Automatic Generation of Metadata for Learning Objects

Konstantin Nosov

Master's Thesis Project Description
Master of Science in Media Technology
30 ECTS
Department of Computer Science and Media Technology
Gjøvik University College, 2011

Avdeling for
informatikk og medieteknikk
Høgskolen i Gjøvik
Postboks 191
2802 Gjøvik

Department of Computer Science
and Media Technology
Gjøvik University College
Box 191
N-2802 Gjøvik
Norway

Automatic Generation of Metadata for Learning Objects

Konstantin Nosov

2011/07/01

Abstract

With the advent of a large number of educational resources in electronic forms, as well as the pervasiveness of online learning management systems, which uses these resources, a number of problems related to search and reuse of existing materials arise. For many years metadata is a good mechanism for describing the electronic resources. Metadata is data about data, information about information, and description of content which allows having the opportunity to find the necessary information, access it in an acceptable form.

The aim of this thesis is the analysis of automatic generation input for increasing the level of discovery and reusability of learning object. We have examined what contribution to the production of metadata for learning objects the curriculum information stored in the form of course syllabus and analysis of the harvestable metadata will provide. The syllabus field's analysis helped us to identify patterns in the structure of syllabus and determine the rules which the information will be extracted from. We studied and discussed what elements of metadata can be harvested and utilized for a detailed description of the learning object in addition to the information extracted from the course syllabus. As a result, we have learned to what extent the automatic generation of metadata for learning object using proposed resources can be utilized as a supplement to the manual input.

Preface

This report is my master thesis for the conclusion of my Master in Media Technology study in the Faculty of Computer Science and Media Technology at Gjøvik University College. I want to thank all the people who helped me during this project.

I especially want to thank my supervisor Rune Hjelsvold who gave me the opportunity to gain such experience, and without whose advices I could not cope with the problem. Thank you for your participation. I also want to thank Vera Nekrasova - my opponent and inspirer during all these months. Thank you for your concern.

And of course I cannot ignore my parents. Thank you very much for the moral support throughout my study in Norway.

Konstantin Nosov, 1st July 2011

Contents

Abstract . . . iii
Preface . . . v
Contents . . . vii
List of Figures . . . xi
List of Tables . . . xiii
1 Introduction . . . 1
1.1 Topic covered by the project . . . 1
1.2 Keywords . . . 2
1.3 Problem description . . . 2
1.4 Justification, motivation and benefits . . . 3
1.5 Research questions . . . 3
1.6 Summary of contributions . . . 3
2 Related work . . . 5
2.1 Learning Object (LO) . . . 5
2.2 Learning Object Repository (LOR) . . . 6
2.3 Metadata . . . 6
2.4 Metadata for Learning Object . . . 7
2.4.1 Metadata categories . . . 7
2.4.2 Metadata Schemas . . . 7
2.5 Automatic generation of Metadata . . . 10
2.6 Resources for Automatic Metadata Generation . . . 11
2.6.1 Course Syllabus . . . 11
2.6.2 Course Description Metadata (CDM) . . . 14
2.6.3 XCRI project . . . 15
2.6.4 OpenSyllabus model . . . 16
2.6.5 Keywords Extraction . . . 16
2.6.6 Language Identification . . . 17
2.7 Manual Metadata Producing . . . 18
2.8 How to define Metadata Quality? . . . 18
2.9 Uncertainty of Metadata . . . 19
2.10 Legal Aspects of Learning Objects . . . 20
2.11 Summary of Related Work . . . 21
3 Experiment Setup . . . 23
3.1 The General Purpose of Experiments . . . 23
3.2 Technical Aspect . . . 23
4 The Correspondence between Resources for Generation and DC Elements . . 25
4.1 Data Set . . . 25
4.2 Element "Description" . . . 25
4.2.1 Element Consideration . . . 25
4.2.2 Results . . . 26

4.2.3 Evaluation . . . 27
4.3 Element "Publisher" . . . 27
4.3.1 Element Consideration . . . 28
4.3.2 Results . . . 28
4.3.3 Evaluation . . . 28
4.4 Element "Creator" . . . 29
4.4.1 Element Consideration . . . 29
4.4.2 Results . . . 29
4.4.3 Evaluation . . . 30
4.5 Element "Contributor" . . . 30
4.5.1 Element Consideration . . . 30
4.5.2 Results . . . 32
4.5.3 Evaluation . . . 32
4.6 Element "Language" . . . 33
4.6.1 Element Consideration . . . 33
4.6.2 Results . . . 33
4.6.3 Evaluation . . . 33
4.7 Elements "Coverage", "Title", and "Relation" . . . 34
4.7.1 Elements Consideration . . . 34
4.7.2 Results . . . 38
4.7.3 Evaluation . . . 40
4.8 Element "Date" . . . 40
4.8.1 Element Consideration . . . 41
4.9 Element "Format" . . . 41
4.9.1 Element Consideration . . . 41
4.10 Element "Type" . . . 41
4.10.1 Element Consideration . . . 41
4.11 Element "Identifier" . . . 42
4.11.1 Element Consideration . . . 42
4.12 Element "Source" . . . 42
4.12.1 Element Consideration . . . 42
4.12.2 Results . . . 43
4.12.3 Evaluation . . . 43
4.13 Element "Rights" . . . 43
4.13.1 Discussion . . . 43
4.14 Element "Subject" . . . 44
4.14.1 Discussion . . . 44
5 Methods Testing . . . 45
5.1 Data Set . . . 45
5.2 Limitations . . . 45
5.3 Prototype . . . 45
5.4 Results . . . 48
5.5 Metadata Quality Evaluation . . . 48
6 Discussion . . . 53
6.1 Syllabus Benefits . . . 53
6.2 Syllabus Challenges . . . 53

6.3 How to Avoid or Improve Challenges 54
6.3.1 Analogy between Course Description Metadata and Dublin Core . . 55
6.4 Harvesting Issues 56
6.5 Whether Automatic Metadata Generation Fixes Human Errors 58
6.6 Generalization of Proposed Methods 58
6.7 Criticism of Chosen Methodology 59
7 Conclusion and Future Work 61
7.1 Sub-question1: What contribution textual resources in form of course syllabus that related to a given learning object can make to the automatic generation of Dublin Core metadata elements? 61
7.2 Sub-question2: What contribution the metadata harvesting can make to the automatic generation of Dublin Core metadata elements? 61
7.3 The main Research Question: To what extent the automatic generation of metadata for learning object using proposed resources can be utilized as a supplement to the manual input? 62
7.4 Future Work 62
Bibliography 63
A 60 Corse Syllabi 69
B 30 Learning Objects 71
C Results of the Second Part of Experiment 73
D Source Code Examples 75
D.1 Element "Title" generation 75
D.2 Element "Contributor" generation (names) 75
D.3 Element "Contributor" generation (e-mail information) 75
D.4 Table processing 76
D.5 Free-text syllabus processing 1 77
D.6 Free-text syllabus processing 2 77
D.7 Textbook information extraction 77

List of Figures

1 A schematic representation of the hierarchy of elements in the LOM data model [25] 8
2 Course syllabus structure 12
3 Detailed example of course syllabus [33] 13
4 CDM example [34] 14
5 Page header example 26
6 Tags statistics 26
7 Tags statistics 27
8 Publisher name extraction method 28
9 Results of Publisher extraction 29
10 Harvesting of “Creator” information 30
11 Course leader information statistics 31
12 Course assistant information statistics 31
13 Results of ”Contributor” extraction 32
14 Language information statistics 33
15 “Temporal information, lecture title, and related materials” representation statistics 35
16 Table structure example 36
17 Table tags structure examples 36
18 List structure example 37
19 List tags structure examples 37
20 Harvesting results of field Title 39
21 Extraction of basic literature information results 43
22 The concept of generation process 47
23 User interface example 47
24 Correctness criteria evaluation 49
25 IMT4931 Semantic Web in CDM format (Partly) 57

List of Tables

1 Dublin Core elements classification [27] 10
2 Information extracted from the table view of the course schedule 38
3 Information extracted from the free text paragraphs of the course schedule 39
4 Summed results 39
5 Applied methods 46
6 Analogy between CDM and DC 56
7 60 Corse Syllabi 70
8 30 Learning Objects 72
9 The second part of experiment results 74

1 Introduction

With the advent of a large number of educational resources in the electronic form, as well as the pervasiveness of online learning management systems, which use these resources, a number of problems related to search and reuse of existing materials arises. Worldwide educational institutions are increasingly aware of the need for rational use of the existing education materials and techniques to unify them available for sharing. Metadata is one of the highest priority mechanisms to improve the quality of search and as a consequence re-use of educational material. Many projects were designed to study a user's input in forms of free text "Folksonomies". But human nature is to tell lies, be lazy, to follow their habits, and have a subjective opinion [1]. This project aims to study the automatic generation of metadata for learning object based on curriculum resources that related to a given learning object and metadata that can be harvested directly from file properties. Also, the aim of this thesis is to explore whether this method of metadata production can be a useful addition to the manual input of metadata, what place it occupies in the process of creating metadata, and to analyze whether it can prevent or correct human errors.

This chapter is an introduction to this master thesis, where the topic and problem area are described. The rest of the project consists of Chapter 2, where related works and consideration of learning objects and metadata from different angles are presented. Chapter 3 presents the main goal of further experiments. Chapter 4 presents the first part of the experiment, where the selected resources and their application for the automatic generation of metadata are examined and discussed. Chapter 5 is the second part of the experiment, where is tested how many metadata records of Dublin Core can be automatically filled by using proposed methods per learning object and assessed the quality of the results. In Chapter 6 we discuss the obtained results, reveal benefits and challenges of the proposed methods and provide alternatives to solve them. Chapter 7 contains the conclusion for this thesis.

1.1 Topic covered by the project

There are a lot of definitions of what a Learning Object is. According to the definition of the IEEE LOM specification "Learning objects are defined as any entity, digital or non-digital, which can be used, reused, or referenced during technology supported learning" [2]. Wiley defines it as "A learning object is a digital resource that can be reused to facilitate learning." [3]. For Polsani, "A learning object is an independent and self-standing unit of learning content that is predisposed to reuse in multiple instructional context." [4]. Scientists agree that the learning object is a multimedia object which facilitates in the learning process. They also agree that the learning object should be reusable [5].

