



Knowledge Base for RTD Competencies in IST



Deliverable D3.1

Data Import/Export Specification as XML Schema

Author(s):	Brigitte Jörg
Identifier:	D3.1
Work package:	WP3 Data Acquisition
Lead Partner:	Deutsch. Forsch. für Künstliche Intelligenz (DFKI)
Partner(s):	all
State of document:	final
Version:	1.0
Dissemination Level:	Public
Date:	2005-08-12

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

This document specifies the formal Import/Export interfaces for the IST World data repository. The specification is implemented as a set of XML Schemas, which are used by the participating partners to validate their national datasets before importing them into the IST World repository. The technical partners agreed that XML Schemas should be used for data validation to ensure a proper data quality and hence improve planned IST World services and functionalities. All partners will provide their national datasets in XML Schema compliant XML formats.

The IST World repository comprises the following content types:

- Person
- OrgUnit
- Project
- ResultPublication

Each content type can be represented by XML element sets, which are defined in type-specific XML Schemas. Moreover the schemas are used for validation. For a better understanding of XML Schemas and how they are used for the IST World repository, we will examine the XML Schema engineering process itself.

Deliverable 1.1 describes the conceptual model of the IST World data repository in which there are two agents (People and Organizations) that cooperate in a number of contexts (Projects, Publications and Events). The major aim of this deliverable is to ensure a correct representation of those agents and their contexts in the IST World data repository in terms of the CERIF data model. With deliverable 1.1 we investigated the CERIF data model and documented the IST World specific extensions to that model. In this document we take this one step further by examining to population of the repository with validated XML data.

The *XML Schema Generation* section describes the transformation of relational data repository tables into modular XML representations of those tables, which result in the definition of mappings.

Content Table

Knowledge Base for RTD Competencies in IST	1
IST World Consortium	2
Abstract	3
Content Table	4
1. Introduction.....	5
1.1. Relational DB Model	6
1.2. Conceptual Model	6
1.3. XML / XML Schema	7
2. XML Schema Generation	8
3. Conclusion.....	12
4. XML Records / XML Schemas.....	13
4.1. Person	14
4.1.1. XML Example Record for Person	14
4.1.2. XML Schema Example for Person	15
4.2. OrgUnit.....	16
4.2.1. XML Example Record for OrgUnit.....	16
4.2.2. XML Schema Example for OrgUnit	17
4.3. Project.....	18
4.3.1. XML Example Record for Project.....	18
4.3.2. XML Schema Example for Project	19
4.4. ResultPublication	20
4.4.1. XML Example Record for ResultPublication	20
4.4.2. XML Schema Example for ResultPublication	20
5. Mappings from XML to RDBMS.....	21
6. Bibliography	23

1. Introduction

The IST World project aims at setting up and populating an information portal with innovative functionalities that help to promote RTD competencies in IST on a local, national and European level. In order to achieve a critical mass of accurate data and comprehensive coverage, we cannot rely on any *single* method of data acquisition. Therefore, there are three major approaches to populate the IST World repository:

- 1. Base:** Existing data is firstly imported from CORDIS, then from domain-specific portals such as LT World and later from CERIF-conformant national and European Current Research Information Systems (CRISs).
- 2. Community:** The portal offers services for community building and maintenance in which organisations, groups and experts register themselves, and their projects to build professional virtual communities. Promotion campaigns in the new member states (NMS) and associated candidate countries (ACC) aimed especially at SMEs will be carried out in order to ensure that research competencies from these countries are well represented in the repository.
- 3. Automatic:** Web and text mining techniques will be used to acquire additional, unstructured data that is not held in a traditional database.

This deliverable is concerned mainly with the first (*base*) method of data acquisition to ensure consistency between incoming national data sets and the IST World data repository. The second (*community*) method of data acquisition is motivated and relies upon an already acquired critical mass of *base* data and is supported by Web forms based on the specifications in this deliverable. The third (*automatic*) method of acquisition does not depend on our specified XML Schemas as it has its own means of data collection and validation.

After a round table discussion among the partners at the Bratislava kick-off meeting it was clear that the national datasets differ markedly in quantity, quality and data format. In order to get a more detailed and realistic view of nationally available content types, formats and the quantity of national data, all partners agreed a questionnaire to survey these issues.

Examining the results of the questionnaire, the following conclusions were reached:

- The early availability of a large quantity of national datasets of the first, *base*, kind is not very realistic in most cases.
- A critical mass of *base* data *can* be acquired for the purposes of *community* activity.
- An early start to *base* data acquisition is required even if the quantity of data is small; regular updates are needed; the partners need to identify additional sources of national *base* data.
- The legal state of national datasets must be considered; even more important is an early, formal specification for offering Web forms to acquire data of the second, *community*, kind.
- Web mining and text mining techniques are of great importance for the acquisition of IST World data.

After serious technical discussion between the main technical partners in the project regarding the different storage and modeling alternatives, the consortium reached consensus that IST World needs to combine a pragmatic view with a more far reaching conceptual view, as already stated in deliverable 1.1. The baseline is that we need a basic version of the portal and thus an operational data store very early-on in the project. At the same time we need to ensure that deeper semantic analysis is possible at a later stage. Therefore, we have decided to start with a combination of two data models:

- **Relational model:** A detailed RDBMS schema, based on an extension of the CERIF 2004 Full Data Model Release 1.1.
- **Conceptual model:** An ontology, allowing for proper conceptualization of the domain and deeper analysis.

