Mapping scheme from Greenstone to CERIF format

Valentin Penca*, Siniša Nikolić*, Dragan Ivanović*

* University of Novi Sad/Faculty of Technical Sciences/Department of Computing and Electronics, Novi Sad, Serbia
{valentin_penca,sinis_nikolic, chenejac} @uns.ac.rs

Abstract— This paper describes the basics of the Greenstone institutional repository and CRIS systems and their data models. The result of this research is mapping scheme of the data from Greenstone to the CERIF standard.

I. INTRODUCTION

Quick development of science and technologies resulted with vast amount of various data. One of the most important tasks is how to preserve and make that data accessible. Institutional Repository (IR) can solve the mentioned issue. In [1], an IR is addressed as an electronic system that captures, preserves and provides access to the digital work products of a community. The three main objectives for having an institutional repository are:

1. creating global visibility and open access for an institution's research output and scholarly materials.
2. collecting and archiving content in a "logically centralized" manner even for physically distributed repositories.
3. storing and preserving other institutional digital assets, including unpublished or otherwise easily lost ("grey") literature.

The availability of open-source technologies affect on the rapid development of IRs worldwide, particularly among academic and research institutions. Thus, it is not surprising the existence of several open-source software platforms available for developing IRs like Greenstone (GS) [2], EPrints [3], DSpace [4], Fedora [5] and Invenio [6]. GS large popularity in large number of countries is addressed in [7]. Although IRs had been used for a long time, they still don’t have a mutually agreed and standardized representation of their data. This can cause difficulties in data exchange between diverse IR systems.

So, to overcome those difficulties in the data exchange, one of the possible solutions is to rely on some predefined standard outside IR domain. Common European Research Information Format (CERIF) standard [8], which is the basis of Current Research Information Systems (CRISs), is used for data exchange from scientific-research domain and can be utilised in IR domain.

In this paper the scheme for mapping data from Greenstone IR to CERIF format is proposed. That scheme can be used as a guideline, supporting the exchange between Greenstone repositories and CRIS systems.

Motivation for this work was also to extend and improve research from [9] [10].

II. GREENSTONE IR

The Greenstone digital library software is a system for construction and presentation of information for digital resources. Digital resources can range from newspaper articles to technical documents, from educational journals to oral history, from visual art to folksongs, etc. So, Greenstone is an IR that can contain data which can vary in types and formats. In GS, each single digital resource is described with appropriate metadata and has a link to a stored physical document of that digital resource.

The software provides a way of organizing information and publishing it on the Web in the form of a fully-searchable, metadata-driven digital resource. Greenstone was one the first IR software packages to appear and has been available for more than 15 years. First widespread version was Greenstone 2.0 that was released under the GNU General Public License in September 1999. Greenstone has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium. It is an open-source multilingual Institutional repository software. One of Greenstone's unique strengths is its multilingual nature where their interface is available in over 60 languages [11].

The Greenstone software operates under most variants of Unix (including Linux, FreeBSD and MacOS X) and all versions of Microsoft Windows. Current Greenstone version is "Greenstone3" that is a complete redesign and reimplementation of the original digital library software developed (Greenstone2) back in 2000 and the last stable release came in 9th September, 2015. So, it can be concluded that Greenstone is continually updated and reliable software that has a long successful history of development.

Popularity of software is confirmed thought its usage on worldwide scale. Greenstone software has spread in over 90 counties (e.g. Canada, Germany, New Zealand, Romania, Russia, United Kingdom, United States, etc.). According to the official data presented in [12] there are roughly 3800 active software instances. Popularization of Greenstone has been been increased by realizing workshops all over the World [13]. Comparison and advantages of Greenstone to other IR repositories is described in [14].

