

Metadata Mapping and Application Profiles. Approaches to providing the Cross-searching of Heterogeneous Resources in the EU Project Renardus

Heike Neuroth <neuroth@mail.sub.uni-goettingen.de>
Lower Saxony State and University Library Göttingen, Germany
and

Traugott Koch <Traugott.Koch@ub2.lu.se>
NetLab - Lund University Library Development Department, Sweden

Abstract

This paper presents the approach and results of a mapping process to define a common metadata format for cross-searching distributed and heterogeneous subject gateways in the EU project Renardus. The outcome in Renardus is a well defined data model with semantic and syntactical definitions of each metadata element. It results in richer and semantically controlled cross-searching. The metadata elements are mainly based on Dublin Core, some further elements and qualifiers are defined in Renardus namespaces. A collection level description schema has also been developed to allow a well structured description of each participating gateway. A Renardus application profile is under development. The Renardus experience and some of its solutions may well be a good basis for similar interoperability efforts.

Keywords: Renardus, Application Profile, Namespace, Metadata Mapping Processes, Collection Level Description, Subject Gateway.

1. Introduction

Renardus [1] is funded (January 2000 – June 2002) through the Information Society Technologies (IST) Programme, 'Promoting a User-friendly Information Society' [2]. This is a major theme of the European Union's 5th Framework Programme [3]. The twelve Renardus partners represent European library and other information-related communities from Finland, Denmark, Sweden, UK, The Netherlands, France, and Germany. They work at the forefront of developing quality-controlled subject gateways, providing access to selected quality resources for the academic and research communities.

The aim of the Renardus project is to provide users with integrated access, through a single interface, to high-quality Internet resources and other

Internet-based, distributed services. The approach being taken is to provide access to distributed subject gateways (high quality metadata collections) that will allow the integrated searching and browsing of distributed resource collections. Further goals are to develop and define organisational models, business models, technical solutions and metadata standards. This paper intends to provide an overview of the development of the Renardus data model, the Renardus namespaces, the Renardus application profile, and the Renardus collection level description.

2. Quality-controlled Subject Gateways

The Renardus broker intends to bring together heterogeneous and distributed quality-controlled subject gateways.

Quality-controlled subject gateways are characterized by high standards of quality control and a rich set of metadata which enables users to search across several metadata elements. Koch [4] defines a quality-controlled subject gateway as follows: "Quality-controlled subject gateways are Internet-services which apply a rich set of quality measures to support systematic resource discovery. Considerable manual effort is used to secure a selection of resources which meet quality criteria and to display a rich description of these resources with standards-based metadata. Regular checking and updating ensure good collection management. A main goal is to provide a high quality of subject access through indexing resources using controlled vocabularies and by offering a deep classification structure for advanced searching and browsing."

The following elements characterize a typical quality-controlled subject gateway:

- selection and collection development,
- collection management,
- intellectual creation of metadata (done by experts),
- resource description (a rich and documented metadata set),

- resource indexing (using a subject classification scheme or controlled vocabulary system, etc.)

To allow interoperability with other Internet services, this recommended that gateways apply open standards.

3. Development of the Data Model

All gateways participating in the Renardus pilot broker service are quality-controlled subject gateways or resource discovery broker services. The crucial characteristics of such subject gateways are, besides selection, collection development, collection management, quality criteria, etc., that they all apply resource descriptions and subject classification to all their records. The Renardus broker cross-searches this human-created metadata, so it needs to be structured in a standardized way. So at the beginning of developing a common Renardus metadata profile there were theoretical considerations about which metadata elements are most useful for cross-searching in such a service. Several of the subject gateway services in Renardus support both English and native language descriptions of metadata elements. For this reason, the multilingual character of a broker service like Renardus needs to be considered in some way. Initial research led to a first formulation of elements that needed to be integrated into the Renardus data model (e.g. subject, description, etc.). There are important questions about how this data model would ideally have to be constructed and how this could optimally be achieved.