For many years metadata is a good mechanism for describing the electronic resources. Metadata is data about data, information about information, a description of the content which allows to have the opportunity to find the necessary information, access it in an acceptable form. The creators of information must be assured that their intellectual property rights will be protected, and administrators and other professionals should be able to support electronic information, such as ensuring its preservation for a long time. Metadata is a key component to solving these problems. Metadata can be created manually or automatically.

Learning Management Systems implicitly store a great amount of information in the context where the learning object is published: usually we have information about the course and lesson where the learning object is, the description of the task to be performed with the learning object, the navigation structure of previous and following material and etc. [6]. Also, every learning object is authored by one or more people. Quite often information about these people is available from different sources. If the learning objects are stored together with their metadata, available metadata can be used as a source for the newly introduced learning objects. This information is typically used if the new object is related to another object already stored in the system (as a new version of the existing one, for example) [7]. All this information can be a resource for automatic metadata generation.

A course syllabus is the skeleton of a course, curriculum subjects, including a description of the studied discipline, its goals and objectives, summary, topics, and the duration of each session, tasks of independent work, the consultation period, the requirements of teachers, evaluation criteria, schedule control, references and so on. Syllabus prepared on the basis of a typical curriculum. Later, a syllabus can be improved by adapting information from other relevant syllabi [8]. The various learning objects that are included in a course offering are created based on the syllabus definition, and are tightly integrated with the reference material (also included in the syllabus) [9].

Harvestable metadata is the metadata, that already exists in the source document. It may be created automatically by the system of manually by content producer. Such data can be harvested automatically by different existing tools (for example Exiftool).

The basis of the study in this project is an automatic approach to identify its positive and negative characteristics. As the sources for metadata generation for learning object the textual resources that related to a given learning object and harvestable information are used.

1.2 Keywords

Metadata, Learning Object, Automatic Generation, Syllabus, Extraction, Harvesting.

1.3 Problem description

Learning object plays a major role in the educational process. For many years electronic repositories of universities have accumulated huge amount of lecture materials. The problem of improving the reachability and reusability of these materials occurs acutely. Practice shows that the use of metadata to describe the resource can significantly increase its level of discovery. However, research results show that the one of the most common metadata standards for the educational purposes is the LOM standard, with its almost seventy elements, some of which are rarely used, is a serious barrier to the authors, and they try to avoid filling in all fields. On the other hand, some information which recorded in these attributes may depend on the subjective opinion of annotators. It is quite difficult to motivate the author of the learning object to the addition of metadata, because sometimes users searching the resource have more advantages than the author of the learning object.

Metadata serves not only to describe the content of the resource, but also open the domain of usage, and relationships to other learning objects. Scientists from article [10] argue that the process of annotation of the learning object cannot be attributed only to humans. They state that creation of structured metadata is too difficult, complicated and time-consuming for the authors of learning objects. Thus, this process requires special skills of the expert group. But with the increasing number of resources, time and cost of professional metadata creators are unacceptable for organizations.

Moreover, the metadata must possess the following quality criteria: completeness, correctness, accessibility, accuracy, provenance, conformance to expectations, logical consistency and coherence [11, 12]. "To err is human" (Alexander Pope). It is the human nature to lie, to be lazy, to follow their habits, have their own subjective opinion [1]. Low quality of metadata that belongs to the particular learning object can greatly reduce the level of discovery and as a consequence the level of reusability of the learning object. It is necessary to determine the sources of metadata.

1.4 Justification, motivation and benefits

Automatic generation of metadata reduce the efforts of the authors of learning object to add the information in those fields, which can be filled without his participation. Moreover based on the information already generated by the user or by the system it may be much easier to complete it at their discretion, confirm or reject it. Automatic generation can avoid grammatical mistakes made by the human, and do not take into consideration the subjective opinion of the person when making annotations for the resource. The high quality of metadata allows students as active users of LMS to obtain more representative information about the context of a learning object. Likewise, the high-quality metadata allows teachers to more effectively reuse the learning object, to create new courses or expand the content of existing courses. Responding to the criteria of quality metadata can increase the level of discovery and reusability of learning materials, systematize data in the Learning Object Repository.

1.5 Research questions

The main Research Question: To what extent the automatic generation of metadata for learning object using proposed resources can be utilized as a supplement to the manual input? Since we propose to use a combination of multiple resources for the automatic generation, we can distinguish the following sub question for greater detalization of the problem:

- Sub-question 1: What contribution textual resources in form of course syllabus that related to a given learning object can make to the automatic generation of Dublin Core metadata elements?
- Sub-question 2: What contribution the metadata harvesting can make to the automatic generation of Dublin Core metadata elements?

1.6 Summary of contributions

The aim of this thesis is the analysis of automatic generation input for increasing the level of discovery and reusability of learning object. It will be examined what contribution to the production of metadata for learning objects the curriculum information stored in the form of course syllabus, and analysis of the harvestable metadata will provide. Syllabus field's analysis will help us to identify patterns in the structure of syllabus and determine the rules by which the information will be extracted from this fields. We will study and discuss what elements of metadata can be harvested and utilized for a detailed description of the learning object as a supplement to the information extracted from the course syllabus. As a result, we will learn whether the automatically generated metadata using our approaches could be the supplement for manual annotation of learning object and to what extent.

2 Related work

2.1 Learning Object (LO)

For the first time the concept of LO was described in 1967 by Gerard. However, the term itself is attributed to Wayne Hogins - futurist who works with educators around the world introducing new models of learning, learning management system, and claiming the role of technology in learning. Throughout history, LO changed many names: chunks, nuggets, reusable information objects, units of learning and others [13]. Today, IEEE Learning Technology Standards Committee [14] defines LO "as any entity, digital or non-digital, that may be used for learning, education or training". Wiley in [15] speaks of the LO as a "digital resource that can be reused to support learning be it large or small". This category includes digital images or photos, live or prerecorded audio or video, text files, and small web applications like a Java calculator. However, in [4] proposed the following description of the LO: "Examples of Learning Objects include multimedia content, instructional content, instructional software and software tools that are referenced during technology supported learning. In a wider sense, Learning Objects could include Learning Objectives, persons , organizations, or events ". According to the authors of this article, any digital object be it a painting or musical composition evokes emotions from the user, forces him to think, thus turning the user into learner.

According to the users of LO community, LO should possess the following properties. The first property - *Accessability*, imply that LO should be annotated with metadata, and thus can be stored and referenced in a database. The second property - *Interoperabiluty* - the LO should be independent of both the delivery media and knowledge management systems. And the third property - *Reusability* - once created, a LO should function in different instructional contexts [4]. This property has a very high value in the production of LO. In [16], the authors present the benefits for authors of LO as well as for learners. For authors, they point out the following benefits: reusable learning object specific templates ensure that design and development is consistent across the organization, author write effective and efficient job / task, authors can reuse any reusable learning object in future development, and authors can combine old and new reusable resources to form larger structures. In the case of learners, divided into the following benefits: Reusable LOs act as a job aid or performance support tool, giving the learners just-in-time access to training and information. Delivery modes are customized to best match the individual learning style of the learner. Custom learning paths are tailored to the knowledge and skills the individual learner needs to acquire.

The authors of [17] relate the concept of LO with the metaphor "LEGO" or with other children's hobbies. With this metaphor, they are trying to present the theory of learning object: "create small pieces of instruction (LEGOs) that can be assembled (stacked together) into some larger learning-facilitating structure (castle or spaceship)". In other words, any learning object (LEGO block) may be combined with other learning object (another LEGO block) to create a new unity of knowledge.

So, to sum up what a Learning Object is, let's use the following specifications proposed by Robert Beck:

- "Learning objects are a new way of thinking about learning content. Traditionally, content comes in a several hour chunk. Learning objects are much smaller units of learning." This tells us that a particular LO is a small component of the lesson.
- "Learning objects are self-contained - each learning object can be taken independently." Each LO can be considered individually without reference to other LO.
- "Learning objects are reusable - a single learning object may be used in multiple contexts for multiple purposes." In other words, LO could be the basis for a new LO or expand existing ones.
- "Learning objects can be aggregated - learning objects can be grouped into larger collections of content, including traditional course structures." In other words, learning object can be presented in a mixture of traditional lectures and LO of other formats.
- "Learning objects are tagged with metadata - every learning object has descriptive information allowing it to be easily found by a search." Quite important feature allowing using and reusing of learning object.

2.2 Learning Object Repository (LOR)

Learning object technology allows us building of repositories that constitute a kind of specialized digital libraries where high quality re-usable materials can be selected by teachers, on the basis of a description of their content (usually called metadata) as to educational features, context of use, technical aspects, and so on. A LOR allows registered or unregistered users to search and retrieve LO's from the repository. A LOR typically supports simple and advanced queries, as well as browsing the material by subject or discipline [18].

2.3 Metadata

The most common definition of metadata is "data about data". This term is used when describing the semantic information about online resources and a concise description of the form and content of the resource [19]. Metadata is also useful to provide textual descriptions for non-textual objects, for example, to enable the representation of multimedia document properties in a structured way, simplifying document management and retrieval [20]. General interest to the practical application of metadata associated with an increase of electronic publications, the emergence of new individual and organizational sites, and plenty of undifferentiated digital data available on the network. Search on metadata, in contrast to the details of information resources is more efficient, since metadata providing looking for information about content of the resource. Properly used metadata can identify the name of the resource, the creator, who reformatted it, and other descriptive information. Metadata plays various roles in digital information system [21]:

Accessibility: Effectiveness of searching can be significantly enhanced through the existence of rich and consistent metadata. Metadata can also makes possible to search across multiple collections or to create virtual collections from materials that are distributed across several repositories, but only if the descriptive metadata are the same or can be mapped across each site.

Interoperability: Describing a resource with metadata allows it to be understood by both humans and machines. Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data accurately. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.

Multi-versioning: Objects enter a digital information system by being created digitally or by being converted into digital format. Multiple versions of the same object may be created for preservation, research, dissemination, or even product development purposes. The creator may include some administrative and descriptive metadata for this purpose.

Legal issues: Metadata allows repositories to track the many layers of rights and reproduction information that exist for information objects and their multiple versions. Metadata also documents other legal or donor requirements that have been imposed on objects – for example, privacy concerns or proprietary interests.

