledge domain, the LT World system is based on a multidimensional ontology which in addition to mere contents represents and supports central tasks needed within the acquisition and

maintenance processes and moreover handles interoperability and user interfaces. The LT World ontology was designed to facilitate relevant functionalities and processes. From our practical experience we have learned that the ontology-based approach has greatly improved and facilitated conceptualization, usability, maintenance, acquisition and data quality of the LT World portal. The accomplished approach of LT World in volume and thematic range could not have been managed by available personnel resources without the ontological basis [4].

Project Intelligence (PI, <http://pi.ijs.si/>) is a web portal built on the top of the Text Garden library for browsing and analysis of FP5 and FP6 project databases. Primary functionality is European Knowledge Map browsing, [3], which enables a user to access the database of projects from different cumulative points of view. More advanced functions enable keyword searching through projects database, semantic searching through to organizations competence profiles, and visualization of collaborations. The underlying data model employed in the Project Intelligence system is a simple one allowing storing of information only on project name, acronym, start and end dates, textual project description and a list of partner institutions. Due to the intense calculations on the data performed by the data mining tools the model had to be kept efficient. This is why a non-relational approach was chosen. All the data is stored in a form of serialized on-disk objects, which enable application of computationally intense calculations in an optimal way. The data processing algorithm can in this way be adapted to the specific algorithms of data mining to achieve best performance. Project Intelligence Data Model thus successfully uses a simple and yet capable data model.

1.2. Requirements

The central datastore of IST World (and thus its data model) should cover in efficient manner the following requirements:

- straight forward import of heterogeneous CRIS data (e.g. such coming from national CRIS databases);
- easy alignment with the LT World and the PI schemata;
- storage and management of huge amount of data – the target scale is 10^7 objects (people, organization, projects, etc.);
- management of temporally sensitive dynamically evolving data;
- provenance tracking – to allow for maintenance, e.g. iterative imports/updates of data;
- statistical analysis, as necessary for social networking and other analysis;
- structured multi-dimensional querying at different levels of generality.

1.3. How IST World is Different from a Typical CRIS

We would like to emphasize here a number of differences between IST World and a typical national CRIS system:

- IST World requires a simple core data model, which is easy to understand and maintain and to allow for scaling up a datastore, using it;
- The target audience of the system is not administrative staff, fully dedicated and trained for usage of a specific system; the people using and populating the IST World data will be mostly researchers, which are not motivated to learn the system, its data model and nomenclatures, as such;

D1.1: Definition of Central Data Structure

- To maximize the acceptance and the social-networking effects, it should require minimal efforts from the users to register, create profiles, update and use them, and provide feedback. This is not likely to happen if there is a complex data model with a numerous “look-up tables”.
- IST World is not an exhaustive archive, storing in reliable fashion the research data. It has the functions of a community-driven index, a large scale network, which facilitates the navigation within a CRIS information space. This is a major difference with the national CRIS systems, which have to preserve similar information as a primary register;
- The IST World model shall allow for partial, inconsistent, noisy information.

These differences have to be accounted to make it more clear how the objectives of the IST World data model differ from those which lead the development of CRIS models such as CERIF.

1.4. Modelling Approach

After serious technical discussion between the main technical partners in the project regarding the different storage and modelling alternatives, the consortium reached consensus that IST World needs to combine a pragmatic view with a more far reaching conceptual view. The baseline is that we need a basic version of the portal and thus an operational datastore very early in the project. In the same time we shall assure that deeper semantic analysis is possible at a later phase. Therefore, we had decided to start with a combination of two data models:

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

During the initial phase of the project, the portal will directly use only the relational model. The datastore will be a conventional RDBMS (relation database management system). The conceptual model will not be formally involved in the portal operations, i.e. there will be no component such as semantic repository or reasoner in the architecture. The conceptual model (the ontology) will be used as a design guidance and ground for development of a proper expertise modelling schema (deliverable D1.3). The plans are to get the ontology involved later on together with a semantic repository which properly integrates with the RDBMS.