In order to develop a common data model (or application profile, see below) for the Renardus broker system it was necessary to undertake a very detailed analysis of the metadata formats of all participating services. Once details of all of these formats were known, it was possible to start the development of a common set of metadata and the mapping of partner gateways' metadata elements to this core set. The practical development of the data model was divided into several steps:

- The agreement of a metadata format that could form the basis of the data model, (i.e. an exchange format). In Renardus we agreed on the Dublin Core Metadata Element Set [5, 6] as the basic metadata format.
- An initial survey survey of the metadata formats looked for all metadata elements used by the participating services. This included elements that could be mapped to the agreed exchange format (Dublin Core), some qualifiers (element refinements and encoding schemes), additional domain specific elements, as well as some

administrative elements. For each element the semantics (definition) and syntax (structure) had to be defined. The answers described which code, standards, and cataloging rules are in use for each element. Information about the obligation of the element (e.g. whether it is mandatory, recommended or optional), the degree of usage of each element, use of the language qualifier (e.g. in the case of title, keywords, description etc.) and general notes are provided.

- Replies to the survey allowed the definition of a core set of elements and qualifiers that are used by most of the participating services. A ranking of these elements showed the importance of each element and its obligation. Those that were used by most partners allowed for cross-searchability in Renardus. Data elements in use by gateway services would then need to be mapped to the core Renardus data model. The process of mapping was balanced between the quality of the common metadata format and the amount of work needed to adapt a single gateway's metadata format to the common metadata format used in Renardus. For example, it was much easier to agree on the use of a code for the language element in Renardus, than on the publisher element because the latter was not used by some partners. For language, most partners either support the ISO 639 two letter code or the ISO 639 three letter code, and a mapping from the two letter code to the three letter code is relatively easy to accomplish. To come up with common cataloging rules for the publisher element, however, would be very difficult, so Renardus has not added this element to its data model so far. This means that mapping processes are much easier in the case that automated or partly automated processes can support the upgrading process.
- After defining a first version of the Renardus data model, some problems with element definitions (semantics and syntax) or qualifiers were left unsolved. A second survey was undertaken to clarify the use of certain elements and made partners aware of specific problems. These problems occurred e.g. in the context of multilinguality in Renardus (which could lead to the introduction of a language qualifier for the title, description, and subject elements) or regarding the creator element. Renardus will support the creator element, but only in the syntax of LastName, FirstName. No further information (Email, URL etc.) will be provided to begin with. In this phase it was also useful to think about common administrative elements (e.g. gateways icons for branding and URL linking to the full metadata record at the participating service) and a

metadata format for describing partners' collections (based on the collection level description schema, see below).

- One of the last steps was the development of a Renardus application profile including the namespaces which define the metadata elements used (see below).

After Renardus had defined the common metadata set, the gateway partners had to map from their formats to the agreed format.

4. Renardus Data Model

The aim of Renardus is to develop a single interface for cross-searching and cross-browsing of distributed subject gateways. Cross-searching is facilitated by the adaption of a common metadata profile, cross-browsing by mapping locally-used classification systems to a common classification system.

The Renardus data model was developed to define the mapping processes which are necessary for proper cross-searching of all resources catalogued in partners' subject gateways.

Very early in the discussion about the development of the Renardus data model it was clear that the data model should, as far as possible, be based on Dublin Core, in order to increase potential interoperability with future partners or services. Only one "content" metadata element is not a DC element nor a DC based element and this is the Country element. All other "content" metadata elements and qualifiers are based on Dublin Core and its recommendations, wherever possible. In case no encoding scheme or refinement from Dublin Core can be used, a Renardus qualifier is introduced to help define elements. These additional elements and qualifiers are part of the Renardus namespaces (see below).

The Renardus partners agreed on a minimal, common set of metadata elements which would need to be supported by each participating subject gateway. In this way, academic users can cross-search distributed collections of high-quality Internet resources. Currently the beta-version (September 2001) of the Renardus broker holds a collection of about 18,500 metadata records which can be searched via one single interface.

The minimal set of metadata contains the following Dublin Core elements: Title, Creator, Description, Subject, Identifier, Language, and Type. The only non DC "content" metadata element is Country. Further detailed investigations were focused on the following characteristics for each metadata element:

- semantic definition
- syntactic definition

- associated qualifiers (based on Dublin Core as far as possible, e.g. refinements, encoding schemes)
 - cataloging rules (e.g. for the elements creator, description, keywords)
 - namespace definition
 - the repeatability of each element
 - the form of obligation (mandatory, strongly recommended, optional)
 - language qualifiers (for title, description, subject) as a possible future implementation

Regarding some "administrative elements" the Renardus partners decided to use the elements "Full Record URL" that could lead Renardus users to the original metadata set of the local subject gateway and the element "SBIG ID" that indicates the acronym of each participating subject gateway. Information about the date of creation or updating of the metadata elements will not be provided.