In Greenstone all digital resources (documents) are organized into collections [Figure 1]. Those collections are additionally organized into a library (called a site), and
GS can have multiple sites in a single Greenstone3 installation. A site can be seen as a collection of collections. Site actually provides one upper level of data organization. For e.g. one site can house all of image collections, another site all of pdf collections, and yet another can be used for all of multimedia collections. Or, perhaps sites can be used to separate collections based on topics. Technically GS could have a separate site for each collection (this may not be particularly useful, but is possible).

The collection database dictates how data about a collection is stored. Greenstone comes with two database options: GBDM and JDBM. They are essentially the same. GBDM (or GNU Database Manager) is a simple flat-file database engine which Greenstone uses as the default database for new collections. JDBM is simply a Java implementation of GBDM. It is also possible to use a relational database system, like Microsoft SQL (MS-SQL) or MySQL to create a collection's database. The collection database dictates how data about collection is stored.

The physical documents for digital resources are stored as such (PDF, DOC, HTMIL, XML...). In GS there are different ‘Plug-ins’ [15] that are used for different types of physical documents in order to extract textual content from those documents. Extraction of words from physical document is essential in digital resource creation. Extracted words represent the basis for constructing metadata (e.g. title, author, creation date, etc.) which is used for describing and indexing of those digital resources. In Greenstone supplying metadata to digital resource is an extremely important part of building digital collections. It provides important contextual and provenance information about that recourse. Essentially, it helps users to navigate collections and find the information/documents they need.

In Greenstone, all metadata fields belong to a metadata set, which is simply a pre-defined collection of metadata fields. Because sets often have metadata fields with the same name (for instance, most sets will have a Title field), namespaces are used to distinguish between metadata from different sets. For instance, all metadata fields in Dublin Core are preceded by word dc followed by dot symbol (e.g. dc.Title, dc.Creator, etc.). In the process of creating a new collection in Greenstone, the default metadata sets that are available are Dublin Core (dc), the Greenstone Metadata Set (gs), and the Extracted Greenstone Metadata Set (ex). The Extracted set is specific because it contains metadata automatically generated during the collection building process and cannot be edited by user. List of all available sets that come with Greenstone installation is found on [16].

For every stored document within Greenstone a XML-format file called metadata.xml is used. That file contains metadata of the electronic resource and applies to all files that are in the same directory, or in one of its child directories.

![Greenstone data model diagram]

### III. CERIF MODEL

CERIF is a standard that describes data model which can be used as a basis for an exchange of data from scientific-research domain. CERIF Standard describes the physical data model [17] and the exchange of XML messages between the CRIS systems [18]. The best feature of CERIF is that it can be expanded and adapted to different needs. In practice, CERIF is often mapped to other standards that also represent the data of scientific-research domain, for example CERIF/MARC21 mapping described in [19]. Authors of [20] recommend an extension of CERIF that incorporates a set of metadata required for storing theses and dissertations. Another example is [21] where authors argue how CERIF can be used as a basis for storage of bibliometric indicators.

Hereinafter we will present main entities of the CERIF data model version 1.5

- **Base Entities** - represent the core (basic) model entities. There are only three basic entities: cfPerson, cfOrganizationUnit and cfProject.
- **Result entities** - A group of entities which includes results from scientific research like publications, products and patents. Representatives of this group are: cfResultPublication, cfResultProduct and cfResultPatent.
- **Infrastructure Entities** - represent a set of infrastructure entities that are relevant for scientific research. The entities which belong in this group are: cfFacility, cfEquipment and cfService.
- **2nd Level Entities** - Entities that further describe the Base Entities and Result Entities. E.g. cfMedium can be physical representation of some Result Entity.

Link Entities - are used to link entities from different groups. Typical entities of this group are: cfOrganizationUnit_OrganizationUnit, cfOrganizationUnit_ResultPublication and cfResultPublication_DublinCore. Link Entities allow for a generic classification mechanism to define their meaning, indicating the role for each entity instance in a relationship. Every Link
entity is described with a role (cfClass, cfClassScheme), timeframe of relation (cfStartDate, cfEndDate), value (cfFraction) and identifiers of elements creating relation (e.g. cfOrgUnit, cfResPublId). The 'role' in link entities is not stored directly as attribute value, but as reference to Semantic layer.