The motivation for this decision was multifold; here follow just few of the arguments:

- Most of the CRIS data is available in RDBMS, many of which in CERIF or a compliant format. The adaptation of CERIF as a basis for the relational format re-uses the modelling efforts invested, saves time and makes possible quick development of the first version of the portal, and enables maximal interoperability.
- The adaptation of an ontology-based storage which can cover the scalability requirements and allows for efficient statistical analysis is a technically very challenging task, which could slow-down the development of the portal and deviate resources from other important aspects, such as networking and trend analysis. The choice of RDBMS as a platform for the IST World datastore allows for a quick start of the development and re-use of experience from the ProjectIntelligence project.

D1.1: Definition of Central Data Structure

- There is experience within the consortium in integration of the ontology infrastructure with RDBMS, which provides confidence that such a move can safely be planned for a later stage.

In the subsequent sections the relational model is presented first, as long, as it is more concrete and detailed. Next the conceptual model provides a generalization on top of it.

2. Relational Data Model

IST World project will provide functionality, which relies heavily on the ability to process vast quantities of text and structured data in a very short time. The portal must therefore build functionalities on top of a fast, advanced and reliable data management system. It was decided that an enterprise-class relational database management system (RDBMS) will be used as a basis for the IST world datastore. The concrete RDBMS is still to be selected based on a pragmatic evaluation of licensing schemata and the features of the major competitors. Considering the fair level of portability of SQL data schemes between different providers and the available tools which support this process (e.g. ERWin), this decision is not crucial at this stage.

The relational data model of IST World is developed in a bottom-up fashion starting from CERIF as basis and making the necessary extensions.

2.1. A Brief History of CERIF

The Common European Research Information Format (CERIF) was developed under the co-ordination of the European Commission. In its attempt to harmonise national Current Research Information Systems (CRIS) the European Commission funded work on the CERIF 1991 standard. In 2000, the European Commission transferred the custodianship of the CERIF standard to euroCRIS (<http://www.eurocris.org/>).

2.1.1. CERIF Features

What follows is a brief, high level description of the latest CERIF release. CERIF has the following design features:

1. Supports people, organisations, projects, funding programmes, publications, patents, products, services, facilities and equipment;
2. Provides a fully connected relational data model with powerful, flexible role-based relationships, including recursive relationships to represent hierarchies of people, organisations, projects and funding programmes;
3. Supports multiple language attributes;
4. Supports the latest Dublin Core standard.

2.1.2. CERIF 2004 Full Data Model Release 1.1

This Full Data Model is presented in five levels of abstraction as the data model is rather complex in its entirety¹. The objects at each level are colour coded (the associated colours are given below in parentheses). The five levels are defined as follows:

- Level 1: **Base Entities** (Green)

These are the primary entities of CERIF: **Person**, **OrgUnit** and **Project**.

¹ See the 'Picture' of the CERIF 2004 Full Data Model at the following location:

http://www.edward.grabczewski.btinternet.co.uk/CERIF/CERIF2004/WWW_CERIF2004_FDM_R1/Local/CERIF2004_FDM_R1Document.htm

D1.1: Definition of Central Data Structure

- Level 2: **Secondary Base Entities** (Blue)

These are the secondary entities of CERIF: **FundingProgramme**, **ResultPublication**, **ResultPatent**, **ResultProduct**, **Classification**, **ClassificationScheme**, **Facility**, **ExpertiseAndSkill**, **Service**, **Contact**, **Event** and **CV**.

- Level 3: **Language-Field Base Entities** (Yellow)

Where Level 1 and Level 2 entities have text attributes translated into multiple languages, these are stored in the Language-Field Base Entity tables. For example, a **Project** may have a **ProjectTitle** stored in English, French and Polish. Three rows, one for each language, are stored in the **ProjectTitle** table and associated to the corresponding project in the **Project** table. The same is done for **ProjectAbstract** and **ProjectKeywords**.

- Level 4: **Lookup Tables** (Grey)

Lookup tables are used to store lists of values for a particular attribute. For example, the lookup table **AcademicTitle** contains values such as 'Professor' and 'Doctor'.

Similar lists of values exist in the following lookup tables for entities: **HonorificTitle**, **OrgUnitType**, **Qualification**, **CVType**, **ResultPublicationType**, **ResultPatentType**, **ResultPatentStatus**, **ResultProductType**, **EquipmentType**, **Country**, **NUTSRegion**, **EventType**, **Language**, **PrizeAward** and **MultimediaType**.