Some consideration about future, additional elements apply to information about rights management, terms and condition, access/restriction conditions, etc. (e.g. the DC element Rights) and the possible implementation of the DC element Publisher. But until now there are no fixed standards to use both elements in a specific way. For Rights there are no Dublin Core qualifiers available and Publisher needs further discussion and standardization of rules, etc. (cf. DCMI Agents Working Group [7]).

The following list provides an overview of the Renardus "content" metadata of the Renardus data model. More detailed and updated information can be found at [8].

The format of entry for each Renardus element looks like this:

Name	Name of metadata element
Choice of Namespace	<ul style="list-style-type: none"> ▪ DCMES version 1.1, ▪ DCMES Qualifiers (2000-07-11), ▪ Renardus Metadata Element Set = RMES version 0.1, ▪ Renardus Metadata Element Set Qualifiers = RMES Qualifiers version 0.1
Dublin Core Refinement(s)	Element refinements used in Renardus: These qualifiers make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope.
Renardus Refinement(s)	Renardus refinement, see above
Dublin Core Encoding	These qualifiers identify schemes that aid in the interpretation of an

Scheme(s)	element value. These schemes include controlled vocabularies and formal notations or parsing rules. A value expressed using an encoding scheme will thus be a token selected from a controlled vocabulary (e.g., a term from a classification system or set of subject headings) or a string formatted in accordance with a formal notation (e.g., "2000-01-01" as the standard expression of a date). If an encoding scheme is not understood by a client or agent, the value may still be useful to a human reader.
Renardus Encoding Scheme(s)	Renardus encoding scheme, see above
Form of Obligation	In the Renardus data model the obligation can be: mandatory (M), strongly recommended (R) or optional (O). Mandatory ensures that some of the elements are always supported. An element with a mandatory obligation must have a value. The strongly recommended and the optional elements should be filled with a value if the information is appropriate to the given resource or provided by a subject gateway, but if not, they can be left blank.
Repeatability	Metadata element is repeatable: yes or no

LQ "LANG"	Language Qualifier: to give information about the language of the content of a metadata field (ISO Code 639, two letter), yes , no (or possible: prototype system). Language qualifiers for the Title, Description and Subject elements will be recommended. If neither a Description nor a Subject element is available in English then the language qualifier is strongly recommended.
DC Definition	Dublin Core definition of the metadata element.
DC Comment	Dublin Core comments to this metadata element.
R Definition	Renardus definition of the metadata element.
R Comment	Renardus comments to this metadata element.

Figure 1 provides a short overview of each Renardus metadata element with indication of the kind of namespace. All namespaces (DCMES for Dublin Core Metadata Element Set simple, DCMESq for Dublin Core Metadata Element Set Qualified, RMES for Renardus Metadata Element Set simple, and RMESq for Renardus Metadata Element Set Qualified) are defined in the Renardus application profile (see below).

Element	Refinement	Scheme	Namespace	O	R	Rule
Title			DCMES	M	no	
	Title.Alternative		DCMESq	O	yes	
Creator			DCMES	R	yes	
Creator		LastName, FirstName	RMESq	R	yes	*
Description			DCMES	M	yes	
Subject			DCMES	M	yes	
Subject		LCSH, MeSH, DDC, LCC, UDC	DCMESq	R	yes	
Subject		Partners' systems	RMESq	M	yes	
Subject		Renardus-DDC	RMESq	M	yes	*
Identifier		URI	DCMESq	M	no	
Language		ISO 639-2	DMESq	R	yes	
Type			DCMES	R	yes	
Type		DCMI Type Vocabulary	DCMESq	R	yes	
Country		ISO 3166-1	RMESq	R	yes	*

Figure 1: Overview of Renardus "content" metadata elements

O = Obligation (M for mandatory, R for Recommended, and O for optional)

R = Repeatability: yes or no

Rule = Renardus specific or additional (cataloging) rules

The following list comments on some Renardus elements:

Subject:

