TCDMeta: a metadata model used to automatically create collections and meta-collections of the academic content in the institutional repository of the Technological Educational Institute of Crete

Nikolaos Tsatsakis, Alexandros Gougousis, Stavroula Varvantaki, Georgia Tzedaki, Mihail Panagiotakis

Technological Educational Institute of Crete, Estavromenos Iraklio Crete Greece GR 71004

Abstract: An academic institutional repository has to provide the ability of structuring its content throughout its existence, preferably according to a dynamically evolving as well as easily adaptive mechanism which categorizes all items to a multilevel structure.

The description and use of TCDMeta as a metadata model to achieve content categorization into collections and meta-collections for the institutional repository of the TEI of Crete is described in this paper.

The model was built during the debugging process of the descriptive metadata for the collection of the students' diploma theses and will give us the ability to create collections and meta-collections other than those related to faculties, academic departments and years of creation which were used up to now.

The systematic consideration of all items provided us with a distilled knowledge of the content in our institutional repository and designated the main elements of the model: subjects' collections related to carefully chosen keywords from a controlled vocabulary, popular topics' collections that can help in a macroscopic study of the methods, tools, and outcomes that are related to those topics, geographic collections that contain studies and experiments held in certain geographic areas, distinguished theses collections and others.

Keywords: Automatic collection creation, Metadata model, Institutional repository, XML

1. Introduction

Institutional repositories are databases of digital archives with academic and research content accessed through Internet. They are implemented and supported by academic/research organizations which collect, organize and

Received: 21.10.2013 / Accepted: 1.3.2014 ISSN 2241-1925

preserve in them the digital material that they produce, for open or controlled access according to related intellectual property rights[1].

According to a study conducted by Spanos, Arkoulis, Stavros, Giannopoulou and Mitrou (2013) the benefits brought in an academic institution through the development and use of an institutional repository are numerous and quite important, namely:

a) increased promotion and diffusion of its research outcomes
b) presentation of a complete view of the institution to future students and staff members, as well as collaborators and researchers
c) integrated gathering, long term preservation, and updated - according to format - delivery (i.e. postscript files may become available in pdf) of the digital production of the institution
d) major contribution in the quantification of the research and academic activities and the overall performance of the institution, which is quite important for the evaluation and excellence in each level of the academic structure (institution, branches, faculties, departments, laboratories, groups)
e) centralized collection of the current institutional research
f) contribution in the processes of claiming research subsidies and participating to consortia for the implementation of major research programs
g) empowering of the cooperation between different institutions and of the interdisciplinary research
h) concentration and delivery of learning material to students
i) support for students during their studies and research by providing access to previous diploma/master/PhD theses.

In the following, the terms *item* and *record* are used to refer to the content of the repository. Although their exact meaning is context sensitive, we use *item* to describe a work in its entirety (i.e. article, diploma thesis) or the expression of a work (i.e. image file representing a painting) and may consist of more than one files[2], while *record* is used to describe the metadata that are related to a work. Also the terms *group of items* or *collection* are used to refer to groups of items formed according to a common characteristic – attribute value (i.e. the group of diploma theses of a department), and aim to a better organization of the repository's content.

In order for the above mentioned benefits to be achieved, an academic institutional repository must have the ability of structuring and presenting its content throughout its existence preferably according to a dynamically evolving, as well as easily adaptive, grouping mechanism. Grouping mechanisms in such repositories are generally hierarchical and categorize all items to a multilevel structure, that is to more than one different collections according to one or more

[1] National Documentation Center, *«Institutional Repositories: About»,* http://www.openaccess.gr/repositories/what.dot/

[2] The terms work, expression and file are defined in the FRBR model

qualitative criteria that are chosen by the institution and applied by the system administrator of the repository. For instance a grouping of items may reflect some organizational structure (i.e. faculties / departments / sectors etc).

A benefits' augmentation from the grouping of items to collections is possible by offering different management and deposit functionality for each collection, properly adapted to the demands of certain user groups. In this way, the same repository system may fulfill the different needs of the members of big and scientifically heterogeneous institutions.