1.1. Relational DB Model

The IST World project provides functionality that relies heavily on the ability to process large quantities of text and structured data in a 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 data store. The relational data model of IST World is developed in a bottom-up fashion, starting with the CERIF¹ [1] data model as a base and making some necessary extensions.

1.2. Conceptual Model

The conceptual model for IST World is specified as an ontology, called *RENO*², which allows for a clear definition of the semantics to enable inferencing over the data. Thus, it has a much higher analytical and predictive potential compared to models based on relational algebra. In the context of systems design, so called 'formal ontologies' have gained popularity in the last few years throughout the computer science research community. Formal ontologies comprise classes or so called 'concepts' and their relations in a multiple inheritance order, going beyond taxonomical "is-a" relations. Each class contains specific value definitions and restrictions for properties that are propagated to subclasses.

For modeling the IST World XML APIs, we started from bottom-up and built a new ontology representing the relational CERIF model. This ontology at its current stage mainly serves as a documentation of CERIF and its extensions specified in deliverable 1.1. It is of use for modeling the XML APIs, as described later, and it can moreover be considered to be the first step on the way to the building the *RENO* conceptual model. *RENO* and the CERIF *ontology* are built in OWL DL.

OWL is the latest "Web Ontology Language"³ recommendation for ontology development by the W3C consortium and is used as a kind of modeling standard. The OWL syntax itself is based on XML format.

¹ Common European Research Information Format (CERIF) [1]

² Research Network Ontology (RENO) specified in deliverable 1.1.

³ W3C: Web Ontology Language (OWL): <http://www.w3.org/2004/OWL/>

1.3. XML / XML Schema

The **Extensible Markup Language (XML)** is a simple, flexible text formatting language derived from SGML⁴ (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere⁵. A characteristic feature of XML is its semi-structured data organization. The data describe themselves with their simple syntax and do not *necessarily* rely on a formal schema [3]. XML instances can simply be represented by any element, as can be seen from the example (Fig. 1):

```

<xmlInstances>
  <xmlInstance>
    <entryType>Organization</entryType>
    <entryRole>commercial</entryRole>
  </xmlInstance>
  <xmlInstance>
    <entryType>Organisation</entryType>
    <entryRole>academic</entryRole>
  </xmlInstance>
</xmlInstances>
  
```

Figure 1: XML Example

Most common data formats can easily be transformed into a list of XML elements. Be it simple text files or Excel tables, Word documents or any proprietary database or file formats. Compared with the complexity of a relational DB schema, XML representations of objects are very simple and much easier to communicate, especially to people that are not familiar with database technologies.

Thinking of large amounts of XML data from different sources, a given set of elements is helpful for a proper data organisation and for consistency control between various datasets. XML Schemas provide a means for defining the structure, content and semantics of XML documents. XML Schemas express shared vocabularies and allow machines to carry out rules made by people. XML Schema was approved on 2 May 2001⁶ as a W3C recommendation.

In the IST World project, we make use of XML Schemas to validate incoming XML datasets from participating partners.

⁴ Standard Generalized Markup Language (SGML): <http://xml.coverpages.org/sgml.html>

⁵ W3C: Extensible Markup Language (XML): <http://www.w3.org/XML/>

⁶ W3C: XML Schema. <http://www.w3.org/XML/Schema>

2. XML Schema Generation

As has been mentioned, XML Schemas are a means of validating XML datasets. As such, they can also be considered to be the meta-level description of specified XML datasets. For the creation of IST World specific XML Schemas the following steps are taken (See also Fig 2):

1. Existing Relational CERIF Model
2. Model Transformation
3. Resulting CERIF Ontology
4. CERIF-ontology-based APIs
5. XML Representation of IST World Objects
6. XML Schemas Generation for XML Validation

Changes in the relational model restart the whole process.

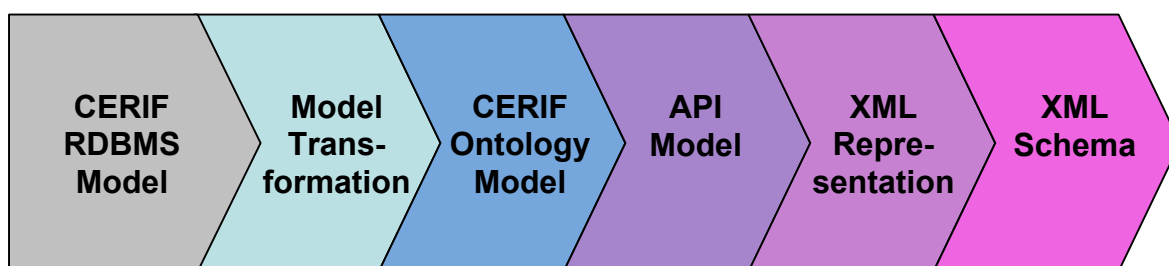


Figure 2: XML Schema generation process

1. Relational CERIF Model

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/>). A brief CERIF history, CERIF features and IST World specific extensions are documented in deliverable 1.1. The latest version of the "CERIF Full Data Model Release 1.1", an implementation, images, documents and SQL scripts can be downloaded from the following address: <http://www.eurocris.org/en/taskgroups/cerif/cerif2004/>

According to the latest CERIF version there are three base entities Person, OrgUnit, Project. Moreover there are secondary base entities like ResultPublication, Contact, CV, Classification, Service etc., there are language field based entities like ProjectTitle, OrgUnitName, PersonResearchInterest etc., there are link tables like Person_OrgUnit, Person_CV, Person_ExpertiseAndSkills etc., and there are the lookup tables like OrgUnitType, Person_ContactRole, AcademicTitle, and many more. All these tables are further specified with columns, as is common practice with relational database schemas.