System improvement and economics: Metadata is also helpful to evaluate and refine systems in order to make them more effective and efficient from a technical and economic standpoint. The data can also be used in planning for new systems.

2.4 Metadata for Learning Object

2.4.1 Metadata categories

Metadata can be classified into four categories. *Discovery Metadata* includes all attributes that support the ability to find the learning objects. *Use Metadata* contains all attributes that are meaningful while a learning object is used. This includes technical information like format or system requirements as well as intellectual characteristics as property rights or restrictions regarding the usage. *Authentication Metadata* involves attributes that guarantee the integrity and the overall trustfulness of a learning object. Attributes like the source of a learning object, its version or the relation to other objects are grouped here. *Administration Metadata* includes attributes supporting the management of a learning object as information about ownership or all meta-metadata (e.g. who created the metadata records) [22].

2.4.2 Metadata Schemas

A metadata record consists of a number of pre-defined elements representing specific attributes of a resource, and each element can have one or more values. These elements form the metadata scheme. Each metadata schema will usually have the following characteristics: a limited number of elements, the name of each element, the meaning of each element. Typically, the semantics is descriptive of the contents, location, physical attributes, type (e.g. text or image, map or model) and form (e.g. print copy, electronic file). Key metadata elements supporting access to published documents include the originator of a work, its title, when and where it was published and the subject areas it covers. Where the information is issued in analog form, such as print material, additional metadata is provided to assist in the location of the information, e.g. call numbers used in libraries. The resource community may also define some logical grouping of the elements or leave it to the encoding scheme [23].

Metadata schema created for only one particular application violates the rules of searchability, extensibility, reusability, and scalability. There is need to have a single standard that allows to use and exchange the media information [19].

IEEE LOM Standard

Standard for describing objects is IEEE Learning Object Metadata. The standards model is presented in a hierarchical structure. At the top of this hierarchy is the "root" element. This root element contains many sub-elements. If a sub-element itself contains additional sub-elements it is called "branch". Sub-elements that do not contain any sub-elements are called leaves ". This hierarchical model is called the "tree structure" of a document. This standard defines the syntax and semantics of Learning Object Metadata [19]. IEEE LOM take a structuralize approach to metadata creation.

The standard has 60 elements which provides a more universal and profound response to learning object. These elements are organized into 9 categories [24]:

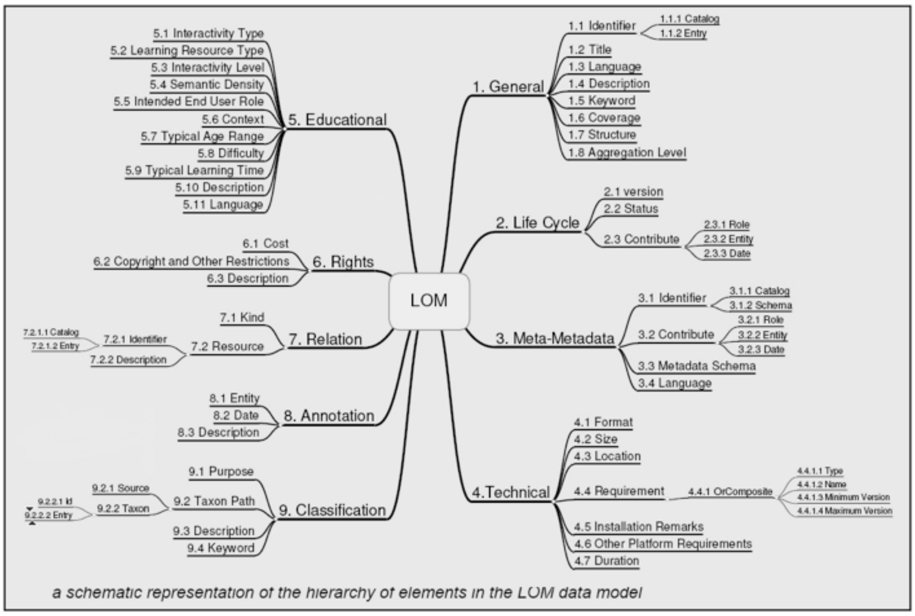


Figure 1: A schematic representation of the hierarchy of elements in the LOM data model [25]

- General. This category covers all general information that describes this learning object as a whole.
- Life cycle. This category describes the history and current state of this learning object and those entities that have affected this learning object during its evolution.
- Meta-metadata. This category describes how the metadata instance can be identified, who created this metadata instance, how, when, and with what references.
- Technical. This category describes the technical requirements and characteristics of this learning object.
- Educational. This category describes the key educational or pedagogic characteristics of this learning object. This is the pedagogical information essential to those involved

in achieving a quality learning experience. The audience for this metadata includes teachers, managers, authors, and learners.

- Rights. This category describes the intellectual property rights and conditions of use for this learning object.
- Relations. This category defines the relationship between this learning object and other learning objects.
- Annotation. This category provides comments on the educational use of this learning object, and information about when and by whom the comments were created.
- Classification. This category describes where this learning object falls within a particular classification system.

Dublin Core

Dublin Core (DC) is a competent and successful projects associated with the development of the structure of Meta descriptions of resources. The Action Group (Dublin Core Metadata Initiative, DCMI) has adopted a number of precise conceptual solutions enable them to find an acceptable compromise between expressiveness and simplicity, naturalness and completeness of meta descriptions. In comparison with LOM Standard, DC standard includes 15 well-defined elements for describing "core" information properties [26]:

- Title - An entity responsible for making contributions to the resource.
- Creator - An entity primarily responsible for making the resource.
- Subject - The topic of the resource. Typically, the subject will be represented using keywords, key phrases, or classification codes.
- Description - An account of the resource. Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
- Publisher - An entity responsible for making the resource available. Examples of a Publisher include a person, an organization, or a service.
- Contributor - An entity responsible for making contributions to the resource. Contributor include a person, an organization, or a service.
- Date - A point or period of time associated with an event in the lifecycle of the resource.
- Type - The nature or genre of the resource.
- Format - The file format, physical medium, or dimensions of the resource.
- Identifier - An unambiguous reference to the resource within a given context.
- Source - A related resource from which the described resource is derived. The described resource may be derived from the related resource in whole or in part.
- Language - A language of the resource.

- Relation - A related resource.
- Coverage - The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
- Right - Information about rights held in and over the resource. Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.

Dublin Core provides the following classification of elements of vocabulary (Table 1):

Content	**Intellectual Property**	**Instantiation**
Coverage	Contributor	Date
Description	Creator	Format
Type	Publisher	Identifier
Relation	Rights	Language
Source		
Subject		
Title		

Table 1: Dublin Core elements classification [27]

Dublin Core Metadata Initiative provides guidelines for encoding Dublin Core metadata in XML and RDF / XML to enable interoperability across different platforms, languages, and systems [28].

2.5 Automatic generation of Metadata

Two main alternatives exist to index the learning objects in a LOR. It could be done either by manual or by automatic indexation. In the first alternative, an expert, after reviewing the learning object, generates the metadata values. In the second alternative, some kind of information extraction system tries to deduce the value of the metadata fields based on the information available about the object [6].

The automatic approach to create metadata allows the use of computing power with a higher level than simply filling the input fields for manual annotation. Automatic metadata generation extracts relevant information from particular learning object and from context that stored in this learning object [22]. Accordingly, distinguished resource-based (content-based) and context-based methods.

Metadata harvesting belongs to the resource-based method and means that the metadata is automatically collected from fields already filled. An example of such information may be size or file format. *Extraction* is a text-based method and means automatic extraction of information from the content of the resource. Keywords or phrase extraction the is most common technique. Having already available information metadata can be obtained from it. For example the source of such information may be course profile or lecture description [10].

Along with the automatic generation of metadata systems there are hybrid systems, which establish a balance between automatic and human metadata generation. Based on the results of automatic analysis of the learning object, three groups of information created: very probable values, which include specifications of the object and usually do not require any user intervention, probable values, which may be considered as an offer systems for metadata entry, but are not reliable and require verification user, and restriction of possible values. The latter group does not serve the sentence, but reduce the scale of possible values [22].

System has to allow easy share, reuse, and group learning object in order to create different context. Interoperability of the system allow to different learning systems to communicate with other systems to share and mix their learning resources [29]. Automatic metadata generation is more efficient, less costly, and more consistent than the human-oriented process [30].

2.6 Resources for Automatic Metadata Generation

For any automatic indexing we need some kind of already available information about the LOs to index. Fortunately, Learning Management Systems implicitly store a great amount of information in the context where the LO is published: usually we have information about the course and lesson where the LO is, the description of the task to be performed with the LO, the navigation structure of previous and following material and etc. [6].

Also, every learning object is authored by one or more people. Quite often information about these people is available from different sources. A creator or a indexer profile groups this information, so that it can be used when generating metadata for a document of that person. If learning objects are stored together with their metadata, available metadata can be used as a source for newly introduced learning objects. This information is typically used if the new object is related to another object already stored in the system (as a new version of the existing one, for example). Moreover, similarity searches can be used to search for similar objects in the system, so that their existing metadata can be used to create new metadata.

Learning management systems can provide rich contextual information, like the courses in which the object is used, how many times the document was used or downloaded, etc. As such, it actually does both document context analysis and document usage analysis [7].

All this information could help in the design and implementation of information extraction systems. Further described an idea of in which form the curriculum information can be presented and available among the learning community.

2.6.1 Course Syllabus

Etymologically syllabus means a "label" or "table of contents." The *American Heritage Dictionary* defines syllabus as outline of a course of study [31]. One of the first steps taken by an educator in planning a course is to construct a syllabus. A course syllabus is the skeleton of a course, curriculum subjects, including a description of the studied discipline, its goals and objectives, summary, topics, and the duration of each session, tasks of independent work, the consultation period, the requirements of teachers, evaluation criteria, schedule control, references and so on. Syllabus are prepared on the basis of a typical curriculum. Later, a syllabus can be improved by adapting information from other relevant syllabi [8].