However, for the collection creation in a repository to be dynamic, accordant to the changing demands of its users and the conceptual content of its items (semantics), which can only be analyzed and described a posteriori, we consider that the grouping of items to collections must be based on content related metadata which apart from content aggregation will lead to the achievement of semantic interoperability.

Metadata are very important for the organization, description and discovery, correlation and grouping, and overall utilization of the items in a repository and have different use according to their type: *descriptive metadata* refer to properties such as type, author, or title, and facilitate items' searching and discovery, *structural metadata* describe the structure of complex items (i.e. a book with chapters saved in different data files) as well as the structure of the repository itself by defining its collections' hierarchy, and *administrative metadata*, which are additionally separated in *technical, access, and preservation metadata* and are necessary for the administration of the items.

Metadata schemes used in academic institutional repositories are quite a lot, with Dublin Core (DC) / ETD-MS / SWAP / XMetaDiss being the most widespread for diploma theses, IEEE LOM for educational material, PRISM for journal publications etc. Recently HEAL – Link [3] recommended the use of healMeta (Spanos, Arkoulis, Stavros, Giannopoulou and Mitrou 2013) by all Greek academic institutions, as the metadata scheme for the description of academic and research production in its entirety. This scheme does not preclude coexistence with DC, empowers semantic interoperability among repositories that use different deposit workflows, and facilitates the collection of metadata through OAI-PMH by content aggregators.

In this paper we present an extension of healMeta with appropriate elements, in order to be used for the collection creation in the institutional repository of the TEI of Crete, which contains items related to diploma theses. The name of the extended scheme, TCDMeta, corresponds to the items' type to whose metadata it refers to: TEI of Crete Diploma theses, while the organization of the paper is as follows: in section 2 we describe the methodology that we followed in order to determine the collections in our repository by means of conceptual analysis of the items, in section 3 we present the metadata that are related to the new collections , and in section 4 we summarize the final conclusions and extensions of the suggested schema.

[3] HEllenic Academic Libraries Link

2. Methodology for the determination of new collections in the institutional repository of the TEI of Crete

Since 2001 an effort aiming to the gathering, preservation, and promotion in a unique way of the intellectual production conducted by students, has been established in the TEI of Crete. The related material that has not been published elsewhere comprises the content of the institutional repository, e-Thesis (http://nefeli.lib.teicrete.gr), where all diploma theses are deposited, described by use of a simple "bibliographic scheme" with elements from DC, and can be delivered to users (full text).

In the currently used institutional repository the grouping of items is simple and based on the faculty, department and year of creation of each item. Additionally users may perform free text searches to the whole bibliographic description of items, or certain fields of it (title, author, advisor, abstract, creation year).

Despite of its simplicity, items' grouping is important and provides different views of the collection to users, but is not considered adequate for research and the substantial use of the repository's content. Thus we studied thoroughly the possibility of grouping items furtherly, by use of other hierarchical structures, so that their searching and presentation is promoted. This enhanced grouping of items is a functionality included in the challenges of the upgrade process of the TEI's institutional repository, a service implemented in the framework of the action "Organization, highlighting and promotion of the academic content of the TEI of Crete" under the operational program "Digital convolution".

The way of creating new items' collections and meta-collections was determined during the debugging process for the descriptive metadata (bibliographic descriptions) of items, in order to be as correct and complete as possible before the migration to the new repository system. Thus we had to study a collection of 5157 theses and their related items and descriptions, and perform corrections in its entirety.

During the process we ascertained the need to constitute a list of suggested terms that will be used as keywords in place of those miss chosen so far, and built a global lexicographic catalogue of keywords to be used during the deposit process of items. Furthermore, we noticed that, certain keywords are repeated, diploma theses with similar or same titles exist, interdisciplinary research with strong interaction has been done etc.

In the following we refer to new items' collections that may result from the current content of the TEI of Crete institutional repository:

- *Popular theses' subjects or subject collections* that are related to the repetition/reuse of keywords and can be derived according to the keywords assigned to theses' descriptions. It is noteworthy that, in the current repository, keywords were not derived from a controlled vocabulary of terms or thesaurus. They have been freely assigned by the author of the work or the staff member that performed the deposit to the repository, thus resulting in often use of the same term in mismatch ways (i.e. plural - singular, spelling errors, abbreviations etc). After the use of the lexicographic catalogue of keywords, terms mismatch eliminated. So, the creation of subject collections from keywords has become straightforward and correct. In the future repository we will additionally use a standard based subject heading for each item (i.e. LCSH or GNLSH[4]) thus enhancing even more the creation of subject collections.