2. Model Transformation

For the transformation of the relational CERIF model into an object-centered CERIF ontology model, first, all available tables are represented as classes to make sure that the whole CERIF schema is captured. Each table is represented as a class in the CERIF ontology whose name equals the name of the table in the relational model. Next, all associated table columns are turned into properties of the class with identical names to those of the columns.

For example, the Person table with columns firstNames, familyNames and otherNames is represented as the Person class with properties firstNames, familyNames and otherNames.

The transformation of the relational model into an ontology is based on the CERIF FDM Release 1.1, which is freely available online [2].

3. Resulting CERIF Ontology

Through described bottom-up transformation, the semantics of the resulted CERIF ontology⁷ is far from being merged with the conceptual model RENO⁸, but as a technical means for modeling the XML APIs and moreover as a means for communicating the original CERIF schema model with its extensions between the technical partners, it is extremely helpful. One can quickly navigate through the classes and read available comments for each class and property. The latest version of the CERIF ontology with IST World specific extensions is available at the collaborative portal: <http://ist-world.dfki.de/downloads/ontologies/cerif.owl>

4. XML API Modeling

In the context of this deliverable, the XML APIs can be considered to be the conceptual specifications for the IST World XML element sets, which directly correspond to the formal specification of XML Schema implementation. Compared with XML Schema, the conceptual models of the XML APIs lack formal constraints and syntax. They can be considered to be a plain list of names, representing the IST World agents and context, not using any rule or type definitions. For modeling the APIs we made use of the CERIF ontology. Because all the CERIF tables with columns are represented as classes with properties, the respective classes have to be brought into such a hierarchical order so that ontological inheritance automatically hands over the concerned properties to the XML API subclasses. This mechanism is also an advantage when new properties are added; they become visible automatically to all classes at deeper levels.

The IST World technical partners agreed that the XML instances are to be represented by a flat, one-level list of XML elements. To make it most easy for the partners to prepare and provide their data we avoid deep hierarchies or complex structures.

The XML API concepts or classes were modeled at the deepest level of the ontology to catch all relevant CERIF properties from super classes, as can be seen in figure 3.

⁷ For modeling the CERIF ontology, we used Protégé, developed at Stanford (<http://protege.stanford.edu>). It is supported by a very active community and offers a lot of features.

⁸ Research Network Ontology (RENO) specified in deliverable 1.1.

D3.1: Data Import/Export Specification as XML Schema

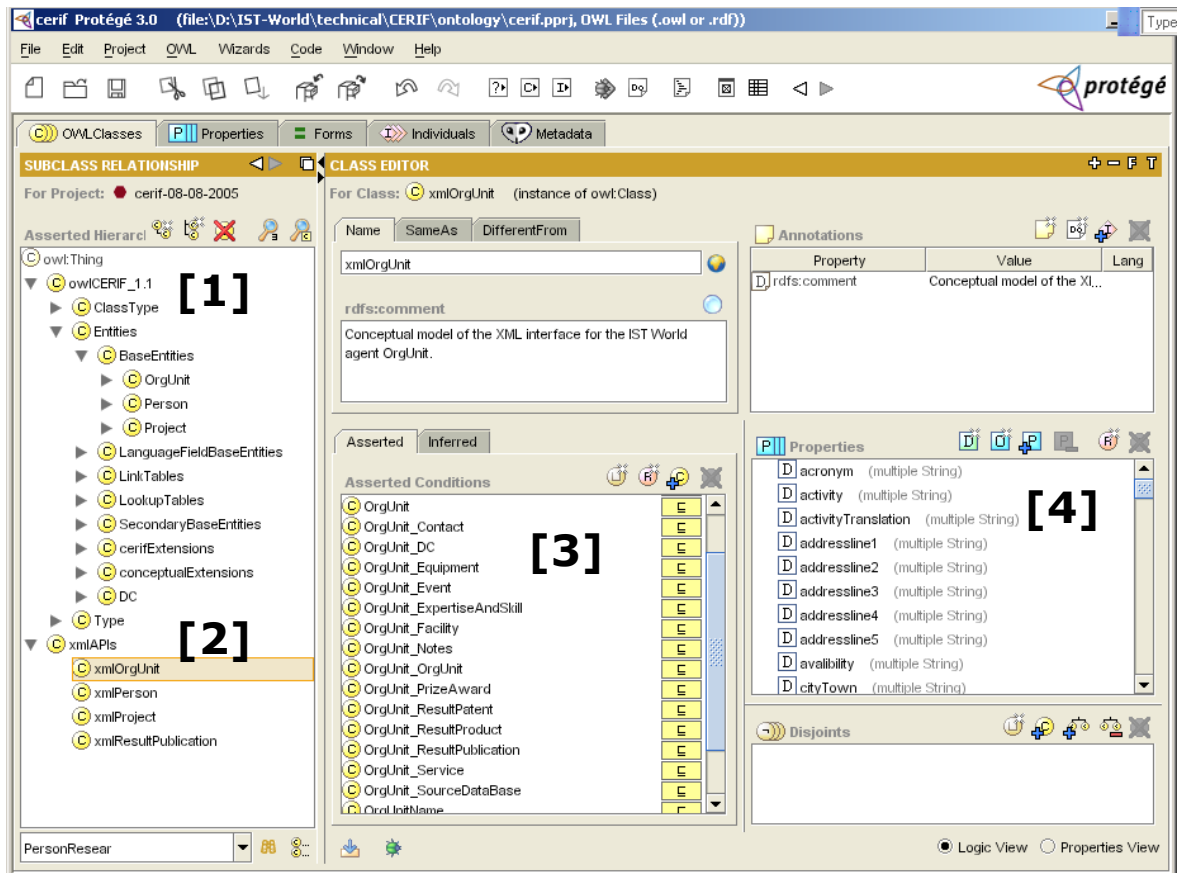


Figure 3: CERIF ontology⁹ at the top [1] and XML API concepts [2] with list of inherited classes [3] of the xmlOrgUnit class and the list of properties [4] at xmlOrgUnit class level.