The authors of [31] classifiy information in syllabus into two groups: information that students need to have "at the beginning of the course" and the second include all information that students need to have "in writing". According to them, the major content areas of syllabus are following:

- **Course Information.** The first items of information in syllabus should give course information: course title, course code, and credit hours. Sometimes it can be prerequisites, language of instruction, and location of classroom.
- **Instructor Information.** Second, the students need information about the instructor: full name, title; office location, office phone number. Many instructors give the students their e-mail address.
- **Reading materials.** Textbook(s) include the title, author, edition, and publisher.
- **Course Descriptions/Objectives.** The treatment of the area (course description, content, goals, objectives).
- **Course Calendar/Schedule.** The topics of daily or weekly lectures is covered and all related materials, such as lecture notes and supplementary reading. The schedule also should include the dates for exams, quizzes, assessments, or due dates for major assignments.
- **Course Policies.** Attendance, lateness, class participation, missed exams or assignments, lab safety/health, academic dishonesty, grading, and available support services.

For clarity, the content areas of syllabus are presented in the Figure 3.

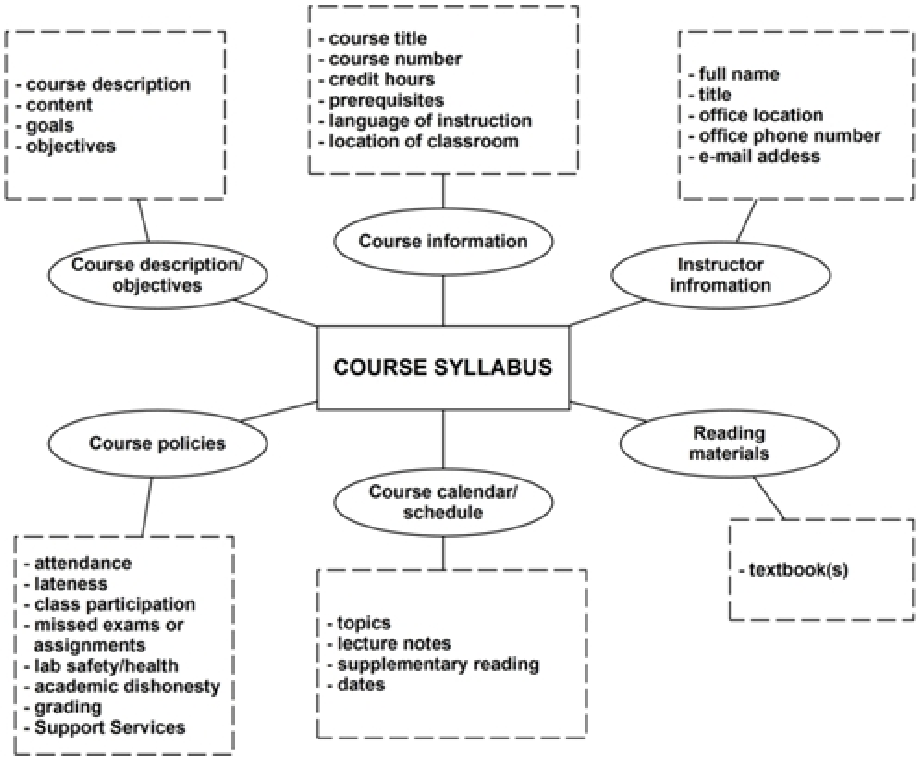


Figure 2: Course syllabus structure

The same idea of syllabus design share [32]. They argue that the syllabus should play three roles: Syllabus as a Contract, Syllabus as Permanent Record, and Syllabus as a Learning Tool. In the article "The Purposes of Syllabus" they describe each element of syllabus and give the most detailed examples of the course syllabus. The example is shown in the Figure 4.

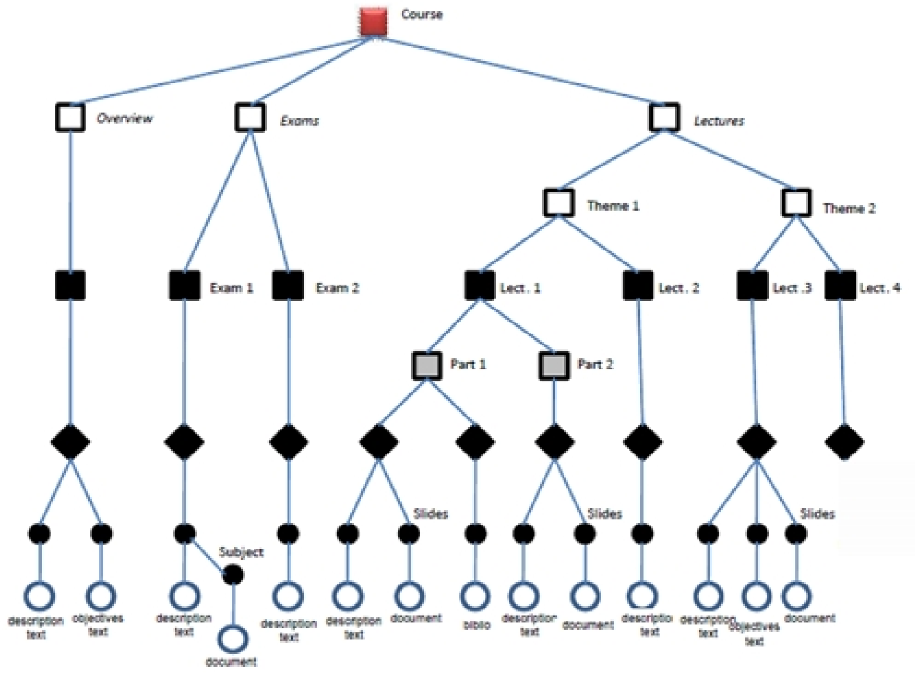


Figure 3: Detailed example of course syllabus [33]

According to the Parkers & Harris (2002), the course syllabus should include a detailed calendar of the course. Having this information syllabus provides information about how to plan for the tasks and experiences of the semester, how to evaluate and monitor one's performance, and how to allocate time and resources to areas in which more learning is needed.

Syllabus Publication

Different schools have their own preferences and rules for syllabus publication. Syllabus may be published by website administrator through the existing management system or presented in paper form. Most universities maintain a standardized course catalog that contains syllabus definitions for all courses offered at that school. Some of them include course summary information in the catalog, while the details are only available in university records systems. This information, however, is not in a standardized format across all universities, and usually contains only the most basic course information without details of learning objects [9]. In the case of online publication syllabus can be in HTML or printable format such as PDF, MS Word format, or Open Office.

Different schools set different access levels to the syllabus. Some allow public reading and viewing from the local subnet or from another network. However, efforts are currently underway to make syllabi and course content available online licensed under an open license [9].

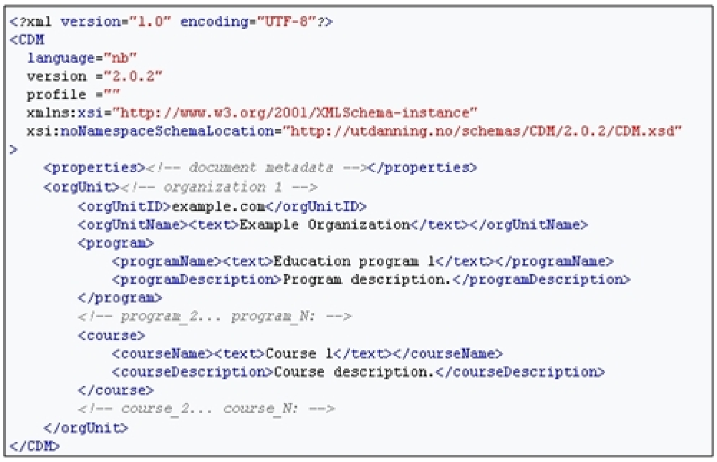
2.6.2 Course Description Metadata (CDM)

The CDM specification was developed in 2001 by USIT's XML group at the University of Oslo for the Norwegian eStandard project, Norway Opening Universities (OUN), a national initiative for change and innovation in Norwegian higher education.

CDM addresses the description of educational course units or other forms of educational offering at all levels. It specifies the structure and semantics of the key concepts used in course descriptions. The metadata are specified as an XML Schema, and guidelines with examples are given to facilitate the generation of course descriptions as XML documents.

The metadata are intended to satisfy the following objectives:

- Facilitate description and exchange of information about educational course units
- Facilitate a standardization of course unit descriptions
- Facilitate the establishment of national and international course catalogues
- Facilitate the establishment of course portals and other services helping students



```
<?xml version="1.0" encoding="UTF-8"?>
<CDM
  language="nb"
  version ="2.0.2"
  profile =""
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://utdanning.no/schemas/CDM/2.0.2/CDM.xsd"
>
    <properties><!-- document metadata --></properties>
    <orgUnit><!-- organization 1 -->
        <orgUnitID>example.com</orgUnitID>
        <orgUnitName><text>Example Organization</text></orgUnitName>
        <program>
            <programName><text>Education program 1</text></programName>
            <programDescription>Program description.</programDescription>
        </program>
        <!-- program_2... program_N: -->
        <course>
            <courseName><text>Course 1</text></courseName>
            <courseDescription>Course description.</courseDescription>
        </course>
        <!-- course_2... course_N: -->
    </orgUnit>
</CDM>
```

Figure 4: CDM example [34]

According to the latest release CDM-2.1 (2009-09-01), CDM are divided into four parts [35, 36]:

- Organisation unit *(orgUnitType)*. An (or part of an) organisation responsible for accomplishment of courses, e.g. university, faculty, institution.
- Study program *(programType)*. A study program comprising a set of course units.
- Course unit *(courseType)*. A course unit with curriculum, time schedule, teaching activities and exam.
- Person *(personType)*. Contact information of persons related to the accomplishment of courses.

Organisational Unit. An element of type *orgUnitType* represents an organisational unit that organises or provides study programs and courses. In general, a given program or course may be organised/provided by multiple organisational units. An organisational unit can have a hierarchical structure with subordinate organisational units (e.g. university, faculty, institute). The concept of organisational unit is supposed to encompass all organisational structures with educational offerings from a traditional university to a loosely defined consortium. Part elements include the following: *orgUnitName* - Full name of the organisational unit, *orgUnitCode* - An organisation unit code according to a codification scheme, contacts - A set of contact information associated with the organisation unit, and other components relevant to a particular organizational unit.