- Multiple Language Entities - These entities provide multilingualism in CERIF for some entities.
- Semantic Layer Entities - Provide different kinds of semantics in CERIF model. The entities in this group are cfClassificationScheme and cfClassification. Those entities are used to describe classes and classification schemes for link and other entities. CERIF prescribes a controlled vocabulary to describe some of the classifications.
- Additional Entities - Currently in this group are classified entities that represent DC record.

**IV. MAPPING SCHEME FOR GREENSTONE TO CERIF**

The motivation of authors for mapping Greenstone data to CERIF data model is found in the fact that they are part of the development team of CRIS UNS system [22], which currently does not have ability to obtain data from the Greenstone system. In the paper [23], a CERIF compatible research management system CRIS UNS is presented, which can be accessed at [22]. Currently, the system stores over 14,500 records of scientific publications (papers published in journals, conference proceedings, monographs, technical solutions and patents etc.). CRIS UNS system is under development since 2008 at the University of Novi Sad in the Republic of Serbia. Former development of that system covered implementation of the system for entering metadata about scientific research results [24]. Later phases in the development of CRIS UNS system included integration of various extensions that rely on CERIF model.

Proposed mapping scheme did not include any customization of Greenstone distribution since the most users use just the default installation.

Greenstone can store various physical documents that came from scientific research domain, such as journal articles, conference papers, whitepapers, monographs, different forms of reports, etc. Unlike CERIF that has different entities (e.g. cfResPubl, cfResProd, cfPers, etc.) for keeping information about different record types (e.g. publications, products, persons, etc.), GS only has digital recourses in which the record type is not explicitly defined. Therefore, it is essential to derive the data mapping in a manner to comply with the structure and semantics of both models. Mapping of the data from GS model [Figure 1] will begin with key GS entity (Document) which carries the relevant data about the document.

Table I shows part of mapping of the GS metadata for digital recourses to the appropriate entities in CERIF. In proposed mapping each digital resource from GS is represented as an instance of CERIF cfMedium entity, in accordance to EuroCRIS suggestion [25]. “GS Metadata Sets” column indicates which metadata set form GS is used for a particular metadata that is displayed in the column “GS Metadata Element”. Columns “multiple” give information whether metadata may appear more than once in the GS and/or model CERIF.

One of the simplest examples of metadata mapping can be shown on document title which is represented in CERIF with the concrete value of title that is stored in cfTitle attribute of entity cfMediumTitle which is certainly linked with cfMedium. Representation of the organization that is resource owner requires creation of CERIF entity cfOrgUnit and cfOrgUnitName which will store a name.
for organization/institution. A connection between the resource (cfMedium) and organization/institution (cfOrgUnit) in CERIF is implemented with link entity cfOrgUnit_Medium (column CERIF Link Entities). CERIF model relies on Link and Semantic Layer Entities to provide additional semantic between entities and for some particular entities. So, it is to be assumed that a large portion of metadata fields from GS documents will be stored as instances of those CERIF entities. In our case it is necessary to classify the link entity cfOrgUnit_Medium with appropriate classification scheme and class. (column Used CERIF Classification). Some of CERIF link entities and classification can be connected with the identifier (e.g. cfMediumId, cfOrgUnitId).

A certain metadata (e.g. Language) from GS cannot be directly mapped to cfMedium and/or its link entities from CERIF model. However, in CERIF there is a group of "Additional Entities" that is used for the purpose of mapping metadata that is coming from other external systems. These CERIF entities are built on model of a widely known DC metadata set [26]. So, language value will be stored in attribute cfDCValue of cfDC_Language entity. In order to adequately map mentioned metadata (e.g. Language) it was needed to add only one linking entity element (cfMedium_DC) between the cfDC and a cfMedium. Also, it was essential to use the appropriate classification (column used CERIF Classification) which provides the semantic for that relation.