At this point one can clearly see the differences between relational and object-centered data models keeping all properties with the object. XML supports object-centered data modeling. In order to entirely represent CERIF based relational entities with object-centered XML, some amendments are needed, when thinking of the properties. With ontological inheritance, properties are handed down from classes to subclasses. But as a logic consequence of inheritance, properties that are assigned at several class levels like i.e. translation, name, description, or title, are now, at the deepest class level, only available once and the associations are lost.

We therefore manually added for all concerned properties additional properties whose property name now indicates their associations also at deeper class levels. Instead of only *name* for example, we now have *orgUnitName*, *contentName*, etc. This can be seen in figure 3 part [4], where *translation*, is now *activityTranslation*. We consider this necessary, as we don't want to invent a deeper and hence more complex XML structure. The CERIF roles and types are set as attributes with the XML elements. Figure 4 shows an example of an XML extract representing the agent OrgUnit, compliant with the xmlOrgUnit API, modeled in the CERIF ontology (See figure 3).

The technical partners agreed that the original CERIF properties at class level will all be kept for a general compatibility and will be regarded later, with respect to RENO.

⁹ Protégé screenshot of the CERIF ontology based on the relational CERIF model version 1.1.

All conceptual extensions are modeled in the class `conceptualExtensions` of the CERIF ontology (see screenshot figure 3 part [1]).

The modeling of the IST World XML APIs resulted in a list of mappings for additional properties due to conceptual changes. The mappings are needed when importing XML datasets into the relational CERIF-based IST World repository. For documentation they are listed in chapter five.

5. XML Representation of IST World Objects

The XML datasets for IST World agents have to comply with the specified XML APIs. Each data provider has to ensure, that his national XML dataset is correctly represented. Correct XML examples for each IST World agent and its context will be given in chapter four, with links to online resources. Figure 4 shows a sample extract for `OrgUnit`.

```
<?xml version="1.0" encoding="UTF-8"?>
<instances>
  <orgUnit type="academic">
    <acronym>DFKI</acronym>
    <activity>DFKI is focusing on the complete cycle of innovation ...</activity>
    <activityTranslation languageCode="DE">Das DFKI konzert ...</activityTranslation>
    <addressline1 role="main">Erwin Schrödinger-Straße</addressline1>
    <addressline2 role="main">Building 57</addressline2>
    <cityTown role="main">Kaiserslautern</cityTown>
    .
    .
    .
    <keywords>artificial intelligence, language technology,
      knowledge management, intelligent user interfaces
    </keywords>
    <notes>Notes about the entry.</notes>
  </orgUnit>
</instances>
```

Figure 4: XML extract for the IST World agent `OrgUnit`

6. XML Schema for Validation

Altova¹⁰ *XML Spy* was used to automatically generate XML Schemas from XML representations. XML Schemas are a meta-level description of the specified XML datasets. To validate the XML dataset, an XML Schema has to be assigned. Technically this is realized by adding the following attributes/values (red-colored) to the XML dataset. This is illustrated in figure 5.

```
<?xml version="1.0" encoding="UTF-8"?>
<instances xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://ist-world.dfki.de/downloads/
  xmlSchemas/OrgUnit.xsd">
  <orgUnit>
    <acronym>DFKI</acronym>
    .
    .
    .
  </orgUnit>
</instances>
```

Figure 5: XML extract for the IST World agent `OrgUnit` with XML Schema reference

XML Schema examples for each IST World agent and its context will be given in chapter four, with links to online resources.

¹⁰ XML Spy from Altova: A commercial tool for work with XML: http://www.altova.com/products_ide.html

3. Conclusion

This document can be regarded as a reference document, which will be changed over the development period.

Changes to the XML dataset definitions will most probably be necessary with each incoming dataset or during development over the lifetime of this project. The described process of XML Schema generation shows the dynamic and flexibility if changes are needed. So we can quickly model API extensions and subsequently build new XML Schemas very easily.

The Integration of the XML datasets into the IST World data repository will be done automatically, taking into account the defined mappings. Our experience with automatically generated program code from defined mappings is very positive and can be more flexible than pure SQL statements.

Next steps will be, to conceptualize and then realize Web forms based on the XML Schema definitions to enable data entries by the community when the portal goes public.

4. XML Records / XML Schemas

The current XML Schemas are available online for being used by all participating partners to validate their national datasets before sending data to the IST World repository.

What now follows are some examples of IST World XML records and their related XML Schemas for the four main IST World content types:

- Person
- OrgUnit
- Project
- ResultPublication

The XML records are kept as a 'one-level' list of elements with *type/role/languageCode* attributes, to make it easy for the partners to prepare and provide their datasets. The *languageCode* is mandatory for translation elements. The *type* attribute is mandatory for OrgUnits and ResultPublications. The *role* attributes are non-mandatory in all cases but can be applied if preferred. To keep it simple and open, we did not yet enforce constraints on role and type attribute values.

Datatype values are defined as unformatted "string" values in most of the cases. This avoids running into validation problems caused by different national datatypes and formats, which are difficult to foresee at this stage in the project.