Study Program. An element of type *programType* contains the description of a study program. Part elements include the following:*programName* - Full name of the study program, *qualification* - Learning outcomes, skills, competencies, marks and/or grades obtained, rights to practise and/or professional status at different levels accorded to the holders of the qualification, *level* - Level of the program, eg undergraduate, bachelor, master, *formalPrerequisites* - Description of any formal prerequisites for the study program, etc.

Course Unit. An element of type *courseType* contains the description of a course unit. The term course refers to a complete unit of instruction that provides the learners with the knowledge or skills required for competence in a subject matter. A course is any academic or vocational course arranged by a course provider. This is the lowest level that can offer credits or recognition within an educational institution. A course usually includes teaching activity and examination. Part elements include the following: *courseName* - Full name of the course unit, *courseDescription* - A general description of the course unit, *instructionLanguage* - Main language of instruction, *syllabus* - Information on syllabus, textbooks/literature prescribed for study, etc.

Person. An element of type *personType* contains the description of a person. The focus of the description is contact information of persons related to the accomplishment of courses. Part elements include the following: *name*, *title*, *role*, *contactData*, etc. [36].

2.6.3 XCRI project

The project was managed by Manchester Metropolitan University[1]. The goal of XCRI project is to provide definitive specifications that describe accurately the learning opportunities that will be offered in particular locations at particular times. This project aims to lay foundations for an open, service-oriented approach to managing and utilizing course information by developing a suitable vocabulary and technology bindings for describing relevant data and demonstrating how such data could be managed, retrieved and transformed for different audiences using web services [37]. The project was motivated by the report of National Committee of Inquiry into Higher Education, which stressed 'the importance of clear and explicit information for students so that they can make informed choices about their studies and the levels they are aiming to achieve' [38].

The XCRI project is built on the Norwegian Course Description Metadata project to define a vocabulary to describe course-related information in a way that fits UK needs. The vocabulary encompasses course marketing, course quality assurance, enrolment and reporting and personal development requirements. XCRI the general requirements of

[1] www.mmu.ac.uk/

formality and interoperability: this is open specification using the XML schema formalism [39].

2.6.4 OpenSyllabus model

OpenSyllabus is a model of syllabus organization supported by an XML representation. For the basis of OpenSyllabus model have been taken the model proposed by [32] described earlier and extended by seven elements found in the major number of syllabi: News, Staff information, Course overview, Assessment, Lecture list, Frequently Asked Questions. The structure model of OpenSyllabus is quite flexible and suitable for each university to easily parameterize the various elements of the model as well as the vocabulary used while keeping a semantic suitable for sharing across platforms [33].

2.6.5 Keywords Extraction

Moreover, learning object by itself can contain a wide range of information to fill in the fields of metadata schema. Keywords can be considered as condensed versions of documents, which can play important role in some text processing tasks such as text indexing, summarization and categorization. Therefore, it is possible to think of them as a set of phrases semantically covering most of the text. However, there are many digital documents especially on the Internet that do not have a list of assigned keywords. Assigning keywords to these documents manually is a difficult task and requires appropriate knowledge of the topic. Automatic keyword extraction process can solve this problem [40]. [41] motivates that the automatic generated metadata in the form of keywords may improve the discovery of potential relevant learning object or can be used to assist or supplement the manual assignment of subject terms. Likewise, keywords extracted from the learning object and related materials can be organized in the ontological structures that can be used in an effective way of navigating within the learning object repository. Ontology can be used to find quite specific concepts or the most extensive (generalization) concepts matching to the specified search.

Preprocessing

The information that is being analyzed very often is irrelevant and redundant, contains noisy and unreliable data, thus Text Mining process is becoming difficult. Data preprocessing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set [42].

Tokenization

The first step in preprocessing is the so-called tokenization, i.e. the process of breaking a stream of text into *tokens*. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. To this were used regular expressions [43].

Stop words removing

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. The general strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often handfiltered for their semantic content relative to the domain of the documents being indexed, as a stop list, the members of which are then discarded during indexing [43].

Words Normalization

Stemming. For grammatical reasons, documents are going to use different forms of a word, such as *organize*, *organizes*, and *organizing*. Additionally, there are families of derivationally related words with similar meanings, such as *democracy*, *democratic*, and *democratization*. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

WordNet Based Method. The dictionary of wordNet consists of four major networks for significant parts of speech: nouns, verbs, adjectives and adverbs. Basic unit in WordNet dictionary is not a single word, a so-called synonymous series ("sinsets"), and combining words with similar meaning and are inherently network nodes [44]. Keyword extraction can be limited by simple nouns processing, simply comparing each word in the document with a set of words presented in WordNet dictionary. Reason for this is that nouns are very informative and could be treated as possible keywords [45].

2.6.6 Language Identification

Without the basic knowledge of the language the document is written in, applications such as information retrieval and text mining are not able to accurately process the data, potentially leading to a loss of critical information [46]. The task of determining the language of the text arose long ago, and many attempts have been committed to solve this problem.

Techniques to identify the language can be classified into three types: «Linguistically-grounded methods», «Similarity-based methods” and «classification methods” [47].

Linguistically-grounded methods. This method includes the following techniques: The investigated text is checked for the presence of a certain set of "special characters" [47]. Another technique is to check the text for the presence in it of the sequence of characters unique to a specific language. The article [48] provides an analysis of such sequences. Also an early attempt was the analysis of the text to identify in it stop words inherent in a particular language (the concept of stop word described earlier). The advantage of such methods for determining the language is its cheapness and simplicity. The disadvantages include the fact that the sets of symbols and words on which the analysis is based, may overlap in various languages, which reduces the accuracy of the results.

Similarity–based Methods. The core of these methods is the application of N-gram counting. An n-gram is a subsequence of n items from a given sequence [49]. Canvar and Trenkle [50] presented their results using rank order statistics. Language of the text is determined by comparing the list of common short string found in the classified text with a sheet of the training set for the different training corpora.

Classification Methods. The basic of this method is to develop a set of character level language models from the training data and then to use this language models to estimate the likelihood that a particular test string might have been generated by each of this language models [48]. This class of methods involves compare Bayesian probability of distribution [50], compare entropy of distribution [51]. The disadvantage of these methods is that for the analysis of much training data is required and classification can be slow.

2.7 Manual Metadata Producing

Search and navigation among the web resources have been largely improved with the introduction of collaborative tagging. The process of collaborative tagging is regarded as the process of adding user metadata in the form of keywords to the content which will then be saved or published [52]. This process involves the introduction of folksonomy. "Folksonomy" is a combination of two words "folk" and "taxonomy". "Folk" mean collaborative and progressive definition of a relaxed categorization and organization of content "taxonomy". The folksonomies has three main components - the user, tag and resource. The user can use the tags for tagging capacity of his favorite resources (photos, videos, web pages) [53]. But finding the right information does not always produce the desired result. Ambiguity of the language implies that one tag may refer to several concepts or several tags may refer to the same concept due to variability of the spelling, the lack of explicit representations of the knowledge contained in folksonomies, the difficulties to deal with tags from different languages [54].

Likewise, some users simply do not use this opportunity to participate in tagging or use it absolutely wrong. This is due to the lack of specific knowledge or the rules of filling tags. Everyone thinks differently and the logic of one person is difficult to understand others.

2.8 How to define Metadata Quality?

The automatic generation has an obvious relevant advantage: the economy of work for not having to “manually” create the metadata, apart from standardizing the value of the metadata. Nevertheless, it has also a serious weakness related to the quality of the generated metadata. Metadata is useful not only when it is complete and accessible but also when it is correct. However, the correctness of some metadata values, especially those related to educational characteristics is very difficult to verify, because such information deals with the intended usage of the material, which remains implicit in the material itself [11].

Further some characteristics of metadata quality are followed:

Metadata Completeness – in order to provide full access to an educational resource, it has to be ensured that all the information is annotated with the metadata. Otherwise, important or useful parts of an information source may be missed or cannot be indexed correctly [55]. Estimating the completeness of metadata requires at least make sure whether all the elements of the scheme completed.

Metadata Correctness - Metadata about an educational resource is only useful if it correctly describes its contents and pedagogical contexts. In fact, inconsistent metadata may be a more difficult problem than missing metadata, “because mechanisms relying on metadata will produce wrong results without warning” [55]. Therefore, validity of metadata should depend on the evaluation of the correctness of this metadata. Correctness evaluation is more complex than completeness assessment because it deals with the semantics of the metadata values [11].

Metadata Accessibility - Metadata for educational resources has to be accessible for people and applications wanting to use them in order to be useful. Accessible metadata has the following features: (1) it is defined within an interoperable format (e.g. XML),

(2) it uses an accessible vocabulary (e.g. RDF), and (3) it is possible to localize [11].

Also [12] bring the following characteristics:

Accuracy. The accuracy is the degree to which the metadata values are “correct”, i.e. how well they describe the object. The correctness could be a binary value, either “right” or “wrong”, for objective information like file size or language, but, in the case of subjective information, it is a more complex spectrum with intermediate values (e.g.: A title of a picture, or the description of the content of a document).

Provenance. Provenance covers the reputation that a metadata record has in a community. For example, a user may trust more metadata generated by a metadata expert that he knows, than metadata generated by a software tool. While the automated generated metadata may be of a better quality (according to the other metrics), provenance is more related to the subjective perception that the user has about the origin of the metadata.

Conformance to Expectations. The conformance to expectations measures the degree in which the metadata record fulfills the requirements of a given community of users. There are several parameters that affect this quality: The vocabularies words used in the metadata record should be meaningful for the user, the metadata fields filled the ones needed to perform the task that the user intended (search, evaluation, integration, etc), the amount of information is enough to describe the learning object.

Logical consistency and coherence. The logical consistency and coherence is the degree to which a metadata record matches a standard definition and the values used in the fields, correlate positively among them.

2.9 Uncertainty of Metadata

In his article "Metacrap: Putting the torch to seven straw-men of the meta-utopia" [1] Cory Doctorow argues that "reliable metadata would be a utopia. It’s also a pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities". He cites seven obstacles for production and use of reliable metadata:

- People lie - in many cases, content producers to use misleading information about their products to attract consumers’ attention.
- People are lazy - quite often the content producers do not use the opportunity to give names to their works.
- People are stupid - "Even when there’s a positive benefit to creating good metadata, people steadfastly refuse to exercise care and diligence in their metadata creation."
- People follow their habits.
- Schemas aren’t neutral - there is no single approach to the categorization of ideas, objects, attributes.
- Metrics influence results - Agreement on the use of common criteria for measuring the important parameters in a certain field must put in the privileged position of those things that get better results when using this method of measurement.
- There’s more than one way to describe something.