Link tables (see Level 5) may also have a lookup table of values, such as the lookup table **Project_PersonRole** for the link table **Project_Person**. This table contains a list of possible roles between a project and a person, for example: 'hasProjectLeader', 'hasInvestigator' and 'hasAccountant'.

- Level 5: **Link Tables** (White)

Link tables implement many-to-many relationships between entities. The link table between the **Person** entity and the **Project** entity is called **Project_Person**.

Every link table contains a **Role** attribute and a period delimited by two timestamps. This period defines when the **Role** attribute of the relationship was valid. For example, project 'IST World' **hasProjectTeamMember** 'Edward Grabczewski'; this role relationship is valid from 1st April 2005 until *now*.

The link tables relating Level 1 entities are: **Project_Person**, **Project_OrgUnit**, **Person_OrgUnit**, **Person_Person**, **OrgUnit_OrgUnit** and **Project_Project**.

Some further examples of link tables relating Level 1 and 2 entities include: **FundingProgramme_OrgUnit**, **FundingProgramme_FundingProgramme**, **Project_ResultPublication**, **Project_Classification**, **Service_Classification** and **OrgUnit_Event**.

The latest release of the CERIF standard is specified on the euroCRIS Web site at <http://www.eurocris.org/en/taskgroups/cerif/> and is called *CERIF 2004 Full Data Model Release 1.1*.