Renardus has four different namespaces for this element. Because Renardus will also develop a cross-browsing structure based on the Dewey Decimal Classification (DDC) we are allowed (OCLC Forest Press licence) to map our local classification systems to DDC. Furthermore, it is allowed to introduce European specific captions (the verbal description of the DDC notation) instead of official DDC captions. For this reason we also distinguish between DDC and Renardus-DDC. In addition to the two Dublin Core namespaces (DCMES and DCMESq), we defined two elements based on the Renardus Metadata Element Set Qualified namespaces:

- RMESq namespace: Ren-DDC to help build up the common browsing structure in Renardus, and
- RMESq namespace: for all other encoding schemes used by Renardus partners for classification systems and controlled vocabularies

Identifier:

It could be possible that Renardus will define further refinements for this element in future (e.g. Archive, Mirror, etc.)

Type:

We are thinking of implementing the DCT2 (Dublin Core Type Vocabulary: Subtypes) list as soon as this list has been agreed.

All encoding schemes are based on international standards as far as possible (e.g. ISO standards). Only the element Country is a Renardus specific element and reflects first the publisher country and second the (so called) cultural context of the resource. This is the reason why this element is repeatable, the encoding scheme is ISO 3166-1 with some Renardus specific extensions (e.g. EU for European Union, XP for international, XF where the element is not applicable).

All these elements are used to realize cross-searching over the distributed metadata collections or for the sorting and/or filtering processes of results.

The main basic index allows a search across the elements Title, Description and Subject. By doing the mapping from partners' metadata format to the agreed Renardus metadata format it is necessary that subject gateways will provide free-text in the Description field (and not e.g. a URL) and that the subject gateways deliver some kind of subject information. Up to now it is an open question whether DDC captions will also be included in the basic index.

The cross-browsing structure is realized through a mapping of each partners' classification system to the Dewey Decimal Classification (DDC). The subject

element containing the DDC mapping is therefore mandatory.

The elements Country, Language, and Type are primarily used as search filters.

5. Namespaces, Renardus Application Profile

According to Heery & Patel [10], an application profile is a type of metadata schema which consists of metadata elements drawn from one or more namespaces combined together by implementors for a particular local application. An application profile gives information about:

- the schemas used and the incorporated elements of a (domain specific) metadata implementation,
- the policies defining how elements should be applied, and
- guidelines explaining how to use each element.

One example of such an application profile is the proposed DC Education schema [11], which consists of various DC based metadata elements, several which are defined in a DC Education namespace (e.g. Audience with the qualifier Mediator, Standard with qualifiers like Identifier, Version, etc.), and various IEEE Learning Object Metadata (LOM) elements [12] (e.g. InteractivityType).

A namespace schema contains all metadata elements defined by a managing body or registration authority (e.g. the Dublin Core Metadata Element Set is defined by the Dublin Core Metadata Initiative); application profiles are defined by implementors for their specific needs and contain one or more namespaces. The advantages of defining namespaces are that each metadata element is well defined by a managing body and can be uniquely identified. This is also true for vocabularies (e.g. Dublin Core Types). Namespaces and application profiles are expected to be registered by authority bodies (e.g. the Dublin Core registry [13] or the Schemas project registry [14]).

The Renardus project will create such an application profile [15]. The metadata elements of the Renardus data model are defined by four namespaces:

- Dublin Core Namespace: [DCMES version 1.1] Dublin Core Metadata Element Set, Version 1.1: Reference Description
- Dublin Core Qualifiers Namespace: [DCMES Qualifiers (2000-07-11)] Dublin Core Qualifiers
- Renardus Namespace: [RMES version 0.1, 2001-04-30] Renardus Metadata Element Set
- Renardus Namespace Qualifiers: [RMES Qualifiers version 0.1, 2001-04-30] Renardus Metadata Element Set Qualifiers

At the moment the Dublin Core community is working on a policy for naming namespaces and single terms (metadata elements, term of a controlled vocabulary, etc.) [16]. The encoding of the two Renardus namespaces and the application profile will be in RDF/XML. This encoding syntax is not standardized so far, although a paper by Kokkelink & Schwänzl [17] submitted to the DC Architecture working group [18] is under discussion. The RDF schema declaration for the Dublin Core Element Set 1.1 [19] and the Qualified Dublin Core Element Set (2000/03/13) [20] will have to be updated if a new standard for such a declaration is formulated.