2.10 Legal Aspects of Learning Objects

The emergence of learning objects in an electronic form gives rise to new issues regarding the ownership and distribution of it in a very public manner. These learning objects are subject to Information Property Rights (IPR) law. Specifying the IPR information is very important for repositories, and should contain records of who owns the resource, who has access to the resource and under what conditions this resource is distributed.

Fair Use. Fair use is an affirmative defense to copyright infringement. Fair use is based on two factors. One of the factors considered in a fair use assessment is whether the use proposed is non-profit or commercial. Another factor asks if the proposed use impacts the market for or value of the original work [56].

Copyright. The owner of the copyright in an intellectual work enjoys the right to grant or withhold the right to others to make copies of the work; copyright is often described as a restrictive right because it is concerned with stopping others doing something with the work. It is very important to point out that when a piece of intellectual work is created and fixed in a material form such as in a drawing, a video recording, notes, or a printed text etc then the person, whom the law identifies as the 'author' and first owner, can enjoy and claim the protection of copyright law immediately. Copyright also applies to collaborative works that do not easily allow identification of a single author. Where two or more people have created a single work protected by copyright, those people are generally joint authors and joint first owners. However, if the work is created during the course of employment, the employer owns the copyright [57].

In [56] considered some cases when professor uses existing materials to create new learning object. The first and most common case is the use of material created and owned by others. Professor uses commercially available material or other material the rights of which are owned by others. Examples of such material would be charts and graphs from books, commercially available maps or images, or other textual content. Professor may not adapt the material (including translation) without permission of the copyright holder. Another case involves the use of materials created by students. Students automatically own the copyright in the works they create, even if the work is created during the course of instruction. Another case described in the article is the use of public domain materials. The public domain material may be freely adapted. The owner of the copyright in the resulting learning object owns only the material added to the public domain material. The original public domain material remains in the public domain forever.

Let's look at the existing license, under which existing materials can be used and new learning objects may be released:

Creative Commons. The purpose of Creative Commons is to allow copyright holders to pass some of the rights to their works to the public, and at the same time retain other rights. The fact that in accordance with the currently in force in most countries copyright laws all rights as property as well as non-proprietary belong to the authors automatically. Creative Commons makes it possible to pass certain rights of the public by the family of ready-made licenses recognized juridical legislation of many countries. Thus, the goal of Creative Commons is to promote the free flow of information, although not all Creative Commons licenses are free licenses. Creative Commons licenses are provided with tags written on XML, which allows a program for viewing Web pages to find this information [58].

GNU General Public License. The purpose of GNU GPL is to give the user the right to copy, modify and distribute (including commercial) programs (by default, is prohibited by copyright law), as well as ensure that users of all derivatives of the above programs receives the rights. The principle of "inheritance"of the rights is called "copyleft" [59].

Design Science License is a copyleft license for free content in the form of text, images and music. The license requires that any modification protected with DSL is published under the same license, without any new restrictions on its distribution / modification, derivative work will be named differently to distinguish it from the original, the new product will be correctly stated with authorship (What specific parts created by original author, what by new authors, as well as all other changes and its dates) [60].

2.11 Summary of Related Work

This chapter presents an overview of existing literature, researches and technologies with respect to research questions defined in Introduction part. Let's summarize all facts and apply them to the defined Research Questions.

Scientists find that the most important characteristic of learning objects is reusability, i.e. it serve as bases for new or extension of existing learning objects. This characteristic is particularly useful for creators of learning objects. Also, it acts as a support tool for students. There is no doubt that metadata play an important role in the reusability of learning objects, storing and maintaining them over time, as well as protection of intellectual property rights.

Automatic generation of metadata requires some information resources. In this project we consider the course syllabus and harvestable metadata as resources for generation. While working with the scientific literature, these resources have been considered from various angles and found their capacity to participate in the automatic generation of metadata.

There are several standards for representation of metadata elements. The current study is based on Dublin Core metadata standard. The choise is based on the following facts: Dublin Core scheme is usable and flexibile, provides a clear semantics of the elements for a wide range of users, provides a description of core feaures of electronic resources that support resource discovery. Furthermore, all Dublin Core elements are optional, but each system is capable to determine its own set of mandatory elements. It can be extended to meet demands of more specialized communities [21]. Also, while working with the scientific literature, we have not found work in the field of application of the course syllabus as a resource for the automatic generation of metadata. Dublin Core metedata element set is the initial step in this area to test this assumption.

If we follow [1], there are some obstacles to the use of metadata as the sole mechanism for describing resources. Automatic generation of metadata is not error free, but is able to solve some problems of "meta-utopia". In this study, we use such approaches to the automatic generation of metadata as metadata harvesting and extraction.

Since the metadata are used for guidance and compliance with intellectual property rights, were considered the highlights of this issue, use cases, and existing licenses during the literature review.

An important factor is not only the existence of metadata, but also its quality. In my project I would based on the following characteristics of metadata quality: metadata completness, correctness, accuracy and others that described in Section 2.8.

3 Experiment Setup

This thesis focuses on studying the contribution of the information resources in the form of courses syllabus and metadata harvesting from learning objects in the process of automating the generation of metadata. For this purpose we need to find the most effective approach to generate metadata by using these resources.

The experiment is divided into two parts. The first part of the experiment includes consideration of methods for extraction of the information suitable to fill the metadata elements of Dublin Core. To do this will be considered course syllabi and PDF document properties. Results and conclusions obtained in the first part of the experiment will be the basis for the second part of the experiment, where will be considered the total generation of elements of Dublin Core per learning object. Findings of the second part of the experiment will allow us qualitatively assess the potential of the proposed methods.

3.1 The General Purpose of Experiments

The purpose of this experiment is to explore and check whether metadata extracted from the course syllabus as well as the metadata harvested directly from learning object may be used for description of electronic learning object. Also as a result of the experiment we can conclude the extent to which the automatic generation of metadata for learning objects can facilitate the process of manual input of annotations using the proposed methods.

3.2 Technical Aspect

Since within this project we are working with the electronic syllabus published on the Internet then to process them, we need a software tool allowing working with the HTTP protocol and the local and remote files. In our case, for realization of tasks PHP[1] scripting language was used.

In the case of direct processing of learning objects and metadata extraction from fields already filled (i.e. harvesting) Exiftool[2] was used, which is free software used for reading, writing, and manipulating file metadata.

[1] http://www.php.net/

[2] http://www.sno.phy.queensu.ca/ phil/exiftool/

4 The Correspondence between Resources for Generation and DC Elements

In this part of the experiment the correspondence between the fields of a typical course syllabus and the Dublin Core Metadata elements will be examined and set. In other words, the mapping of syllabus fields in the Dublin Core metadata scheme will be conducted. Harvesting aspect of information extraction will also be considered, proposed methods by which information can be retrieved, and discussed its degree of applicability. The proposed methods are tentative and will be tested in the second part of the experiment. The process of description is based on the consideration of each Dublin Core element separately.

4.1 Data Set

60 course syllabi from different schools and universities took place in analysis. We used the popular search engine Google to find schools and collect documents which are the course syllabi. Schools course syllabi of which have been considered are located in various countries such as Norway, Sweden, Denmark, USA, England, Ireland, Iran, and Turkey. The consideration of various schools and universities of various countries is an important factor, because each institution has its own style of syllabus design. Our goal is to determine the common and useful elements among all syllabi. We considered syllabi of such courses as the Semantic Web, Media Management, Internet Technology, Programming etc. (A complete list of schools and courses can be found in the Appendix A). *30 learning objects* in form of PDF lecture slides are the set to identify metadata elements that can be generated automatically after program processing and for identification of harvestable elements. Consideration of learning objects from different courses of different schools will also determine the most frequent and similar features in indication of metadata for learning objects. (List of schools and courses from which learning objects are taken can be found in the Appendix B).

4.2 Element "Description"

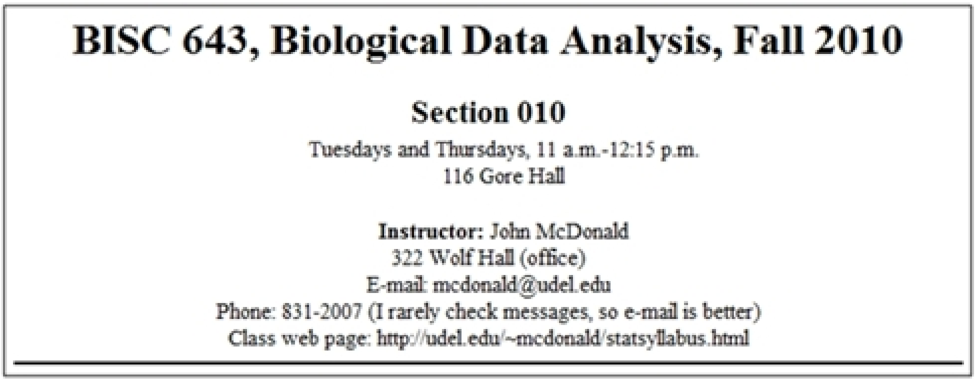
Dublin Core specification allows describing of electronic resource by free text. In the case of learning object in the "Description" element can be recorded the title and course code, which owns the learning object. This allows to determine the place of learning object in the learning process.

Using considered in Section 2.6 parts of typical syllabus we have found that each syllabus may contain the title and code of the course.

4.2.1 Element Consideration

We are considering where and how information on the course title and code is presented in the syllabus.

This element is variously represented in syllabi. In some syllabi the title and code of the course is placed in the page header, in some cases, indicated by sub elements such as "Title", or "Syllabus for". Figure 5 shows the example of page header.



