

**510035**

**CACAO**

## **User Requirements**

<b>Deliverable number</b>	<i>D.7.1</i>
<b>Dissemination level</b>	<i>Public</i>
<b>Delivery date</b>	<i>18 April 2008</i>
<b>Status</b>	<i>Final</i>
<b>Author(s)</b>	<i>Claude Roux</i>



***eContentplus***

This project is funded under the *eContentplus* programme<sup>1</sup>,  
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

---

<sup>1</sup> OJ L 79, 24.3.2005, p. 1.

<b>1</b>	<b>GOAL</b>	<b>3</b>
1.1	PROTOCOLS	3
1.2	METADATA	3
1.3	SOME DEFINITIONS	4
1.3.1	<i>Subject Heading</i>	4
1.3.2	<i>Subject Heading List</i>	4
1.3.3	<i>Authority File</i>	4
1.3.4	<i>Classification Scheme or Classification System</i>	4
1.3.5	<i>Bibliographic Format</i>	5
1.3.6	<i>Cataloging Rules (or Catalog Code)</i>	5
1.4	SUBJECT HEADING AND CLASSIFICATION SCHEME	6
1.5	AGGREGATION	6
1.6	MOTIVATION	6
1.7	IMPLEMENTATION	6
1.8	BIBLIOTHÈQUE NATIONALE DE FRANCE (BNF)	8
1.9	PROTOCOLE	8
1.10	METADATA	8
1.11	CLASSIFICATION	8
1.12	AGGREGATION	8
<b>2</b>	<b>CITE DES SCIENCES ET DE L'INDUSTRIE (CSI)</b>	<b>8</b>
2.1	PROTOCOLS	8
2.2	METADATA	8
2.3	CLASSIFICATION	9
2.4	SPECIFIC LANGUAGE RESOURCES	9
2.5	AGGREGATION	9
<b>3</b>	<b>LINGUISTICS RESEARCH INSTITUTE – HUNGARIAN INSTITUTE (LRI)</b>	<b>9</b>
3.1	PROTOCOLS	9
3.2	METADATA	9
3.3	CLASSIFICATION	10
3.4	SPECIFIC LANGUAGE RESOURCES	10
3.5	AGGREGATION	10
<b>4</b>	<b>FREE UNIVERSITY OF BOLZANO</b>	<b>10</b>
4.1	PROTOCOLS	10
4.2	METADATA	10
4.3	CLASSIFICATION	11
4.4	SPECIFIC LANGUAGE RESOURCES	11
4.5	AGGREGATION	11
<b>5</b>	<b>KORNIK</b>	<b>12</b>
5.1	PROTOCOLS	12
5.2	METADATA	12
5.3	CLASSIFICATION	12
5.4	SPECIFIC LANGUAGE RESOURCES	12
5.5	AGGREGATION	12
<b>6</b>	<b>GOETTINGEN STATE AND UNIVERSITY LIBRARY</b>	<b>13</b>
6.1	PROTOCOLS	13
6.2	METADATA	13
6.3	CLASSIFICATION	13
6.4	SPECIFIC LANGUAGE RESOURCES	14
6.5	AGGREGATION	14
<b>7</b>	<b>SUMMARY TABLE</b>	<b>15</b>

## WP7 Business activities

Work package number :	7.1	Start date:	0	End date:	24					
Work package title:	<b>Business activities</b>									
Applicants involved:	1	2	3	4	5	6	8			

### 1 Goal

This delivery is dedicated to gathering the user requirements from the main libraries of their own country. The goal of this specific delivery consists of answering questions about metadata, standards and protocols. We have collected data from the different partners about the following points:

- Protocols
- Metadata
- Classification
- Specific Language Resources
- Aggregation

We provide below a brief description of each of these points.

#### 1.1 Protocols

The library catalogues are accessed through different protocols, such as:

- Z39.50: which the most ancient. It was defined in an era of character terminals. The Z39.50 has also be enriched with a complete query language called: *Common Query Language*.
- SRU/SRW: These are two WEB protocols. SRU embeds the query into the *URL* while SRW is based on *SOAP*.
- OAI-PMH (*Open Archives Initiative's Protocol for Metadata Harvesting*): harvests (or collects) the metadata descriptions from the archive records so that services can be built using metadata from many archives. The goal of CACAO is to have most of the libraries converging to this standard to simplify the development of an interface to library catalogues.