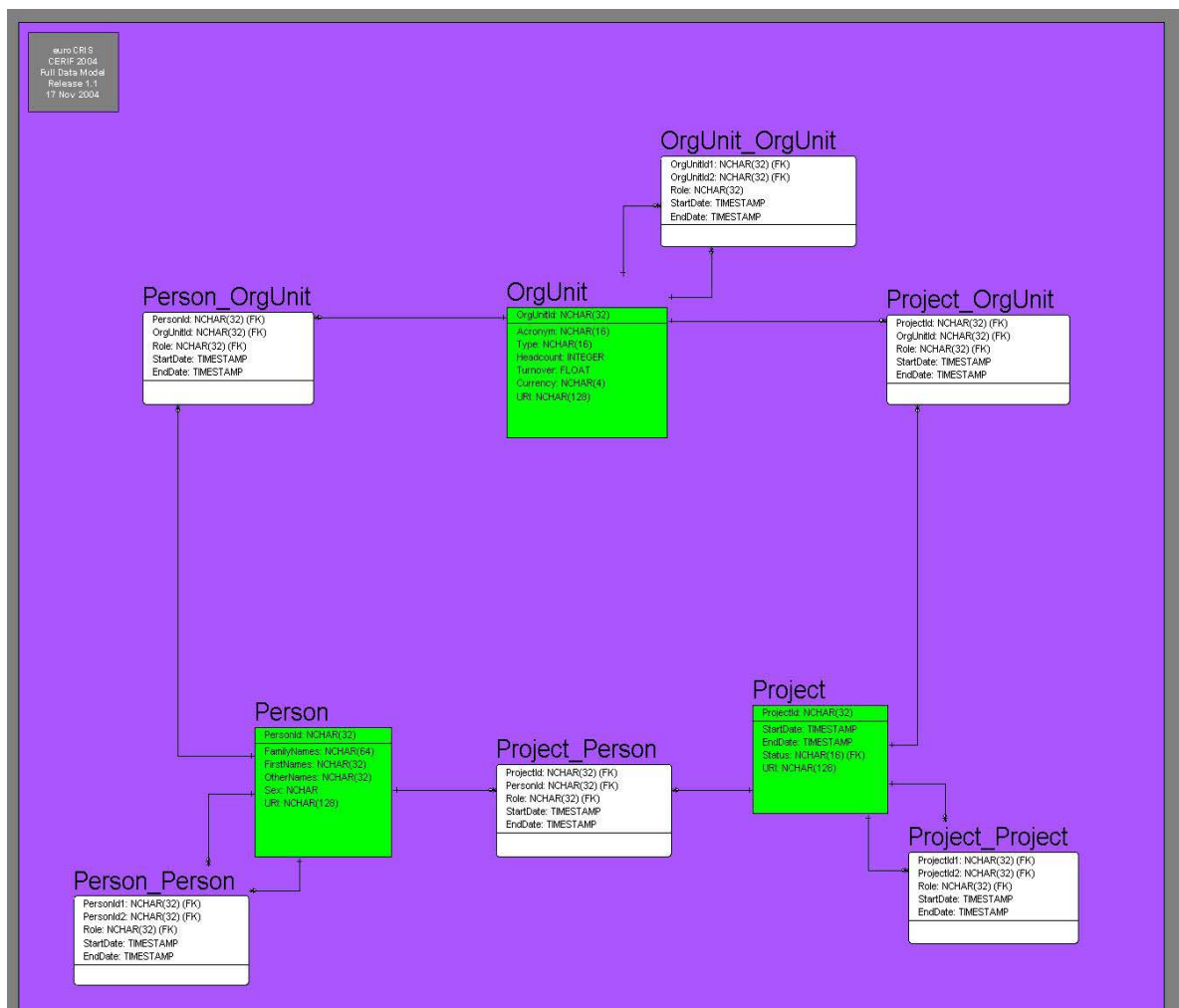


Figure 1. euroCRIS CERIF 2004 Full Data Model Release 1.1 Base Entities and their Relationships

Figure 1 shows a high-level diagram of the Base entities and relationships of the CERIF data model.

2.1.3. CERIF and IST World

The euroCRIS *CERIF 2004 Full Data Model* is a relational model that forms the basis for an OWL ontology defined by the euroCRIS *CERIF Task Group*. The OWL ontology derived from this CERIF data model will be used by the IST World project to ensure a consistent semantics for the IST World portal.

2.2. CERIF Extensions

After careful review of the CERIF model in light of the functionality of the IST World portal it was found that modifications to the CERIF model are necessary in order to meet the data storage requirements of the IST World.

The required CERIF modifications can be separated according to the requiring functionality:

- Functionality of providing information, trends and prediction on the current formal (de jure) and informal (de facto) state of the European and national research activities.
- Functionality of providing computer aided social networking.

Functionality of providing information, trends and prediction on the current formal (de jure) state of the European and national research activities. People, organizations, projects and results involved in scientific research must be modeled and described. Therefore modifications of the CERIF are necessary for capturing the required detail of overall research activity. Information on trends and prediction on the current informal (de facto) state of the European and national research activities will also be provided. Data providing this information will be collected autonomously by the IST World Portal by repeated search, collection and analysis of the relevant publications from the World Wide Web. Therefore the CERIF model must be extended to allow for storing of the data and meta-data about discovered publications. This data includes storing of the actual publication, extracted text and references.

Functionality of providing computer aided social networking. IST World will enable its users search and collaboration of people by providing use of the existing social network with regards to privacy, trust and interests of the members of the network. Therefore the CERIF data model must be extended in a number of ways to provide support for storing data for social networking.

In order to discover the required modifications of the CERIF model we have compared the existing CERIF model to the schema model of five different data sources that the future IST World portal will use to provide the described functionality. The SICRIS database holds information on all de-facto Slovenian research activity. The COBISS database is a Slovenian national on-line library system, which includes meta-information on all the literature in Slovenian libraries. FP6DB is an acronym for a database with information on European research activity. These sources and their schemas were used to assess the CERIF model with regards to storing the data required for providing formal state of the research in scope.

Google Scholar is an online resource that can be used for harvesting of most recently published research results. It was used to assess the CERIF model with regards to providing the informal state of the research activity in scope. Linked-In is a well known online resource used for computer aided social networking. It was used to assess the CERIF model with regards to its ability to store data relevant for social networking.

The analysis has showed the necessary extensions of the CERIF model. After consideration of the possible solutions to the described problems we have formed a list of suggested CERIF extensions and other solutions to the problems. Problems and the necessary modifications of the CERIF schema as well as other solutions are now listed:

D1.1: Definition of Central Data Structure

1) Problem/solution: storing original source database identifiers:

a. New table: **SourceDataBase**

with columns:

SourceDataBaseId	NCHAR (32)
URI	NCHAR (128)
Type	NCHAR (32)
SchemaName	NCHAR (128)

b. New tables: **Person_SourceDataBase,
 Project_SourceDataBase,
 OrgUnit_SourceDatabase,
 ResultPublication_SourceDatabase**

with columns:

PersonId/ProjectId/OrgUnitId/ ResultpublicationId	NCHAR (32)
SourceDatabaseId	NCHAR (32)
SchemaElement	NCHAR (128)
Identifier	NCHAR (128)

2) Problem/solution: storing research area descriptions in plain text:

Add to the existing table: **PersonResearchInterest**
 the additional column:

Description	NCHAR (1024)
-------------	--------------

3) Problem/solution: representing all roles of CERIF in a language independent way.