BISC 643, Biological Data Analysis, Fall 2010

Section 010

Tuesdays and Thursdays, 11 a.m.-12:15 p.m.
116 Gore Hall

Instructor: John McDonald
322 Wolf Hall (office)
E-mail: mcdonald@udel.edu
Phone: 831-2007 (I rarely check messages, so e-mail is better)
Class web page: http://udel.edu/~mcdonald/statsyllabus.html

Figure 5: Page header example

Course title and code, from the viewpoint of HTML tags, in most cases is prescribed in <title> tag and can be automatically extracted with PHP script. In the current project for this purpose DOMDocument object was used, which is created to extract information from HTML tags. DOMDocument object converts the HTML document located at the input link in the document tree. And then using the getElementsByTagName() the contents of the required tags is obtained. Figure 6 represents the statistics of the presence of information about title and course code in the HTML tags.

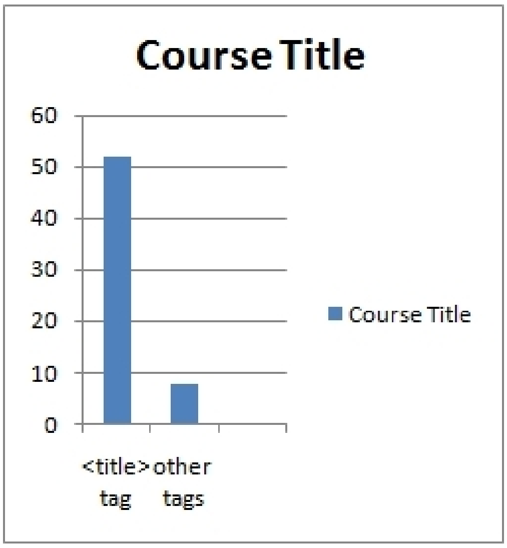


Figure 6: Tags statistics

4.2.2 Results

We relate information about title and code of the course with element “Description” of Dublin Core metadata element set. Extraction of such information from the 60 syllabi of courses yielded the results are presenter on the Figure 7.

The experimental results show that proposed method to retrieve information about course title and code using data extraction from HTML tag <title> gives a positive result in 90% of cases. Failure in 10% of cases is due to the absence of such information in

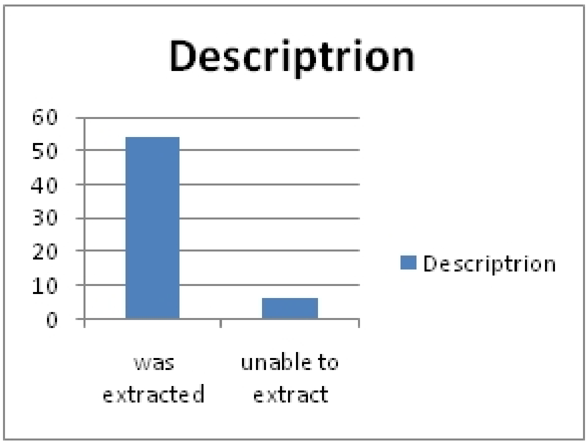


Figure 7: Tags statistics

the tag <title>. Moreover, in some cases the tag <title> contained a side text, such as "Syllabus", "Syllabus for", "Course Syllabus", "Homepage", "Course Materials", punctuation, or date information. To obtain the correct data side text was removed on a phase of preprocessing (Appendix D.1). However, confident trend of presence of information about the title and code of the course exists.

4.2.3 Evaluation

In considering the metadata element of Dublin Core "Description" have been proposed information, by which this element can be filled. Since Dublin Core allows filling this element with the free text as an alternative to use information about the title and code of the course prescribed in the syllabus have been proposed. It was determined that the relevant information contains in the HTML tag <title>. Extraction of information about the title and code of the course from syllabus yielde positive results in 90% of cases. This rate shows the stability and success of this method. However, some challenges have been discovered during the extraction. The presence of a side-text in <title> tag requires the setting of a word list of exceptions, which participates in preprocessing of the text of tag. If the tag <title> contains only the text "Course syllabus", then after the removal of this a side-text element "Description" will remain empty.

Nevertheless, the proposed method leads to correct automatic generation of metadata element "Description". This method ensures objectivity in filling the metadata field "Description".

4.3 Element "Publisher"

According to Dublin Core element "Publisher" "includes a person, an organization, or a service responsible for making the resource available." Thus, in the case of learning object, in the element Publisher can be recorded information about the institution or service providing the course to which the described learning object belongs. Using the previously mentioned components of the syllabus can be concluded that the course syllabus include the name of school, university, or department, which teaches a described course.

4.3.1 Element Consideration

Next will be described in what form and how "Publisher" information can be presented in the course syllabus.

From the viewpoint of analysis of HTML tags, the name of the school has no strict trend and therefore need to find another way to extract this type of information using text processing techniques. Visual detection shows that information about university may be included in page header. During the consideration of course syllabi have been found a list of key phrases that can be used in regular expressions in text processing. Have been discovered the following key phrases: “University of”, “School of”, “College”, “Department of”, and “Institute of”.

Although the use of regular expressions can give a good result, in our case for direct retrieving of the name of the institution, we used the following approach: Since all syllabi used for analysis are published on Internet and have URL (Uniform Resource Locator) we have found a way to extract information about "Publisher" using web address of resource. This approach reduces the cost of resources and reduce time to process information.

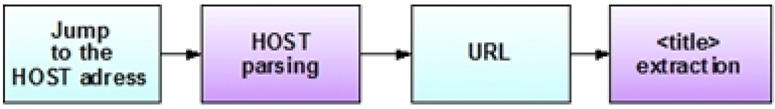


Figure 8: Publisher name extraction method

Figure 8 shows the workflow for retrieving the name of the institution responsible for publishing of electronic documents. The essence of this process is following: take the web address of the course syllabus and determine its host address, i.e. homepage of the institution. Next, the program passes the address and extracts name of the institution with using of DOMDocument object from HTML tag <title>.

In some cases the tag <title> contained a side text, such as "Home page", "Welcome to", or punctuation. Thus, preprocessing should be involved. We used the same techniques like in element "Description" case.

4.3.2 Results

Information about organization responsible for publishing the learning object we relate with the element "Publisher" of Dublin Core metadata element set. The results of extraction of such information from the 60 course syllabi are presented on the Figure 9.

To extract information about the organization responsible for the publication of learning object method involving URL of the syllabus page is used. In 84% of cases the name of the university, college, school or department was determined. Negative result in 12% of cases associated with features of e-learning courses and organizations offering them. Some organizations do not specify their name or specify but in an implicit form, which requires more advanced approach.

4.3.3 Evaluation

In the case of learning object Dublin Core element "Publisher" it have been proposed to fill by name of the organization within which was created and published these learning object. It have been suggested to use the information located in the tag <title> from home page of the organization. During the experiment the proposed method has shown a positive result in 84% of cases. Negative result refers to cases where the course syllabus

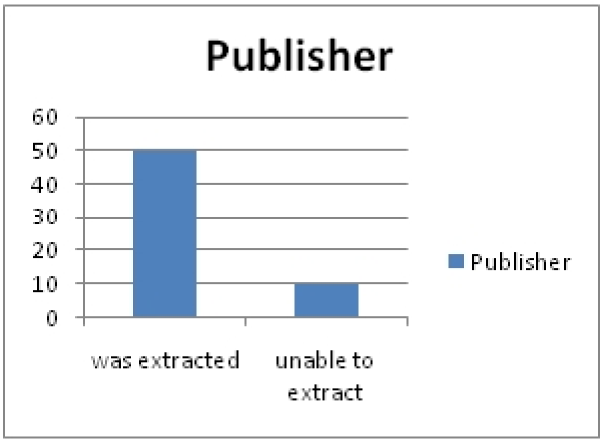


Figure 9: Results of Publisher extraction

is not presented in the domain of educational organizations, but on the home page of the teacher.

Thus we can see that this approach is relevant for those cases where learning objects are published within the institution and not relevant to private courses. Nevertheless, this method of extracting information about the educational institution to fill the metadata element "Publisher" is real and can be used for automatic generation.

4.4 Element "Creator"

According to the Dublin Core "Creator" is "an entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organisation, or a service ". Thus, to fill the fields "Creator" we need to find information relevant to person that created the describing learning object.

4.4.1 Element Consideration

It is possible to assume that the creator of learning object is the head of the course. Previously described structure of the syllabus includes information about the course leader. However, quite difficult to determine the case when learning object is created directly by course instructor, and when the learning objects is created by leader of the guest lectures and seminars or by a group of students. Moreover, we have previously described the "LEGO" metaphor (Section 2.1)regarding learning objects. Under this approach, learning object can be built from combinations of small parts taken as the basis from already existing learning objects created by different authors.

Thus, to accurately extract information about the author of learning object the course syllabus cannot be used. But we can use the harvesting approach to extract information about the author of learning object. In particular the Portable Document Format (PDF) contains a set of key fields for storing metadata. In the case of disputed authorship of the document key field "Author" will be used [61].

4.4.2 Results

Extraction of information about the author of 30 learning objects in our data set have given the results shown in Figure 10.

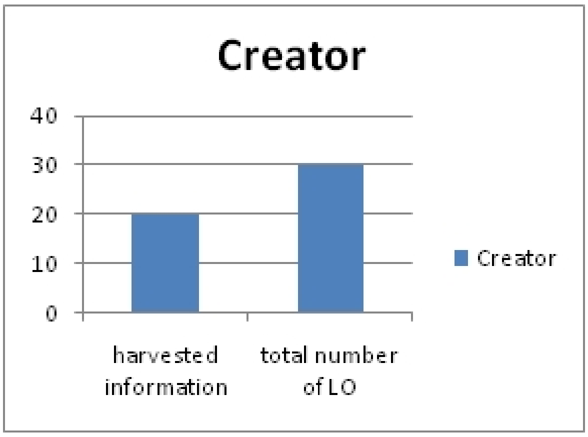


Figure 10: Harvesting of “Creator” information

4.4.3 Evaluation

Previously, it was considered that the course syllabus is not a source of correct information about the creator of learning object. Therefore, the basis for generating the metadata element "Creator" is proposed harvesting approach. The results of harvesting has shown the presence of information about the author of learning object in 67% of cases, but the reliability and correctness of this information is difficult to judge as it is impossible to determine whether the information is provided manually or assigned by the system automatically. Moreover, in the case of joint authorship may present the subjective opinion of a person. In order to eliminate errors in the generation of information about authorship we propose to apply manual input. Wrong indication of the author may raise legal issues such as uncertainty in the allocation of rights and conditions of distribution of learning objects.