#### 1.2 Metadata

The metadata are the data which are associated to each document to yield different information about this document, such as the Author, the Publication Date, the Language etc. Different formats of metadata exists such as:

- Dublin Core
- Extended Dublin Core
- FRBR
- UNIMARC
- MARC21
- MARCXML

### **1.3 Some Definitions**

#### **1.3.1 Subject Heading**

*German: Schlagwort*

"The most specific word or phrase that describes the subject, or one of the subjects, of a work, selected from a list of preferred terms (controlled vocabulary) and assigned as an added entry in the bibliographic record to serve as an access point in the library catalog. A subject heading may be subdivided by the addition of subheadings (example:

Libraries--History--20th century) or include a parenthetical qualifier for semantic clarification, as in Mice (Computers). The use of cross-references to indicate semantic relations between subject headings is called syndetic structure. The process of examining the content of new publications and assigning appropriate subject headings is called subject analysis. In the United States, most libraries use Library of Congress subject headings (LCSH), but small libraries may use Sears subject headings." Joan M. Reitz, ODLIS — Online Dictionary for Library and Information Science, [http://lu.com/odlis/odlis\\_s.cfm](http://lu.com/odlis/odlis_s.cfm)

#### **1.3.2 Subject Heading List**

*German: Schlagwortliste*

"A list of authorized controlled vocabulary terms or phrases together with any references, scope notes, and subdivisions associated with each term or phrase." (Arlene G. Taylor, Introduction to Cataloging and Classification, 10th ed.)

Examples: Schlagwortnormdatei (SWD), Soggettario italiano, Rameau, Library of Congress Subject Headings (LCSH), Medical Subject Headings (MeSH)

#### **1.3.3 Authority File**

*German: Normdatei*

An authority file is a grouping of authority records, which records the decisions made during the process of performing authority work. "An authority record contains all the forms used for a particular name, title, or subject, and usually designates one of the forms as the 'authorized' or 'default' one to use in catalog records." (Arlene G. Taylor, Introduction to Cataloging and Classification, 10th ed.)

Examples: Personennamendatei (PND), Gemeinsame Körperschaftsdatei (GKD), Library of Congress Authorities

#### **1.3.4 Classification Scheme or Classification System**

*German: Klassifikationsschema, Klassifikationssystem, System, Systematik*

"An organized framework for the systematic organization of knowledge, usually organized by subject." (Arlene G. Taylor, *Introduction to Cataloging and Classification*, 10th ed.)

"A list of classes arranged according to a set of pre-established principles for the purpose of organizing items in a collection, or entries in an index, bibliography, or catalog, into groups based on their similarities and differences, to facilitate access and retrieval. In the United States, most library collections are classified by subject.

Classification systems can be enumerative or hierarchical, broad or close. In the United States, most public libraries use Dewey Decimal Classification, but academic and research libraries prefer Library of Congress Classification. See also: Classification Society of North America, Colon Classification, and notation." Joan M. Reitz, ODLIS — Online Dictionary for Library and Information Science, [http://lu.com/odlis/odlis\\_c.cfm](http://lu.com/odlis/odlis_c.cfm)

Examples: Library of Congress Classification (LCC), Dewey Decimal Classification (DCC), Universal Decimal Classification (UDC), Göttinger Online-Klassifikation (GOK), Basisklassifikation (BK)

### 1.3.5 Bibliographic Format

*German: Katalogisierungsformat*

"The standardized sequence and manner of presentation of the data elements constituting the full description of an item in a specific cataloging or indexing system. The machine-readable MARC record format has become the standard for library catalogs in many countries of the world." Joan M. Reitz, ODLIS — Online Dictionary for Library and Information Science, [http://lu.com/odlis/odlis\\_b.cfm](http://lu.com/odlis/odlis_b.cfm)

Examples: MARC21, UNIMARC, USMARC, Pica, International Standard Bibliographic Description (ISBD), MAB (See more examples at: <http://www.ifla.org/VI/3/p1996-1/appx-h.htm>)

### 1.3.6 Cataloging Rules (or Catalog Code)

*German: Katalogisierungsregeln*

"A detailed set of rules for preparing bibliographic records to represent items added to a library collection, established to maintain consistency within the catalog and between the catalogs of libraries using the same code. In the United States, Great Britain, and Canada, libraries use the Anglo-American Cataloguing Rules developed jointly by the American Library Association, Library Association (UK), and Canadian Library Association. Synonymous with cataloging code." Joan M. Reitz, ODLIS — Online Dictionary for Library and Information Science, [http://lu.com/odlis/odlis\\_c.cfm](http://lu.com/odlis/odlis_c.cfm)

Examples: Anglo-American Cataloguing Rules (AACR2), Regeln für die Alphabetische Katalogisierung (RAK), Regeln für die Alphabetische Katalogisierung an wissenschaftlichen Bibliotheken (RAK-WB) (See more examples at: <http://www.ifla.org/VI/3/p1996-1/appx-h.htm>)

#### **1.4 Subject Heading and Classification Scheme**

Libraries utilize existing subject heading hierarchies or classification schemes (see above) to classify their documents, according to the domain to which the book belongs.