- *Collections of theses with a repeated subject* that are derived from the theses' titles and related items that have same or similar titles. Despite of the fact that their existence was at first considered as mishandling of the advisors, we eventually decided that such collections will contribute to the macroscopic study of the outcomes they deal with. For instance, the potential researcher will be able to examine the productivity related to the application of particular methods or experiments to certain scientific fields, be aware of the evolution of the systems used in various application domains, and study the impacts of climate change while practicing different scientific solutions. Such collections have the essence of a meta-research and aware the role of "observatories for the science and technology".
- *Collections regarding geographical areas,* which may include works related to experiments, studies and methods, or the implementation of systems or services that took place to a certain geographical area. By use of the toponyms that appear in titles, abstracts or keywords collections related to them may be created. Further grouping inside them according to the department in which the work took place or the scientific field it is related to, may occur. For example a researcher may be able to study theses regarding water pollution, or the cultivation of vegetable species to certain geographical areas.
- *Collections of works conducted by certain scientific sectors* of the departments of the institution. These may contain diploma or master theses carried out to each sector, or even laboratory, of each department, and can be used for its promotion and the specification of evaluation factors. Their creation can be based on a matching process of advisors of theses to the sectors / laboratories they belong to.
- *Collections of distinguished theses* which may be derived from the value in a certain metadata element that will be computed by the application of criteria defined by advisors to the values of other descriptive metadata elements. Such collections are very important for the recognition of the authors of the items they contain (i.e. can help students in applications to postgraduate programs or jobs).
- *Collections of published works* that will be derived from the value in a certain metadata element which will be filled appropriately during the deposit process, or a related post processing of the records in the repository.
- *Collections of surveys* for different scientific areas, in which are grouped items with the terms "study", "comparative study", or "survey" inserted as values to the title, abstract or keyword elements of the metadata schema.
- *Collection of items with similar format in accompanying material* where for

[4] Library of Congress Subject Headings, and Greek National Library Subject Headings

example items that are accompanied by autocad designs or mp4 videos will be grouped. These collections will be useful for potential users that search for certain material types, as well as for the system administrator of the repository during the process of format update of its content (for preservation reasons).

One may easily deduce that the automated creation of the above mentioned collections can essentially be based on the consistent annotation of items with suitable metadata, thus, on the adoption and systematic use of a scheme that contains them. In the section that follows we present the elements of healMeta along with others contained in TCDMeta that must be used for the collection creation in the repository of the TEI of Crete.

3. 3.HealMeta and TCDMeta elements that must be used for the automated creation of new collections

A detailed presentation of the necessity as well as the elements healMeta contains can be found in a study conducted by Spanos, Arkoulis, Stavros, Giannopoulou and Mitrou (2013). From all its elements we have to use at least the following for the creation of the collections mentioned in the previous section:

Work Type
Element: heal:type
Possible values: (bachelorThesis, masterThesis, doctoralThesis, conferenceItem, journalArticle, bookChapter, book, report, learningMaterial, dataset, other)
Mandatory: Yes
Repeatable: No
Content description: The type of each work which is expressed as a value from a predefined vocabulary.
XML syntax: <heal:type>Work Type</heal:type>
Example: <heal:type>bachelorThesis</heal:type>

Title
Element: heal:title
Attribute: xml:lang= " RFC 5646 code" (mandatory)
Mandatory: Yes
Repeatable: Yes
Content description: The main title of each work as provided by its creator. The xml:lang attribute denotes the language in which the title is expresses and is declared by a code related to RFC 5646[5] standard which may be searched to the IANA[6] registry (i.e. “en” for English and “el” for Greek).