Automatically translate *all* the roles into English

4) Problem/solution: storing the country where the item was published.

Add to the existing table: **ResultPublication**
 the additional column:

CountryCode	NCHAR (4)
-------------	-----------

5) Problem/solution: storing the original language of the publication.

Add to the existing table: **ResultPublication**
 the additional column:

LanguageCode	NCHAR (2)
--------------	-----------

6) Problem/solution: storing the content type, keywords and description.

New table: **ResultPublication_Content**

with columns:

ResultPublicationId	NCHAR (32)
LanguageCode	NCHAR (2)
Type	NCHAR (32)
Keywords	NCHAR (1024)
Description	NCHAR (1024)

7) Problem/solution: storing a publication physical description.

New table: **ResultPublication_PhysicalDescription**

D1.1: Definition of Central Data Structure

with columns:

ResultPublicationId	NCHAR (32)
LanguageCode	NCHAR (2)
PhysicalDescription	NCHAR (1024)

8) Problem/solution: storing remarks and notes for various entities.

New tables: **Person_Notes, Project_Notes, OrgUnit_Notes, ResultPublication_Notes**

with columns:

PersonId/ProjectId/OrgUnitId/	
ResultPublicationId	NCHAR (32)
LanguageCode	NCHAR (2)
Notes	NCHAR (1024)

9) Problem/solution: storing raw resources e.g. publications.

a. New table: **Resource** for storing raw resources

with columns:

ResourceId	NCHAR (32)
Name	NCHAR (1024)
Size	INTEGER
Type	NCHAR (32)
Data	BINARY

b. New table: **ResultPublication_ResultPublicationResource** for many-to-many relation with the resource table

with columns:

ResultPublicationId	NCHAR (32)
ResourceId	NCHAR (32)
Role	NCHAR (32)
StartDate	TIMESTAMP
EndDate	TIMESTAMP

c. New table: **ResultPublication_ResultPublicationResourceRole**

with columns:

Role	NCHAR (32)
RoleFull	NCHAR (64)

10) Problem/solution: storing extracted text (Full text, Body text, Citations and Abstract).

a. New table: **Text**

with columns

TextId	NCHAR (32)
LanguageCode	NCHAR (2)
Length	INTEGER
Encoding	NCHAR (32)
Text	LONG

b. New table: ResultPublication_Text

with columns:

ResultPublicationId	NCHAR (32)
TextId	NCHAR (32)
Role	NCHAR (32)
StartDate	TIMESTAMP
EndDate	TIMESTAMP

c. New table: ResultPublication_TextRole

with columns:

Role	NCHAR (32)
RoleFull	NCHAR (64)

11) Problem/solution: storing Citation references of one article to another.

a. New table: ResultPublication_ResultPublication

with columns:

ResultPublication1	NCHAR (32)
ResultPublication2	NCHAR (32)
Role	NCHAR (32)
StartDate	TIMESTAMP
EndDate	TIMESTAMP

b. New table: ResultPublication_ResultPublicationRole

with columns:

Role	NCHAR (32)
RoleFull	NCHAR (64)

In the course of the development of the IST World datastore, the different variants for solving the modelling gaps will be evaluated and the best solutions will be adopted.

3. Conceptual Model: Research Networks Ontology (RENO)

The conceptual model is specified as an ontology, which allows for a clear definition of the semantics that enables inferencing over the data. Thus, it has a much higher analytical and predictive potential as compared to models based on relational algebra.

We call the conceptual model of IST World “Research Network Ontology” or shortly RENO. It is designed in a top-down fashion, on the basis of a conceptual analysis of the domain and the tasks essential for the IST World portal. RENO starts from a few basic concepts and their relations and then specifies their mapping to the CERIF entities.