Examples of RDF schemas for application profiles can be found at the SCHEMAS project site [21]. For example a draft RDF schema for the RSLP-CLD (see below) application profile is already available [22].

The RDF/XML encoding of the Renardus application profile is under development and will be available soon.

6. Renardus Collection Level Description (RCLD)

A simple collection description is used to describe the collections (subject gateways) of Renardus partners. This schema is based on the Collection Level Description (CLD) schema developed by the Research Support Libraries Programme (RSLP) Collection Description project [23, 24]).

The aims of Renardus to develop a Renardus Collection Level Description Element Set are:

- to support the selection of subject gateway(s) for searching,
- to provide background information about the participating subject gateways for both humans and software,
- to promote/register the individual subject gateway(s) as high quality resources in the Internet, and
- to allow software in the future to use subject information for systems in the selection of services.

The Renardus Collection Level Description schema (RCLD) is based on the RSLP schema with some additional elements and rules for several elements, some guidelines e.g. for the description field will be developed in the near future to ensure a more or less standardized form of description. The RCLD schema is based on three kinds of formats (and in this way also namespaces):

- Dublin Core (based) elements (e.g. dc:title)
- Collection Level Description elements based on the RSLP schema (e.g. cld:country)
- Renardus specific Collection Level Description elements (e.g. ren-cld:language)

All elements, except DC.Relation are mandatory, the following list enumerates all metadata elements

which are part of the Renardus Collection Level Description.

Dublin Core (based) elements:

Title: The name of the collection (in our case the name of the participating Subject Gateway in Renardus).

Identifier: An unambiguous reference to the collection within a given context (encoding scheme: URI).

Description: An account of the content of the collection (in future with a standardized structure of the content of description with information about granularity of collected resources, type of subject indexing etc.).

Language: The main language(s) of the metadata in the collection with quantitative indication (free text).

Publisher: The organization etc. who is responsible for the intellectual (not technical) distribution of the collection.

Format.Extent: The size of the collection.

Date.Issued: Date of formal issuance (e.g. publication) of the collection.

Subject: The topic of the content of the collection, main DDC captions for the subjects represented in the Subject Gateway.

Relation: A reference to a related collections.

Collection Level Description elements based on the RSLP schema:

Country: The country in which the collection is physically located.

Renardus specific Collection Level Description elements:

Acronym: The acronym of the collection.

Resource Language: Language(s) of the described resources with quantitative indication.

In addition the Renardus CLD contains several internal technical elements.

The HTML form for creating the RCLDs is based on the RSLP tool [25]. It is a WWW based form that can create CLDs encoded in RDF, RDF/XML, and text.

The files are saved locally by the subject gateways, so that each partner is able to update his description at every time. The Renardus broker gathers all descriptions and gives access to them at a prominent part of the user-interface.

7. Conclusions

Experiences with developing the Renardus data model demonstrates that the development of core metadata formats for cross-searching can be very

time-consuming. Nevertheless such effort is necessary for proper cross-searching in a distributed and heterogeneous service.

The Dublin Core metadata format is a good basis for the development of such a core set of metadata, although Dublin Core has no set of cataloging rules like AACR2 (often used for MARC 21). But the main goal of defining a common metadata set in Renardus is not resource description, but resource discovery. An even richer set of metadata for describing resources as developed for example by the subject gateways themselves is not necessary here. With eight "content" metadata elements the Renardus data model is very rich, compared with other information services, and users have a lot of possibilities for fielded searches as well as for sorting and filtering results.

The Renardus data model can also be seen as a model for other subject gateways which are under development. Some Renardus partners are now adapting this data model for their local services to enrich their metadata. The model may have to be updated if, in some future development, the Renardus project intended to include resources other than the mainly freely accessible online resources currently described by subject gateways. To include online journals, for example, some new metadata elements which provide information about rights (terms and condition, access/restriction conditions etc.), publishers and formats of the document will be important.

In this context the development of a library application profile (under development by the DCMI Libraries Working Group [26]) is very interesting. For future international co-operation, such an application profile may support interoperability between a range of metadata implementors.