4.1. Person

XML example files and XML schemas for the IST World agent Person can be downloaded from the IST World project portal: <http://ist-world.dfki.de/>.

4.1.1. XML Example Record for Person

Download: <http://ist-world.dfki.de/downloads/xmlExamples/Person.xml>

```
<?xml version="1.0" encoding="UTF-8"?>
<instances xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://ist-world.dfki.de/downloads/xmlSchemas/Person.xsd">
  <Person>
    <personId>http://www.dfki.de/~hansu/</personId>
    <URI>http://www.dfki.de/~hansu/</URI>
    <familyNames>Uszkoreit</familyNames>
    <firstNames>Hans</firstNames>
    <sex>male</sex>
    <qualification>Professor</qualification>
    <academicTitle>Prof.</academicTitle>
    <personResearchInterestTranslation languageCode="EN">Language Technology</personResearchInterestTranslation>
    <addressline1>DFKI GmbH</addressline1>
    <addressline2>Stuhlsatzenhausweg 3</addressline2>
    <addressline3>Building 43.8</addressline3>
    <email>sek-hu@dfki.de</email>
    <telephone>+496813025282</telephone>
    <fax>+6813025388</fax>
    <postCode>66123</postCode>
    <cityTown>Saarbrücken</cityTown>
    <stateOfCountry>Saarland</stateOfCountry>
    <countryCode>DE</countryCode>
    <expertiseAndSkillId role="teaching">http://www.coli.uni-saarland.de/~hansu/</expertiseAndSkillId>
    <orgUnitId role="affiliatedWith">http://www.dfki.de/</orgUnitId>
    <orgUnitId role="headOfDepartment">http://www.dfki.de/lt/</orgUnitId>
    <orgUnitId role="affiliatedWith">http://www.coli.uni-sb.de/</orgUnitId>
    <orgUnitId role="inAdvisoryBoardOf">http://www.acrolinx.de/</orgUnitId>
    <projectId role="investigatorOf">http://collate.dfki.de/</projectId>
    <projectId role="coordinatorOf">http://ist-world.dfki.de/</projectId>
    <resultPatentId>Reference to ResultPatent 1</resultPatentId>
    <resultPatentId>Reference to ResultPatent 2</resultPatentId>
    <resultProductId>Reference to ResultProduct 1</resultProductId>
    <resultProductId>Reference to ResultProduct 2</resultProductId>
    <resultPublicationId role="listOfPublications">http://www.coli.uni-saarland.de/~hansu/publ.html</resultPublicationId>
    <sourceDataBaseId>http://www.lt-world.org/</sourceDataBaseId>
    <identifier>lt-world</identifier>
    <keywords>language technology, computational linguistics, knowledge management</keywords>
    <notes>To be continued.</notes>
  </Person>
</instances>
```

4.1.2. XML Schema Example for Person

Download: <http://ist-world.dfki.de/downloads/xmlSchemas/OrgUnit.xsd>

```

<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XML Spy v4.4 U (http://www.xmlspy.com)-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="instances">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Person" maxOccurs="unbounded">
          <xs:complexType>
            <xs:choice maxOccurs="unbounded">
              <xs:element name="personId" type="xs:string"/>
              <xs:element name="URI" type="xs:string"/>
              <xs:element name="familyNames" type="xs:string"/>
              <xs:element name="firstNames" type="xs:string"/>
              <xs:element name="sex" type="xs:string"/>
              <xs:element name="qualification" type="xs:string" minOccurs="0"/>
              <xs:element name="academicTitle" type="xs:string" minOccurs="0"/>
              <xs:element name="honorificTitle" type="xs:string" minOccurs="0"/>
              <xs:element name="availability" type="xs:string" minOccurs="0"/>
              <xs:element name="conditions" type="xs:string" minOccurs="0"/>
              <xs:element name="personResearchInterestTranslation" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
            </xs:choice>
            <xs:element name="addressline1" minOccurs="0">
              <xs:complexType>
                <xs:simpleContent>
                  <xs:extension base="xs:string">
                    <xs:attribute name="role" type="xs:string"/>
                  </xs:extension>
                </xs:simpleContent>
              </xs:complexType>
            </xs:element>
            <xs:element name="PersonId2" minOccurs="0">
              <xs:complexType>
                <xs:simpleContent>
                  <xs:extension base="xs:string">
                    <xs:attribute name="role" type="xs:string"/>
                  </xs:extension>
                </xs:simpleContent>
              </xs:complexType>
            </xs:element>
            <xs:element name="expertiseAndSkillId" minOccurs="0">
              <xs:complexType>
                <xs:simpleContent>
                  <xs:extension base="xs:string">
                    <xs:attribute name="role" type="xs:string"/>
                  </xs:extension>
                </xs:simpleContent>
              </xs:complexType>
            </xs:element>
            <xs:element name="skillReading" type="xs:string" minOccurs="0"/>
            <xs:element name="skillSpeaking" type="xs:string" minOccurs="0"/>
            <xs:element name="skillWriting" type="xs:string" minOccurs="0"/>
            <xs:element name="facilityId" type="xs:string" minOccurs="0"/>
            <xs:element name="identifier" type="xs:string"/>
            <xs:element name="keywords" type="xs:string"/>
            <xs:element name="notes" type="xs:string" minOccurs="0"/>
            <xs:element name="prizeAwardId" type="xs:string" minOccurs="0"/>
            <xs:element name="resultPatentId" type="xs:string" minOccurs="0"/>
            <xs:element name="resultProductId" type="xs:string" minOccurs="0"/>
            <xs:element name="sourceDataBaseId" type="xs:string"/>
            <xs:element name="orgUnitId" minOccurs="0">
              <xs:complexType>
                <xs:simpleContent>
                  <xs:extension base="xs:string">
                    <xs:attribute name="role" type="xs:string"/>
                  </xs:extension>
                </xs:simpleContent>
              </xs:complexType>
            </xs:element>
            <xs:element name="projectId" minOccurs="0">
              <xs:complexType>
                <xs:simpleContent>
                  <xs:extension base="xs:anyURI">
                    <xs:attribute name="role" type="xs:string"/>
                  </xs:extension>
                </xs:simpleContent>
              </xs:complexType>
            </xs:element>
          </xs:choice>
        </xs:complexType>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>