[5] http://tools.ietf.org/html/rfc5646

[6] IANA Language Subtag Registry: http://www.iana.org/assignments/language-subtag-registry

XML syntax: <heal:title xml:lang=*"*code from RFC 5646*"*>Work Title</heal:title>
Example: <heal:title xml:lang=*"en"*> Corporate crisis management </heal:title>
Subject Classification
Στοιχείο: heal:classification
Attribute: scheme= (LCC, DDC, UDC, NLM, ACMCCS, MSC, PACS, other) (optional)
Attribute: xml:lang= " RFC 5646 code" (mandatory)
Mandatory: No
Repeatable: Yes
Content description: The broader subject classification in which the work belongs to, which may consist of an authority from an encoding scheme or of free text. In case an authority is used the attribute scheme must have an appropriate value denoting the encoding scheme of the subject classification. For the sake of semantic interoperability among Greek institutional repositories the use of LCC and DDC is suggested, for which matching tables exist. The element’s value is comprised by the subject classification title followed by the corresponding code of the encoding scheme used in brackets. In case free text is used for the subject classification, the attribute scheme is omitted. In both cases the attribute xml:lang denotes the language of the subject classification. It is notable that the element subject classification refers only to the broader subject to which the work belongs to, while the element keyword refers to more specialized authorities.
XML syntax: <heal:classification xml:lang=*"* code from RFC 5646 *"* scheme=*"*encoding scheme id*"*> Subject classification [Subject classification code]</heal:classification>
Example: <heal:classification xml:lang=*"en"* scheme=*"LCC"*> Surveying engineering science [Q161]</heal:classification>

Keyword (Subject)
Element: heal:keyword
Attribute: scheme= (LCSH, MeSH, STW, AAT, other) (optional)
Attribute: xml:lang= " RFC 5646 code" (mandatory)
Mandatory: No
Repeatable: Yes
Content description: A keyword that describes a work, which may either be an authority from a thesaurus or vocabulary, or free text. In the first case the element must contain the attribute scheme denoting the encoding scheme of the authority. The element’s value is the keyword followed by the code of the encoding scheme used in brackets. In case free text is used as a keyword, the attribute scheme is omitted. In both cases the attribute xml:lang denotes the language of the keyword.
XML syntax: <heal:keyword xml:lang=*"*code from RFC 5646*"* scheme=*"*encoding scheme id*"*> keyword [Keyword code]</heal:keyword>
Example: <heal:keyword xml:lang=*"en"* scheme=*"LCSH"*>Linear models (Statistics) [sh85077177]</heal:keyword>

Included File Format
Element: heal:fileFormat
Mandatory: No
Repeatable: Yes
Content description: The format of the files that represent a work regarding the format vocabulary of IANA[7]. It is mandatory in order to cover the existence of items without accompanying material (files).
XML syntax: <heal:fileFormat>File Format</heal:fileFormat>
Example: <heal:fileFormat>application/pdf</heal:fileFormat>

Advisor's Name
Element: heal:advisorName
Attribute: xml:lang= "code from RFC 5646" (mandatory)
Mandatory: Yes
Repeatable: No
Content description: The name of the advisor of a thesis.
XML syntax: <heal:advisorName xml:lang=*"*code from RFC 5646*"*>Advisor name</heal:advisorName>
Example: <heal:advisorName xml:lang=*"en"*> Theodoros Antoniou</heal:advisorName>
In addition to the above there is a need to define elements for the description of theses as distinguished or surveys. These elements that are included in TCDMeta metadata scheme are defined as follows:

Characterized of a Work as Distinguished
Element: TCD:Distinguished
Mandatory: Yes
Repeatable: No
Content description: Indication whether the work is distinguished or not. The characterization is denoted by a true value following the specification of data type boolean in XML Schema.
XML Syntax: <TCD:Distinguished> {true, false} </ TCD:Distinguished >
Example: < TCD:Distinguished > true </ TCD:Distinguished >

Characterized of a Work as Survey
Element: TCD:Survey
Mandatory: Yes
Repeatable: No
Content description: Indication whether the work is a survey or not. The characterization is denoted by a true value following the specification of data type boolean in XML Schema.