Renardus also tries to follow the new developments in defining namespaces and application profiles. Both methods seem to be very useful for a standardized kind of communication and exchange of metadata between services. Also it is helpful as a more systematic way of defining and publishing a data model. Up to now, there are no official registries for such an application profile and for namespaces. RDF/XML encoding schemes for application profiles and namespaces are under development as well as for the declaration of collection level descriptions.

The project Renardus is one of the first attempts to use the full suite of Dublin Core based metadata features for the creation of a service which offers access to distributed, heterogeneous information services via one single interface. The concept of Dublin Core metadata and of collection level descriptions, the definition of specific namespaces as well as the development of an application profile can be explored in a working environment. The

experience of Renardus and some of its solutions may be a helpful basis for other interoperability efforts and services to build upon.

Acknowledgements:

The authors wish to thank Michael Day (UKOLN) for his invaluable enhancement suggestions.

References:

- [1] Renardus - <http://www.renardus.org/>
- [2] CORDIS: IST - <http://www.cordis.lu/ist/home.html>
- [3] CORDIS: 5th Framework Programme <http://www.cordis.lu/fp5/home.html>
- [4] Koch, Traugott. Quality-controlled subject gateways: definitions, typologies, empirical overview. Online Information Review - The International Journal of Digital Information Research and Use, Volume 24, Number 1, 2000: 26. <http://www.mcb.co.uk/oir.htm>
- [5] Dublin Core Metadata Element Set, Version 1.1: Reference Description <http://www.dublincore.org/documents/dces/>
- [6] Dublin Core Qualifiers <http://www.dublincore.org/documents/dcmes-qualifiers/>
- [7] DCMI Agents Working Group <http://www.dublincore.org/groups/agents/>
- [8] Renardus project at SUB Göttingen: Academic Subject Gateway Service Europe <http://renardus.sub.uni-goettingen.de/>
- [9] DCT2: Dublin Core Type Vocabulary: Subtypes Working Draft <http://epub.mimas.ac.uk/DC/subtypes.html>
- [10] Heery, Rachel, and Manjula Patel. Application profiles: mixing and matching metadata schemas. Ariadne, Number 25, 2000. <http://www.ariadne.ac.uk/issue25/app-profiles/>
- [11] DC-Education Proposal to the Advisory Committee <http://www.ischool.washington.edu/sasutton/dc-ed/Dc-ac/DC-Education.html>
- [12] IEEE Learning Technology Standards Committee's Learning Object Meta-data Working Group. Version 3.5 Learning Object Meta-data Scheme. <http://ltsc.ieee.org/wg12/>
- [13] DCMI Registry Working Group <http://www.dublincore.org/groups/registry/>
- [14] SCHEMAS Registry <http://www.schemas-forum.org/registry/>
- [15] Renardus Application Profile (SUB Göttingen, Germany) <http://renardus.sub.uni-goettingen.de/renap/>
- [16] DRAFT Namespace Policy for the Dublin Core Metadata Initiative (DCMI)

- <http://dublincore.org/documents/2001/03/09/dcmi-namespace/>
- [17] Kokkelink, Stefan, and Roland Schwänzl. Expressing Qualified Dublin Core in RDF/Draft/Version-2001-5-3
<http://www.mathematik.uni-osnabrueck.de/projects/dcqual/qual21.3.1/>
- [18] DCMI Architecture Working Group
<http://www.dublincore.org/groups/architecture/>
- [19] RDF Schema declaration for the Dublin Core Element Set 1.1 [2000/03/13]
<http://dublincore.org/2000/03/13/dces>
- [20] RDF Schema declaration for the Qualified Dublin Core Element Set [2000/03/13]
<http://dublincore.org/2000/03/13/dcq>
- [21] SCHEMAS: Forum for Metadata Schema Designers and Implementers
<http://www.schemas-forum.org/>
- [22] RDF schema for the RSLP-CLD Application Profile
<http://www.schemas-forum.org/registry/schemas/RSLP-CLD/index.html>
- [23] RSLP Collection Description
<http://www.ukoln.ac.uk/metadata/rslp/>
- [24] Collection Level Description
<http://www.ukoln.ac.uk/metadata/cld/>
- [25] RSLP Collection Description Tool
<http://www.ukoln.ac.uk/metadata/rslp/tool/>
- [26] DC- Library Application Profile
<http://dublincore.org/documents/2001/08/08/library-application-profile/>