```

4.2. OrgUnit

XML example files and XML schemas for the IST World agent OrgUnit can be downloaded from the IST World project portal: <http://ist-world.dfki.de/>.

4.2.1. XML Example Record for OrgUnit

Download: <http://ist-world.dfki.de/downloads/xmlExamples/OrgUnit.xml>

```
<?xml version="1.0" encoding="UTF-8"?>
<instances xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://ist-world.dfki.de/downloads/xmlSchemas/OrgUnit.xsd">
  <OrgUnit type="academic">
    <orgUnitId>http://www.dfki.de/</orgUnitId>
    <URI>http://www.dfki.de/</URI>
    <acronym>DFKI</acronym>
    <orgUnitName languageCode="DE">Deutsches Forschungszentrum für Künstliche Intelligenz GmbH</orgUnitName>
    <orgUnitNameTranslation languageCode="EN">German Research Center for Artificial Intelligence</orgUnitNameTranslation>
    <activity languageCode="DE">Beschreibung der Aktivitäten des DFKI.</activity>
    <activityTranslation languageCode="EN">Description of DFKI activities.</activityTranslation>
    <orgUnitResearchInterestTranslation languageCode="EN">Innovative Software</orgUnitResearchInterestTranslation>
    <headCount>280</headCount>
    <currency>Euro</currency>
    <addressline1 role="administrative">Erwin-Schrödinger-Strasse</addressline1>
    <addressline2 role="administrative">Building 57</addressline2>
    <email role="administrative">info@dfki.uni-kl.de</email>
    <telephone role="administrative">+49 (0) 631 205 3211</telephone>
    <fax role="administrative">+49 (0) 0631 205 3210</fax>
    <postCode role="main">67608</postCode>
    <cityTown role="administrative">Kaiserslautern</cityTown>
    <stateOfCountry role="administrative">Rheinland-Pfalz</stateOfCountry>
    <countryCode role="administrative">DE</countryCode>
    <addressline1 role="mainResearch">Stuhlsatzenhausweg 3</addressline1>
    <addressline2 role="mainResearch">Building 43.8</addressline2>
    <email role="mainResearch">info@dfki.de</email>
    <telephone role="mainResearch">+49 (0)681 302 5151</telephone>
    <fax role="mainResearch">+49 (0)681 302 5341</fax>
    <postCode role="mainResearch">66123</postCode>
    <cityTown role="mainResearch">Saarbrücken</cityTown>
    <stateOfCountry role="mainResearch">Saarland</stateOfCountry>
    <countryCode role="mainResearch">DE</countryCode>
    <orgUnitId2 role="hasShareholder">http://www.daimlerchrysler.de/</orgUnitId2>
    <orgUnitId2 role="hasShareholder">http://www.telekom.de/</orgUnitId2>
    <orgUnitId2 role="hasSpinn-Off">http://www.xtramind.com/</orgUnitId2>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/lt.en.html</expertiseAndSkillId>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/dmas.en.html</expertiseAndSkillId>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/km.en.html</expertiseAndSkillId>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/iupr.en.html</expertiseAndSkillId>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/iui.en.html</expertiseAndSkillId>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/iwi.en.html</expertiseAndSkillId>
    <expertiseAndSkillId role="hasResearchFocus">http://www.dfki.de/web/research/zmmi.en.html</expertiseAndSkillId>
    <prizeAwardId>http://www.dfki.de/zkp_eng/</prizeAwardId>
    <serviceId role="newsletter">http://www.dfki.de/web/news/newsletter.en.html</serviceId>
    <serviceId role="languageTechnologyPortal">http://www.lt-world.org/</serviceId>
    <sourceDataBaseId>http://www.lt-world.org/</sourceDataBaseId>
    <identifier>lt-world</identifier>
    <keywords>artificial intelligence, knowledge management, language technology</keywords>
    <notes>Entry to be completed.</notes>
  </OrgUnit>
</instances>
```