3.1. Knowledge Representation Approach

We choose as a core data model for the ontology RDF. It bears the following advantages:

- It is a flexible, semi-structured, triple-based data model;
- Allows for easy aggregation of information from different sources;
- Allows for modelling of multi-dimensional information spaces, with non-scalar dimensions (e.g. spatial regions, expertise areas, etc.);
- It is a W3C standard – the core of the so-called Semantic Web;
- There are many tools for efficient management of RDF data.

While RDF (together with its schema language RDFS, [1]) provides good base for structuring of heterogeneous data, it is a very basic knowledge representation (KR) schema, missing many simple, but useful primitives. Basic semantics of the IST World conceptual model will be specified in OWL Lite, [2], because:

- It is compatible with the RDF data model – the standard representation of OWL is RDF, the standard syntax is the XML syntax for RDF;
- It provides a standard way for defining generalization and aggregation rules (e.g. transitive, symmetric, inverse properties);
- Staying in a limited sub-set of OWL allows for efficient reasoning, while using the ontology as an advanced database schema.

3.2. Design Principles

Here we summarize some principles underlying the design of RENO in order to ensure efficient reasoning:

- follow a clear two-layered extensional semantics. Which means to avoid abuse of the `rdfs:subClassOf` property in order to avoid the computational penalties related to the RDF meta-modelling freedom;
- keep the expressivity in the so-called OWL-DLP fragment, which guarantees that the reasoning over it does not require satisfiability checks (as in OWL DL). This means that tractable reasoning algorithms can be used, which is an absolute must, considering the cardinality of the data which are to be managed in the IST World store;

- Enforce a clear separation between:
 - Schema (schema-ontology, the core data model); and
 - Topic hierarchy (subject hierarchies, classification hierarchies, etc)

These design principles are the same as those crafted and proven in the development of the PROTON upper-level ontology. Description of the latter can be found in [5]; section 4 there represents the design principles.

3.3. Conceptual Grounds

The analysis of the goals and scope of IST World led us to the following generalizations: the essential entities in the model are **Agents** (**Persons**, **Organizations**) which co-operate with other agents in different sorts of **Contexts** (**Projects**, **Publications**, **Events**), thus determining, or providing evidence for, social-networks. We introduce **Object** as a common super-class for agents and contexts. The objects can be associated with **Topics** and located in **Locations**. Further, the objects have a temporal extend (lifespan, duration, etc.).