A specific case is the mapping metadata that indicates a logical connection between resources. So, metadata Relation_HasPart from the extended DC metadata set [27] is actually defining the link between two physical resources (cfMedium), where the relationship is represented with entity cfMedium_Medium that is classified with the appropriate classification scheme (Inter-Medium Realations) and class (has part).

As previously mentioned in the description of the GS model, several electronic resources can form a single collection (Figure 1). Metadata that describes the collection is stored in file "collectionConfig.xml". Table 2 shows a part of metadata mapping for GS collection. Keeping in mind that collection in GS represent only a hierarchical level of organization for electronic resources, collections in CERIF are represented with CERIF semantic layer. For that purpose a new classification scheme (cfClassScheme) GS_Collection is defined, and each collection is stored as a new classification (cfClass). Each GS collection has its own name that can be multilingual. The collection name is preserved in cfTerm attribute of cfClassTerm entity Language in which the name is stored and preserved via attribute cfLangCode. Metadata field "description" is mapped to an entity cfClassDescription, in similar manner as metadata element "name" to cfClassTerm. Metadata creator of the collection can be equally mapped to either cfPers or cfOrg_Unit, depending on the contents of the field. Collection is assigned to person or organisation with entities cfPers_Class or cfOrg_Unit_Class. The public status of GS collection is defined in metadata field

<table>
<thead>
<tr>
<th>GS Metadata Sets</th>
<th>GS Metadata Element</th>
<th>multiple</th>
<th>CERIF Core result 2nd level entities</th>
<th>CERIF Link Entities</th>
<th>Used CERIF Classification</th>
<th>multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLS</td>
<td>Title</td>
<td>X</td>
<td>cfMedium (cfMediumId) cfMediumTitle(cfMediumId, cfLink, cfTrans=&quot;T&quot;)</td>
<td>cfOrgUnit_Medium (cfOrgUnitId, cfMediumId)</td>
<td>Scheme:CERIF Entities Class: Organization</td>
<td>X</td>
</tr>
<tr>
<td>DLS</td>
<td>Organization</td>
<td>X</td>
<td>cfMedium (cfMediumId) cfOrgUnit (cfOrgUnitId) cfOrgUnitName (cfOrgUnitId, cfName)</td>
<td>cfMedium_Medium (cfMediumId, cfMediumId)</td>
<td>Scheme:CERIF Metadata Class: metadata</td>
<td>X</td>
</tr>
<tr>
<td>DLS</td>
<td>Language</td>
<td>X</td>
<td>cfDC(cfDCId, cfDCSchema=&quot;DLS&quot;, cfDCLanguage(cfDCId, cfDCSchema=&quot;DLS&quot;, cfDCValue))</td>
<td>cfMedium_DC (cfMediumId, cfDCId)</td>
<td>Scheme:Medium Metadata Class:has metadata</td>
<td>X</td>
</tr>
<tr>
<td>DCE</td>
<td>Relation_HasPart</td>
<td>X</td>
<td>cfMedium_Medium (cfMediumId, cfMediumId2)</td>
<td>Scheme:Inter-Medium Realations Class:has part</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GS Collection XML Element</th>
<th>multiple</th>
<th>CERIF</th>
<th>Used CERIF Classification</th>
<th>multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>collection</td>
<td>X</td>
<td>cfClassScheme, cfClass</td>
<td>Scheme:GS_Collection Class: NEW CLASS INSTANCE</td>
<td>X</td>
</tr>
<tr>
<td>&quot;name&quot; &quot;lang&quot;</td>
<td>X</td>
<td>cfPers, cfPersName (cfFirstName, cfLastName, cfOtherName) cfPers_Class cfOrgUnit cfOrgUnitClass</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;description&quot; &quot;lang&quot;</td>
<td>X</td>
<td>cfPers, cfPersName (cfFirstName, cfLastName, cfOtherName) cfPers_Class cfOrgUnit cfOrgUnitClass</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>&quot;creator&quot; &quot;lang&quot;</td>
<td>X</td>
<td>cfPers, cfPersName (cfFirstName, cfLastName, cfOtherName) cfPers_Class cfOrgUnit cfOrgUnitClass</td>
<td>Scheme:Inter-Medium Realations Class:has part</td>
<td></td>
</tr>
<tr>
<td>public</td>
<td>X</td>
<td>cfClassScheme cfClass cfClass_Class (cfClassId, cfClassShemeid, cfClassShemeid1, cfClassShemeid2, cfClassId2)</td>
<td>Scheme:GS_Collection_TYPE Class: Public Access or Private Access Scheme: GS_Collection_Type Class: Public Access or Private Access e Scheme: Inter_Collection Realations Class:has part</td>
<td></td>
</tr>
</tbody>
</table>
“public” in GS. In CERIF model the mentioned GS metadata information requires a creation of a new classification scheme (GS_Collection_TYPE) and class (Access Public or Private Access). Also, it was necessary to link that new classification to one that represent the collection. The interconnection between two classifications in CERIF is achieved with link entity cfClass_Class. Assigning a digital recourse to collection is achieved by classifying cfMedium instances with a newly created CERIF classes that are GS collections.