4.2.2. XML Schema Example for OrgUnit

Download: <http://ist-world.dfki.de/downloads/xmlSchemas/OrgUnit.xsd>

```
<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XML Spy v4.4 U (http://www.xmlspy.com)-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="instances">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="OrgUnit" maxOccurs="unbounded">
          <xs:complexType mixed="true">
            <xs:choice minOccurs="0" maxOccurs="unbounded">
              <xs:element name="orgUnitId" type="xs:string"/>
              <xs:element name="URI" type="xs:string"/>
              <xs:element name="acronym" type="xs:string" minOccurs="0"/>
              <xs:element name="orgUnitName">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="orgUnitNameTranslation" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="activity" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="activityTranslation" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="orgUnitResearchInterestTranslation" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              .
              .
              .
              <xs:element name="expertiseAndSkillId" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="role" type="xs:string"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="identifier" type="xs:string"/>
              <xs:element name="keywords" type="xs:string"/>
              <xs:element name="notes" type="xs:string" minOccurs="0"/>
              <xs:element name="orgUnitId2" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="role" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="prizeAwardId" type="xs:string" minOccurs="0"/>
              <xs:element name="resultPatentId" type="xs:string" minOccurs="0"/>
              <xs:element name="resultProductId" type="xs:string" minOccurs="0"/>
              <xs:element name="resultPublicationId" type="xs:string" minOccurs="0"/>
              <xs:element name="sourceDataBaseId" type="xs:string" minOccurs="0"/>
            </xs:choice>
            <xs:attribute name="type" type="xs:string" use="required"/>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

4.3. Project

XML example files and XML schemas for the IST World context Project can be downloaded from the IST World project portal: <http://ist-world.dfki.de/>.

4.3.1. XML Example Record for Project

Download: <http://ist-world.dfki.de/downloads/xmlExamples/Project.xml>

```
<?xml version="1.0" encoding="UTF-8"?>
<instances xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://ist-world.dfki.de/downloads/xmlSchemas/Project.xsd">
  <Project>
    <projectId>http://collate.dfki.de/</projectId>
    <URI> http://collate.dfki.de/</URI>
    <projectTitle>COLLATE</projectTitle>
    <projectTitleTranslation languageCode="EN">Computational Linguistics and Language Technology for Real Life
      Applications.</projectTitleTranslation>
    <projectAbstract>Building a competence center for language technology.</projectAbstract>
    <keywords languageCode="DE">Kompetenzzentrum Sprachtechnologie</keywords>
    <keywordsTranslation languageCode="EN">competence center language technology</keywordsTranslation>
    <startDate>2001-04-01</startDate>
    <endDate>2005-12-31</endDate>
    <ProjectId2 role="hasPart">http://www.lt-world.org/</ProjectId2>
    <ProjectId2 role="hasPart">http://www.lt-demo.org/</ProjectId2>
    <ProjectId2 role="hasPart">http://www.lt-eval.org/</ProjectId2>
    <fundingProgrammeId role="BMBF">http://www.bmbf.de/</fundingProgrammeId>
    <orgUnitId role="isCoordinatedBy">http://www.dfki.de/</orgUnitId>
    <eventId role="hasOrganised">http://www.lt-cc.org/lt_summit.html</eventId>
    <orgUnitId role="isPartner">http://www.uni-sb.de/</orgUnitId>
    <personId role="investigatedBy">http://www.dfki.de/~hansu/</personId>
    <personId role="investigatedBy">http://www.dfki.de/~wahlster/</personId>
    <personId role="investigatedBy">http://www.coli.uni-saarland.de/~pinkal/</personId>
    <personId role="hasParticipant">http://www.dfki.de/~declerck/</personId>
    <personId role="hasParticipant">http://www.dfki.de/~brigitte/</personId>
    <personId role="hasParticipant">http://www.dfki.de/~gulrajani/</personId>
    <personId role="hasParticipant">http://www.dfki.de/~kekl/</personId>
    <serviceId role="informationPortal">http://www.lt-world.org/</serviceId>
    <serviceId role="germanLTDemonstrationCenter">http://www.lt-demo.org/</serviceId>
    <equipmentId role="mobileDemonstrationCenter">http://www.lt-demo.org/</equipmentId>
    <sourceDataBaseId>http://www.lt-world.org/</sourceDataBaseId>
    <identifier>lt-world</identifier>
    <notes>To be continued.</notes>
  </Project>
</instances>
```

4.3.2. XML Schema Example for Project

Download: <http://ist-world.dfki.de/downloads/xmlSchemas/Project.xsd>

```
<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XML Spy v4.4 U (http://www.xmlspy.com)-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="instances">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Project" maxOccurs="unbounded">
          <xs:complexType>
            <xs:choice maxOccurs="unbounded">
              <xs:element name="projectId" type="xs:string"/>
              <xs:element name="URI" type="xs:string"/>
              <xs:element name="projectTitle" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" />
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="projectTitleTranslation" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="projectAbstract" type="xs:string" minOccurs="0"/>
              <xs:element name="projectAbstractTranslation" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="keywords">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="languageCode" type="xs:string" use="required"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              .
              .
              .
              <xs:element name="ProjectId2" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="role" type="xs:string"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="fundingProgrammeId" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="role" type="xs:string" />
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="facilityId" type="xs:string" minOccurs="0"/>
              <xs:element name="identifier" type="xs:string"/>
              <xs:element name="notes" type="xs:string" minOccurs="0"/>
              <xs:element name="resultPatentId" type="xs:string" minOccurs="0"/>
              <xs:element name="resultProductId" type="xs:string" minOccurs="0"/>
              <xs:element name="resultPublicationId" type="xs:string" minOccurs="0"/>
              <xs:element name="serviceId" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="role" type="xs:string"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="sourceDataBaseId" type="xs:string"/>
            </xs:choice>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

4.4. ResultPublication

XML example files and XML schemas for the IST World context ResultPublication can be downloaded from the IST World project portal: <http://ist-world.dfki.de/>.