For instance:

- DDC is a Classification Scheme
- LCC is a Classification Scheme
- UDC is a Classification Scheme
- LCSH is Subject Heading list
- MESH is Subject Heading list

#### **1.5 Aggregation**

Aggregation is the harvesting of data from different sources, in this case from different library catalogues. The CACAO consortium proposes Dspace (<http://www.dspace.org/>) as a way to harmonize queries across different catalogues.

#### **1.6 Motivation**

From the beginning in the CACAO project, we have focused on a business approach. Our main objective is to design a system which would both simplify and normalize the access to library catalogues no matter their access protocols, the document metadata format or the languages. We have established some relations from the beginning with the BNF (Bibliothèque Nationale de France) which has a long experience in implementing Web interfaces to their catalogues (Gallica). The BNF does not belong to the CACAO project, however we have kept strong contacts with them in order to keep our implementation in line with their own goals. For this reason, the BNF appears among the different organizations from which we have requested more detailed information about their communication protocols and metadata. We have also established some strong connections with the European Digital Library project (EDL), which is a natural recipient to test and use our tools.

#### **1.7 Implementation**

With such a huge diversity in protocols and metadata standards, there is a need to unify all these protocols and metadata under one specific umbrella. We suggest Dspace as an answer to this issue. Dspace was launched by the MIT and Hewlett Packard as an open source platform to access and preserve digital content. It is highly customizable to fit the needs of any clients. It has one of the largest communities of users and developers worldwide. It has been successfully implemented in many institutions such as Toronto University or the MIT.

A Dspace platform will be implemented with some specific features which will be specific to CACAO, such as query translations or query expansions. The goal of Dspace in our architecture is to get all libraries to converge to one single entry-point, without the need for these libraries to give up their own protocols and meta-data. The consortium will provide the necessary gateways or translation tools to plug their own system into a Dspace platform. These gateways for instance will know how to connect to an external system which is based

on Z39.50. These gateways will also provide the necessary tools to interpret the different metadata. A user will query through Dspace to access the catalogues without knowing anything about which protocols to comply with or which metadata to keep track of. The Dspace platform will provide a unique portal to many incompatible libraries' access points.

If a new customer is interested in using our technology, then two cases can occur: either the client already uses one of the protocols, for which the consortium has already built a gateway, or we need to design and implement a new gateway. In both cases, the Dspace platform will remain intact and the customer won't need to modify her/his own architecture.

## **1.8 Bibliothèque Nationale de France (BNF)**

### **1.9 Protocole**

The BnF uses both OAI-PMH and Z39.50.

### **1.10 Metadata**

- InterMarc which is BnF marc21 implementation
- unimarc
- ISBD
- ISO2709 used for certain online transfers
- Dublin Core : The whole BnF catalogue is accessible online through OAI and the OAI-PMH protocole.

All the BnF bibliographic description are accessible are « harvestable » by everyone in a Dublin Core format. Collections are visible through the OAI sets.

### **1.11 Classification**

The BnF offers different Subject Heading List such as:

- Rameau, which is linked to LCSH. Rameau provides some authority list for Geographical names, Commercial Brand and Standardized Book Titles.

And also Classification Scheme such as:

- Dewey Decimal Classification

### **1.12 Aggregation**

The BnF does not have any multilingual resources. However the BnF does have some specific collections on Medieval Literature, Mathematics, Geography, which would be mainly available in French.

The Search Engine which the BnF uses is *Lucene*. However, the BnF has started a new project on DSpace on *Fedora*.

## **2 Cité des Sciences et de l'Industrie (CSI)**

### **2.1 Protocols**

They only use Z39.50, however they have programmed a portage to OAI-PMH. They need to acquire the OAI module for ALEPH.

### **2.2 Metadata**

Their metadata are displayed in Dublin Core and MARCXML.

### **2.3 Classification**

They use their own classification schema (CSI classification), which is available.

### **2.4 Specific Language Resources**

They only have a specific collection on mathematics.

### **2.5 Aggregation**

They do not have any experience with DSpace.

## **3 Linguistics Research Institute – Hungarian Institute (LRI)**

### **3.1 Protocols**

Their electronic library has Z39.50 and OAI-PMH protocol or server.  
They were the first Hungarian institution to move to a OAI-PMH server.