As mentioned before multiple collections may be additionally organized within entities called “sites”. (Figure 1). Each site is described with particular metadata that is stored in file siteConfig.xml. Similar to collections, the mapping of site (Table 3) entity can be done by relying on CERIF semantic layer. The main role of the site entity is to describe a group of collection. Like a collection site has the multilingual metadata name, description, and creator that can be mapped on the model CERIF in a same manner as GS collection. All Greenstone mapping are presented on [28].

The main purpose of the presented scheme is mapping data from GS to a concrete CRIS UNS. The mere fact is that GS can export/import data to/in certain format. According to that, this can open up a couple of potential opportunities for transferring data from one system to another. The first potential opportunity is the fact that it has the ability to export data by MARCXMLPlugin. CRIS UNS has MARC21 compatible data model [19] that could enable mapping the exported data from the GS. Exported data to DSpace model is also acceptable for CRIS UNS considering that the authors have already suggested mapping of DSpace data to CERIF [9].

There is another potential direction of interoperability between these two systems, from CRIS UNS to GS, where Greenstone is able to obtain data from CRIS UNS system. As a matter of fact, the Greenstone has SRU/W client so it can possibly connect to CRIS UNS, considering that the system provides the data in SRW XML format in accordance to the specific profile [29]. Import of such data can be easily accomplished with GS SRU client and adequate plug-in in GS [30]. Some type of data (doctoral dissertation) from CRIS UNS that is available in OAIPMH format [31], can be useful for GS that have OAIPMH import plug-ins. Last but not the least, Greenstone can import a data (records) directly from the database (MySQL) using DatabasePlugin (DBPlug). The exported data from the CRIS UNS MySQL database could serve as a starting point for DBPlug. Greenstone has a large number of plugins [15] to import different kinds of data, so it has the potential to take over data from other systems that are not only CERIF like.

V. CONCLUSION

The importance of institutional repositories and CRIS systems for scientific research data is enormous. Making data accessible between these systems is unavoidable. Therefore, this paper presents a mapping scheme for Greenstone data to CERIF model where all GS data is available to CRIF like systems (CRISs, IRs that support CERIF, etc.).

The main contributions of this research are:
• Proposal for mapping data from Greenstone repository to the current 1.5 CERIF model
• Potential possibility for creation Import/Export plug-in for making full interoperability between this systems.

Future work will be directed towards mapping the data from other IRs like Fedora, and Invenio to CERIF format.

ACKNOWLEDGMENT

Results presented in this paper are part of the research conducted within the Grant No. III-47003, Ministry of Science and Technological Development of the Republic of Serbia.

REFERENCES

[5] “Fedora Repository | Fedora is a general-purpose, open-source digital object repository system.” [Online],