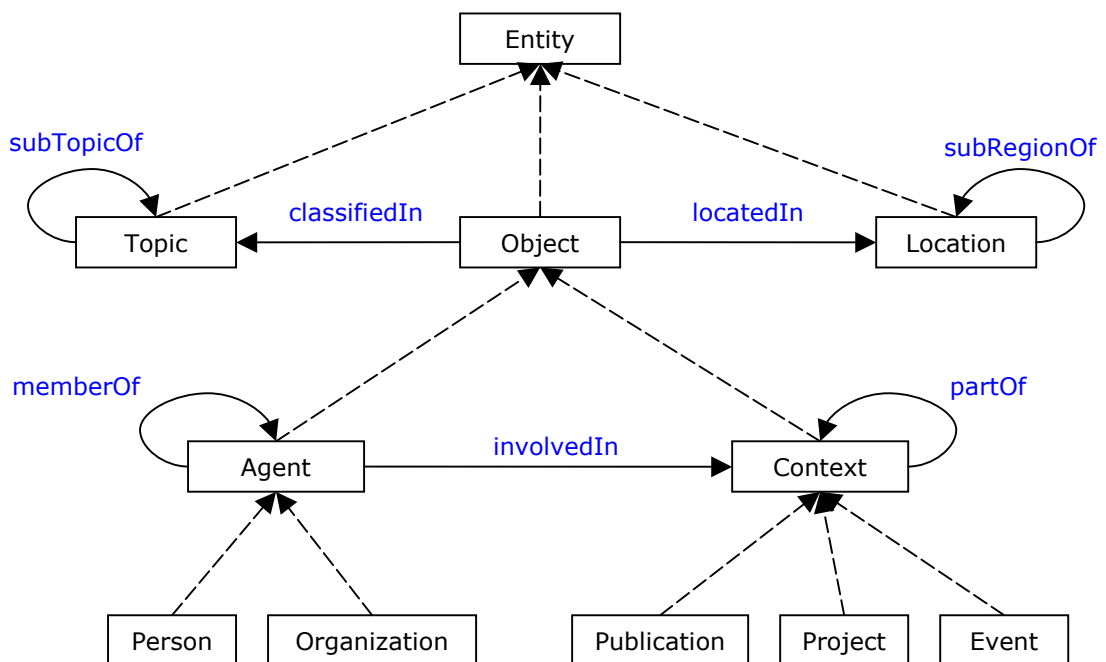


Figure 2. RENO Conceptual Model

Figure 2 depicts the conceptual model behind RENO. The dashed arrows represent sub-class relationships. Below we present this model in a semi-formal fashion together with relations to the corresponding CERIF tables and columns. Unless otherwise stated, all relations are not mandatory and have cardinality many-to-many.

3.3.1. Entity

Entity is used to model individuals in the domain of discourse, excluding system information such as Forms or auxiliary data.

- Sub-classes: **Object**, **Location**, **Topic**.
- CERIF equivalent: no direct equivalent, corresponds to the in-formal notion of entity in the definitions, formalized as Entity class within the CERIF OWL ontology.

3.3.2. Object

Object is an artificial super-class handling the commonalities between agents and contexts; thus it is a sub-class of Entity.

- Sub-classes: **Agent**, **Context**.
- CERIF equivalent: none.
- Objects can be associated with locations through the **locatedIn** relation. This is a weak and rather general primitive for modelling of the spatial extent of Objects. Examples and special cases:
 - A person can be located in specific city;
 - An organization can be located at multiple places (points of presence) with one or more countries, directly or through its child organizations;
 - It can be considered that a research project is located in all the locations where the agents, participating in it, are located;
 - There is no easy way to define location for some publications.
- Objects can be associated with topics through the **classifiedIn** relation. The modelling should allow for specification of the strength of the association between an object and a topic. Examples and special cases:
 - A faculty of mathematics and informatics of an university can be classified under "Computer Science" and "Mathematics";
 - A particular researcher can be classified in all the specific research fields where s/he is active;
 - An event, e.g. a conference, can be classified under the research fields of its scope.
- Objects have a lifespan, or a temporal extent, determined through the **startDate** and **endDate** attributes. Examples and special cases:
 - For projects and events the lifespan is obvious and corresponds to **StartDate** and **EndDate** columns in the respective tables in CERIF;
 - For persons these are the natural lifespan dates, which are not that relevant for IST World, but still can be modeled;
 - The start date of an organization is the establishment date, the end date is the date when it ceases existence (getting closed, transformed, etc);

- Typically, for publications the start and end dates will both match the publication date (**Pub_date** column in the **Result_Publication** table of CERIF).

3.3.3. Agents

Agents are typically persons or organizations. They co-operate between each other in various contexts (discussed later on). The Agents can be nested within each other, in the sense that people and organization can be part of organizations and represent them. For instance, the fact that one department from a university has participated in a project also means that the university participates in the project.

- Sub-classes: **Person, Organization**;
- CERIF equivalents: **PERSON, ORGUNIT**;
- Agents build a nesting hierarchy based on a transitive **memberOf** relationship; **childOrganizationOf** is a special case of the latter one;
- Agents can appear in different contexts, which is modeled here through the **involvedIn** relation. Examples and special cases:
 - Persons or organizations, which participate in a project;
 - The authors/contributors of a publication;
 - Organizers and speakers at conferences. It is a matter of interpretation whether the regular attendants of a conference should be considered involved in it as a context – the fact that two persons attended the same conference (without being organizers or speakers) provides a rather weak evidence for co-operation;

3.3.4. Topic

Topic is a general class, which represents different sorts of classifications, taxonomies and subject hierarchies. The commonality between all of them is that they are used to classify objects in. It is important to realize that topics are different from the classes, as the latter are defined in most of the OO paradigms and the extensional KR formalisms. The major difference is that you can say for a particular person X that he is an instance of Person and further infer that he is an instance of all of its super-classes. On the other hand, considering a topic such as AI (as a field in the science), one can associate a person with this field, but saying that s/he is an instance of AI would cause a lot of tangleness and paradoxes. Still, in a fashion similar to the OO-classes, the topics are often ordered in hierarchies.