### **3.2 Metadata**

Their metadata can be displayed in many different formats:

Examples might be:

- a) Dublin Core
- b) Extended Dublin Core

They use the DC format in the source of our open page of documents, e.g.

<http://mek.oszk.hu/03900/03966>

- c) FRBR
- d) UNIMARC
- e) MARC21
- f) MARCXML

They can also display their metadata:

- In HUNMARC <http://mek.oszk.hu/03900/03966/hunmarc.html>
- In USMARC <http://mek.oszk.hu/03900/03966/usmarc.html>
- In XML <http://mek.oszk.hu/03900/03966/index.xml>

Their entire catalogue is downloadable in many formats actualized daily:

- In HTML eg. For Excel
- In field tagged
- In XML
- In USMARC (in one file and separate too one record one file)

<http://mek.oszk.hu/html/export.html>

### 3.3 Classification

They use a simple, proprietary Classification Schema with 3 level., which has been translated in English.

<http://mek.oszk.hu/html/muszakieng.html>

<http://mek.oszk.hu/html/tarsadalomeng.html>

<http://mek.oszk.hu/html/humaneng.html>

<http://mek.oszk.hu/html/kezikonyveng.html>

### 3.4 Specific Language Resources

They use thesauri, especially a general thesaurus which has been developed by the national library: <http://mek.oszk.hu/00700/00769>

Their Graphical Interfaces have been all translated into English, and are all available both in Hungarian and in English.

Their acquisition policy does not put a high priority on science documents, which represent only 10% of their all catalogue.

### 3.5 Aggregation

Their opinion is that Dspace or Greenstone would be appropriate for managing harvested data.

## 4 Free University of Bolzano

### 4.1 Protocols

Bolzano utilizes the Z39.50 format. They are ready to converge to OAI-PMH.

### 4.2 Metadata

They used as a format: UNIMARC.

Bolzano is ready to disclose the following data:

Available Metadata	Definition
Type of document	Audio, Multi-level description, Loose-leaf publication Monography, Game, Video etc.
Personal name (author)	Person responsible for the creation of a work
Personal name (contributor)	Person contributing to the creation of content of a work (editor, translator, author in case of more than 3 authors...)
Art of corporate body	Synonyms, Formerly etc.

Title proper	Title in the form of the document
Uniform title	Used when a work is entered directly under title and the work has appeared under varying titles, necessitating that a particular title be chosen to represent the work
Parallel title	The title proper in another language
Statement of responsibility	Information relating to the identification and/or function of any persons or corporate bodies responsible for or contributing to the creation of the intellectual content of a work

### 4.3 Classification

Bolzano uses the following Subject Headings Lists:

- Writing in Italian: Soggettario per i cataloghi delle biblioteche italiane (SCBI)
- Writing in German: SWD Schlagwortnormdatei (SWSD)
- Writing in English: Library of Congress Subject headings (LCSH)
- Writing in other languages: one of the above mentioned

They also use the following Classification Scheme for the German language:

RVK: Regensburger Verbundklassifikation (<http://www.bibliothek.uni-regensburg.de/Systematik/systemat.html>)

RVK – SWD mapping is available in Word format.

N.B.: The Library of Congress offers access to the correlations LC Subject Headings - Dewey Classification Numbers and LC Classification - Dewey Classification Numbers in Classification Web (<http://www.loc.gov/cds/classweb/>). Contents are not freely available

### 4.4 Specific Language Resources

Bolzano has access to many resources, which for most of them are freely accessible on the Web.

Bolzano also follows the following policy which to index writings in their own language. Bolzano does not provide any specific resources for the Medieval, Mathematics and Geographic domains.

### 4.5 Aggregation

They do not have any specific experience in using DSpace.

## 5 KORNİK

### 5.1 *Protocols*

All digital assets from Kornik Library are available within Wielkopolska Digital Library ([www.wbc.poznan.pl](http://www.wbc.poznan.pl)) and can be accessed through OAI-PMH protocol. Catalog system might be accessed through Z39.50. Our will is to use OAI protocols standards.

### 5.2 *Metadata*

For digital objects metadata are stored in Dublin Core Metadata Schema v.1.1 Catalog system is based on MARC21 metadata schema. There are no detailed preferences for common set of attributes disclosed per entry/collection/library.

### 5.3 *Classification*

Kornik Library does not use any classification system.