[7] IANA list of MIME Media Types: http://www.iana.org/assignments/media-types/index.html

XML Syntax: <TCD:Survey> {true, false} </ TCD:Survey>
Example: < TCD:Survey > false </ TCD:Survey >
It is obvious that the two metadata schemes differ only in those two last elements. However in case a need of some new collection arises (either by furtherly studying the repository content or by conceptual characteristics of new content) we may easily cover this need by adding new elements to healMeta scheme.

4. Conclusions – future extensions

The basic conclusion of this work is the fact that in order to profit the most form the content of an academic institutional repository for the sake of its creators, the institution itself, and the final users – researchers, it is of great importance to create collections of items by use of metadata schemes as:

- these schemes are extended easily and in a controlled manner and can be used for the semantic annotation of items, and
- their use facilitates the automated creation of the demanded collections.Additionally, we ascertained the utility of the study of a collection of items a posteriori in order to make decisions regarding their optimal grouping to subcollections, as despite of the value of an a priori study of items, it is always an estimate as far as data (the items) have (and in our belief they will always have) their inherent dynamics and meaning.

We are now in a process to design and implement the use of healMeta and TCDMeta to the new institutional repository for the TEI of Crete, and thus automate as possible the creation of the previously mentioned collections of items in the repository.

References

Allard, S., Mack, T. R., Felther-Reichert, M. (2005), "The librarian's role in institutional repositories: A content analysis of the literature", *Reference Services Review*, Vol.33 Iss: 3 pp.325-336.

Burk, A., Al-Digeil, M., Forest, D., Whitney, J. (2007) "New possibilities for metadata creation in an institutional repository context", OCLC Systems & Services, Vol. 23 Iss: 4, pp.403 – 410

Jain, P., (2011), "New trends and future applications/directions of institutional repositories in academic institutions", *Library Review*, Vol.60 Iss: 2 pp.125-141.

International Federation of Library Associations and Institutions (IFLA), Functional Requirements for Bibliographic Records, 1997.

Lagoze C., van de Sompel H., Nelson M., Warner S ., "The Open Archives Initiative Protocol for Metadata Harvesting", 2008, available online at: http://www.openarchives.org/OAI/openarchivesprotocol.html.

M. Nilsson, A. Powell, P. Johnston, and A. Naeve, "Expressing Dublin Core metadata using the Resource Description Framework (RDF)", 2008, available online at: http://www.dublincore.org/documents/dc-rdf.

Σφακιανάκης Μ., Καπιδάκης Σ. (2007): Ενισχύοντας Σημασιολογικά τις Διαδικασίες Αναζήτησης σε Ένα Περιβάλλον Μετα-Αναζήτησης. *Πρακτικά του 16ου Πανελληνίου*

Συνεδρίου Ακαδημαϊκών Βιβλιοθηκών: Ο Ανθρώπινος Παράγοντας στη Διαμόρφωση της Σημερινής και της Μελλοντικής Βιβλιοθήκης, 640 – 651.

Δ. Σπανός, Σ. Αρκουλής, Π.Σταύρου, Ε. Γιαννοπούλου και Ν. Μήτρου, «ΛΕΙΤΟΥΡΓΙΚΕΣ ΚΑΙ ΤΕΧΝΙΚΕΣ ΠΡΟΔΙΑΓΡΑΦΕΣ ΙΔΡΥΜΑΤΙΚΩΝ ΑΠΟΘΕΤΗΡΙΩΝ» Έκδοση 1.1, 2013, διαθέσιμο μέσω του: http://seab.lib.ntua.gr/index.php?option=com_docman&task=doc_download&gid=12&Itemid=53&lang=el

Εθνικό Κέντρο Τεκμηρίωσης, *«Ηλεκτρονικά Αποθετήρια: Τι Είναι»,* http://www.openaccess.gr/repositories/what.dot/.

Dublin Core Collections Application Profile, 2007, available online at: http://dublincore.org/groups/collections/collection-application-profile/.

ETD-MS standard: http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html

Scholarly Works Application Profile, http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

XMetaDiss format: http://www.d-nb.de/eng/standards/xmetadiss/xmetadiss.htm

Final 1484.12.1-2002 LOM Draft Standard: http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html

PRISM Specification: http://www.idealliance.org/specifications/prism/

DCMI Metadata Terms: http://dublincore.org/documents/dcmi-terms/