4.4.1. XML Example Record for ResultPublication

Download: <http://ist-world.dfki.de/downloads/xmlExamples/ResultPublication.xml>

```
<?xml version="1.0" encoding="UTF-8"?>
<instances xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://ist-world.dfki.de/downloads/xmlSchemas/ResultPublication.xsd">
  <ResultPublication type="inProceedings">
    <resultPublicationId>http://www.dfki.de/dfkibib/publications/docs/ws1305FrankA.pdf</resultPublicationId>
    <URI>http://www.dfki.de/dfkibib/publications/</URI>
    <reference>http://www.dfki.de/lt/publications/bibtex.php3?id=725</reference>
    <referenceType>BibTex</referenceType>
    <resultPublicationDate>2005</resultPublicationDate>
    <textId>http://www.dfki.de/dfkibib/publications/docs/ws1305FrankA.pdf</textId>
    <countryCode>US</countryCode>
    <sourceDataBaseId>http://www.lt-world.org/</sourceDataBaseId>
    <identifier>lt-world</identifier>
    <notes>To be continued.</notes>
  </ResultPublication>
</instances>
```

4.4.2. XML Schema Example for ResultPublication

Download: <http://ist-world.dfki.de/downloads/xmlSchemas/ResultPublication.xsd>

```
<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XML Spy v4.4 U (http://www.xmlspy.com)-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="instances">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="ResultPublication" maxOccurs="unbounded">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="resultPublicationId" type="xs:string"/>
              <xs:element name="URI" type="xs:string"/>
              <xs:element name="reference" type="xs:string"/>
              <xs:element name="referenceType" type="xs:string"/>
              <xs:element name="resultPublicationDate" type="xs:short"/>
              <xs:element name="resultPublicationId2" minOccurs="0">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute name="role" type="xs:string"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              <xs:element name="textId" type="xs:string"/>
              <xs:element name="countryCode" type="xs:string"/>
              <xs:element name="sourceDataBaseId" type="xs:string"/>
              <xs:element name="identifier" type="xs:string"/>
              <xs:element name="notes" type="xs:string"/>
            </xs:sequence>
            <xs:attribute name="type" type="xs:string" use="required"/>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

5. Mappings from XML to RDBMS

Mappings needed for the import of XML datasets into the CERIF-based IST World data repository:

CERIF table	CERIF column	CERIF ontology propertyname
OrgUnit	name	orgUnitName
Content	name	contentName
Content	translation	contentTranslation
EquipmentName	name	equipmentName
EquipmentDescription, EquipmentName	translation	equipmentTranslation
EquipmentDescription	description	equipmentDescription
EventDescription	description	eventDescription
EventDescription	translation	eventTranslation
EventName	name	eventName
EventName	translation	eventNameTranslation
ExpertiseAndSkillDescription	description	expertiseAndSkillDescription
ExpertiseAndSkillDescription, ExpertiseAndSkillName	translation	expertiseAndSkillDescriptionTranslation
ExpertiseAndSkillName	name	expertiseAndSkillName
ExpertiseAndSkillName	translation	expertiseAndSkillNameTranslation
FacilityDescription	description	facilityDescription
FacilityDescription	translation	facilityDescriptionTranslation
FacilityName	name	facilityName
FacilityName	translation	facilityNameTranslation
FundingProgrammeName	name	fundingProgrammeName
FundingProgrammeName	translation	fundingProgrammeNameTranslation
FundingProgramme_Description	description	fundingProgrammeDescription
FundingProgramme_Description	translation	fundingProgrammeDescriptionTranslation
OrgUnitName	name	orgUnitName
OrgUnitName	translation	orgUnitNameTranslation
OrgUnitResearchActivity	translation	ActivityTranslation
OrgUnitResearchInterest	translation	orgUnitResearchInterestTranslation

D3.1: Data Import/Export Specification as XML Schema

CERIF table	CERIF column	CERIF ontology propertyname
PersonResearchInterest	translation	personResearchInterestTranslation
PersonResearchInterestDescription	description	personResearchInterestDescription
PersonResearchInterestDescription	translation	PersonResearchInterestDescriptionTranslation
ProjectAbstract	abstract	projectAbstract
ProjectAbstract	translation	projectAbstractTranslation
ProjectKeywords	translation	keywordsTranslation
ProjectTitle	title	projectTitle
ProjectTitle	translation	projectTitleTranslation
ResultPatentAbstract	abstract	resultPatentAbstract
ResultPatentAbstract	translation	resultPatentAbstractTranslation
ResultPatentTitle	title	resultPatentTitle
ResultPatentTitle	translation	resultPatentTitleTranslation
ResultProductDescription	description	resultProductDescription
ResultProductDescription	translation	resultProductTranslation
ResultProductName	name	resultProductName
ResultProductName	translation	resultProductNameTranslation
ResultPublicationTitle	title	resultPublicationTitle
ResultPublicationTitle	translation	resultPublicationTitleTranslation
ResultPublicatin_Content	description	resultPublication_ContentDescription
ServiceDescription	description	ServiceDescription
ServiceDescription	translation	ServiceDescriptionTranslation

6. Bibliography

- [1] Anne Asserson, Keith G. Jeffery, Andrei Lopatenko. CERIF: Past, Present and Future: An Overview. In Proceedings: CRIS2002 – 6th International Conference on Current Research Information Systems, August 29 - 31, Kassel, Germany. <http://www.ub.uib.no/avdeling/fdok/cris/taskgroups/CERIFPastPresentFuture200205021.pdf>

- [2] Online Documentation: CERIF 2004 - Full Data Model - Release 1.1. http://www.edward.grabczewski.btinternet.co.uk/CERIF/CERIF2004/WWW_CERIF2004_FDM_R1/Local/CERIF2004_FDM_R1Document.htm

- [3] Abiteboul, Serge; Buneman, Peter; Suciu, Dan: Data on the Web – From Relations to Semistructured Data and XML. Morgan Kaufmann Publishers, 2000.