### 5.4 *Specific Language Resources*

Catalog system in Kornik Library (MARC21) and metadata for digital objects (DC) are based on National Library subject headings authority files (<http://mak.bn.org.pl/info/info19a.htm>) For catalog system all metadata are stored only in national language. For digital objects (within Wielkopolska Digital Library) it is possible to extend metadata for additional languages – currently there are no entries for other languages than Polish Kornik Library posses resources (digital objects or catalog entries) for Medieval Literature and Mathematics (Teofil Żebrowski and unpublished manuscripts of Jozef Maria Hoene-Wronski).

NOTICE: Kornik Library posses lots of XIX century books, manuscripts, newspapers, maps and others which could be provided for a **Historical portal** (broader than Medieval only) or a **Literature portal**

### 5.5 *Aggregation*

Polish digital resources are aggregated (through OAI-PMH) within Federation of Digital Libraries (<http://fbc.pionier.net.pl/>) based on dLibra Digital Library Framework (<http://dlibra.psnc.pl/>). Kornik Library is publishing their digital objects in Digital Library of Wielkopolska which is based also on dLibra (with multilingual GUI). It can be considered also for CACAO

## 6 Goettingen State and University Library

### 6.1 Protocols

- The library catalogue is accessible by Z39.50, SRU and OAI-PMH (at the moment under revision)
- A document repository and a special Database of retrodigitised journals are available via OAI-PMH

### 6.2 Metadata

- The library catalogue interfaces offer
  - o Dublin Core
  - o Unimarc
  - o Marc21
  - o Pica (internal format, richest source of information)

### 6.3 Classification

actively used in the library

- GOK (Göttinger Online Klassifikation)
  - o universal home grown Classification system
  - o German captions
- BK (Basisklassifikation)
  - o universal classification widely adopted in the library union Gemeinsamer Bibliotheksverbund GBV
  - o German captions
- GFDC (Global Forest Decimal Classification)
  - o only used for forestry related literature
  - o German and English captions, French and Spanish version are work in progress at the moment, russian version is discussed by the board
- Dewey Decimal Classification
  - o universal system, but used only for a small number of subjects and/or media types
  - o multilingual

present by import from foreign sources

- Library of Congress Subject Classification (LCC)
- Dewey Decimal Classification (DDC)
- Regensburger Verbundklassifikation (RVK)

Mappings

- Mappings between GOK and DDC are available of selected parts of the systems (produced and owned by UGOE)
- Mappings between SWD Terms and DDC Classes are developed in the Criss-Cross Project at the German National Library
- There are other mapping activities, especially in the domain of social sciences at Gesis, Bonn

#### **6.4 *Specific Language Resources***

The items are catalogued in their original language or in transcriptions.

Special Subject Collections are present in the domains Anglo-American Culture, Estonian, Finno-Ugric, Korean and Altaic and Paleoasiatic Languages and Culture, Geography, Thematic Maps, Geophysics, Astronomy, Pure Mathematics, Forestry and library, book and information sciences.

#### **6.5 *Aggregation***

There is some experience with DSpace at the Goettingen Library: it is used as a repository for born digital objects while their metadata are located in the library catalogue.

## 7 Summary Table

Institution	Protocol	Metadata	Classification	Subject Heading	Domains	Multi-lingual	Search Engine
BnF	OAI-PMH Z39.50	Intermarc Unimarc ISBD ISO2709 Dublin Core	Dewey	Rameau	Medieval Geography Mathematics	No	Lucene DSpace
CSI	Z39.50	Dublin Core MARCXML	CSI (proprietary)		Mathematics	No	No
LRI	OAI-PMH Z39.50	Dublin Core Extended Dublin Core FRBR UNIMARC MARC21 MARCXML	proprietary		Limited	Yes	No
Kornik	OAI-PMH Z39.50	Dublin Core MARC 21	No	National Library subject headings authority files	Mathematics Medieval	No	Dlibra
Bolzano	Z39.50	UNIMARC	RVK (for German language materials)	SCBI SWSD LCSH	No	Yes	No
Goettingen	Z39.50 SRU OAI-PMH	Dublin Core MARC 21 Unimarc Pica	GOK BK GFDC also LCC, DDC, RVK		<ul style="list-style-type: none"> <li>• Anglo-American Culture</li> <li>• Estonian</li> <li>• Finnougristic</li> <li>• Korean and Altaic and Paleoasiatic Languages and Culture</li> <li>• Geography</li> <li>• Thematic Maps</li> <li>• Geophysics</li> <li>• Astronomy</li> <li>• Pure Mathematics</li> <li>• Forestry</li> <li>• Library, Book and Information Sciences</li> </ul>	Yes	Pica (index of the library catalogue) Lucene for a sperate collection