- Sub-classes: none;
- CERIF equivalents: no direct equivalent; it is a generalization for **Expertise_Skills, Project_Classification, Classification** (indirectly);
- Topics build a hierarchy based on a transitive **subTopicOf** relationship.

3.3.5. Location

- Sub-classes: none in the current version (otherwise all the sub-classes of **ptop:Location** from PROTON's Top and Upper modules, e.g. **Region, Country, Province, PopulatedPlace, City**, etc.). It is also related to a number of columns in the **Contact** table: **Con_City_Town, Con_Province_State, Region_Code**;

- CERIF equivalents: no direct equivalent; it is a generalization of **Country**;
- The nesting of locations within each other can be modeled with the **subRegionOf** relationship, which is transitive.

3.3.6. Context

Context is introduced here as a class specific to the IST World and research networks analysis. It is a super-class for all the typical forms of co-operation between agents (persons or organizations) in a research community.

- Sub-classes: **Project**, **Publication**, **Event**;
- CERIF equivalents: none;
- The contexts can also form part-whole (meronymical) hierarchy through the transitive relationship **partOf**.

3.3.7. Person, Organization, Project, Publication, Event

There are number of classes which have straightforward correspondence to the CERIF tables:

- **Person: Person**;
- **Organization: OrgUnit**;
- **Project: Project**;
- **Publication: Result_Publication**;
- **Event: Event**.

3.3.8. Clusters

The model can be extended to support Clusters as groups of Agents, related through co-occurrence in Contexts. Unlike the other entities, Clusters are derived or inferred on the basis of analysis of the research networks. The definition of the part of the schema related to clusters is postponed for a later stage; below we only list some relevant considerations:

- Sub-classes: none;
- CERIF equivalents: none;
- Clusters are automatically derived based on the level of association between agents in general or over a specific period of time;
- Once derived, they can be stored in an explicit form, given a name, time-span, strength (weight). The strength can vary over the time.
- The relation between a cluster and each of the Agents, part of it, has strength (weight);
- Clusters can be associated with Topics, based on the topics associated with the Agents;
- Clusters can be associated with Locations, based on the location of the Agents, which are part of them.

3.4. Specification of the Ontology

The ontology is formalized as an OWL file in RDF/XML format and provided separately. This formalization can benefit from integration with OWL ontology derived from CERIF as well as with the PROTON ontology. However, we preferred not to implement the integration at this stage in order to allow for maximum clarity and easy of understanding.

To reduce the modelling effort we had derived (in terms of copying, instead of import) a number of definitions from PROTON System and Top modules.

4. Conclusion

We have presented a pair of complementary data models for the central data store of IST World:

- Relational model – extension of the well-established CERIF model. It allows for easy import of CERIF data and efficient manipulation in a relational database;
- Conceptual model – a generalization and formalization of the relational model as an ontology. It allows access to the information at different levels of abstraction, thus

We do expect that over the duration of the project both data models will develop, following the needs and the specifics of the IST World portal implementation. To allow for quick implementation of the first version of the portal, the IST World central storage will be a relational database using the relational data model. The major goal in terms of storage technology will be to make possible simultaneous usage of the same data through both models. This effectively means, allowing for semantic (ontology-aware) querying and manipulation of the data stored in the relational database.

5. Bibliography

1. Brickley, D; Guha, R.V, eds. *Resource Description Framework (RDF) Schemas*, W3C. <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>
2. Dean, M.; Connolly, D.; van Harmelen, F.; Hendler, J.; Horrocks, I.; McGuinness, D.; Patel-Schneider, P.; Stein, L.A. *Web Ontology Language (OWL) Reference Version 1.0*. W3C Working Draft 12 Nov. 2002. <http://www.w3.org/TR/2002/WD-owl-ref-20021112/>
3. Grobelnik, M.; Mladenić, D.; Jermol, M. *Towards the EU IST Projects Knowledge Map and Project Partners Competence Directory*. In proceedings of The 4th European Conference on Knowledge Management. Oriel College, Oxford University, UK; 18-19 September 2003, pp 387.
4. 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
5. Terziev, Iv; Kiryakov, A; Manov, D. *Base upper-level ontology (BULO) Guidance*. Deliverable D1.8.1 of the SEKT project. July 2007. <http://proton.semanticweb.org>.