4.5 Element "Contributor"

According to Dublin Core “Contributor” is “an entity responsible for making contributions to the content of the resource. Examples of a Contributor include a person, an organization or a service.” The main difference between fields contributor and creator is that the creator can publish, edit, delete his work, but all changes of contributor must be approved by creator. Role of contributor in lifecycle of learning object could be as follows: changes in the content of learning object, filling in the fields of metadata to describe the learning object, and publication of learning objects for public use and reuse. It is possible to assume that these responsibilities can be performed by head of the course, at least partially. Information about the course leader quite common for the structure of a typical syllabus and can be used to fill the field "Contributor" of Dublin Core.

4.5.1 Element Consideration

During the consideration of syllabus content it was determined that information about course leader may be indicated by the concepts: "Instructor", "Professor", or "Doctor". Later these concepts will be used as keywords for regular expressions that will enable us to extract information about leader of the course from the free text of syllabus.

Statistics shows that 88% of the considered syllabi contain a field indicating the in-

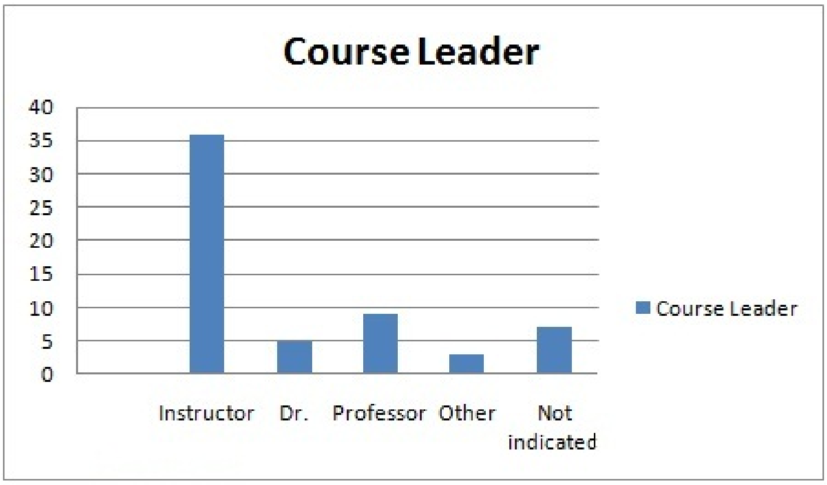


Figure 11: Course leader information statistics

formation about course leader and 60% of them are often call this field as Instructor.

In addition to the name of course leader syllabus stated his personal contact information. Necessary part of the contact information is e-mail of the person. From the standpoint of presenting this information in HTML format e-mail address is written using HTML tags <A> and attribute "mailto:". Extraction of this information will allow us to link learning object with the head of the course.

Some schools also provide the teacher assistants; this information may also be useful when filling the attribute "Contributor". Information about the teaching assistant is denoted by the concepts: “Assistants” or “Teacher Assistants”. However during syllabi analysis Course Assistant met less often (78% of the analyzed syllabi do not refer or do not have a field Course Assistant). Therefore this information was excluded from further consideration.

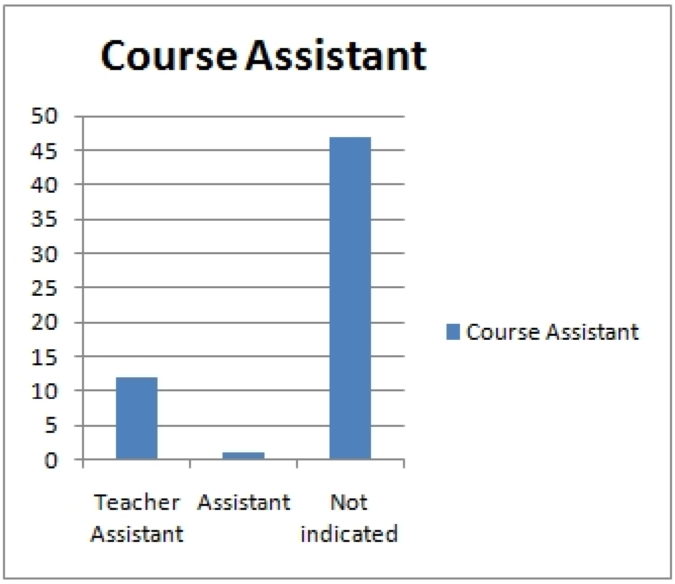


Figure 12: Course assistant information statistics

Let’s summarize the methods to be used in the ongoing project to extract information from the course syllabus for element "Contributor" of Dublin Core:

- By using specific keywords "Instructor", "Doctor", and "Professor", with using regular expressions will be determined name and title of course head (Appendix D.2).
- If using regular expressions do not give a positive result then attribute "mailto:" of HTML tag <A> will be used as an identifier for the head of the course (Appendix D.3).

4.5.2 Results

Extraction of Course leader information from 60 course syllabi yielded the results represented on Figure 13

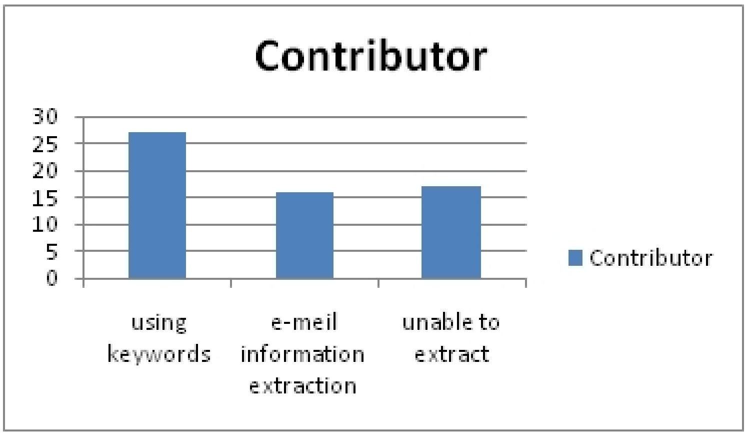


Figure 13: Results of ”Contributor” extraction

Previously, we identified the list of keywords ("Instructor", "Doctor", and "Professor") that are relevant for use in extraction of information on the course leader using regular expressions. It is also proposed to extract identifier of course leader by using his/her email address. The experimental results show that in 84% of cases it is possible to extract this information using proposed method. Failure to extract such information in 16% of cases was due to lack of data in the course syllabus or ambiguous complicated way of displaying this information on the HTML page of syllabus.

4.5.3 Evaluation

Was considered an element of Dublin Core “Contributor” and it was found that extraction of information about the head of the course that relevant to filling this element can be conducted by the proposed methods. We have identified a list of specifications keywords to designate the head of the course, which are the most common for the majority the existing syllabi. It was also proposed to use the e-mail of course leader as an identifier of his or her person. Individually, these two methods do not always give certain results due to the lack or difficulty of its extraction. However, the combination of these techniques can improve accuracy and provide a more stable result.

In terms of automating the generation of metadata element "Contributor" we may conclude that the use of course syllabus as a source of information for the generation as well as application of the proposed methods, and its results are objective and fair. Accordingly, the manual entry and human control are not required and element can be generated automatically.

4.6 Element "Language"

Element "Language" of Dublin Core is defined as "language of electronic resource". In the case of learning object language of instruction will allow the student as a user of an electronic resource to better define how the resource is suitable for him to solve his learning tasks. Also, the authors of new learning objects will be able to determine how pre-existing educational information will be suitable for context of new learning objects and prepared lecture.

4.6.1 Element Consideration

Consideration of the syllabus of courses has shown that some schools and universities provide education on several languages. In this case they indicate what language is used for course. Mostly this applies to non-english-speaking countries and schools working on international programs, such as Norway, Finland, and Denmark.

4.6.2 Results

Visual cobsideration the language of the course in the structure of syllabus shown that the language of instructions can be specified by using such concepts as "Language" or "Language of instructions". The Figure 14 shows results on the visual consideration of the course language.

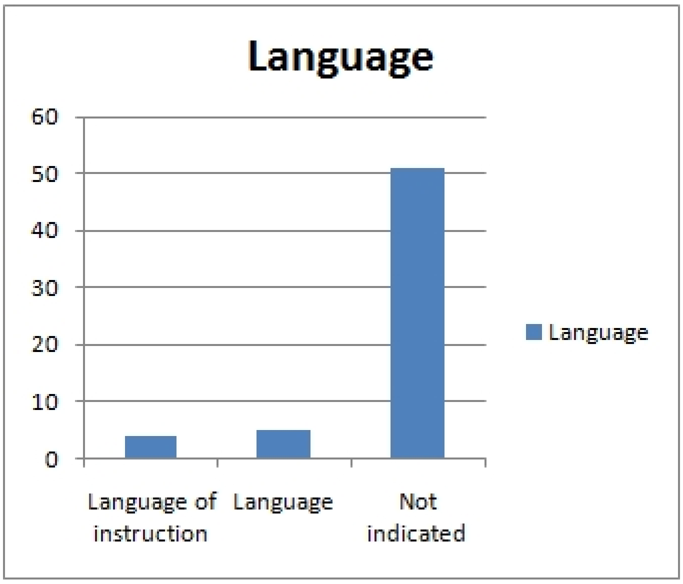


Figure 14: Language information statistics

4.6.3 Evaluation

Visual consideration of the element to represent the language of the course from syllabus among the data set shows that a small number of schools clearly indicates language of course instructions. It would be desirable to note that the english-speaking countries do not indicate the language of instruction of the course in syllabus at all. To compile our data set were mostly used course syllabus and learning objects belonging to them from the english-speaking countries. This can cause low level of indications of language instruction in the structure of the course syllabus. Thus, it is necessary to find techniques for language identification. To do this is required to apply Language Identification from the text of the learning object. The most popular approaches have been presented previously in chapter Related Work.

Existing approaches can accurately determine the language of the text of an electronic document. Accordingly, we can conclude that the automatic generation of the element Language can be done automatically.

4.7 Elements "Coverage", "Title", and "Relation"

It is necessary to briefly mention why we consider elements "Coverage", "Title", and "Relation" as conjunction. It was found that in the course syllabus more often, these three elements form a single structure and the extraction of one element of the structure depends on the remaining elements. More detailed definition of the elements of this triple and review of the structure is presented below.

Element "Coverage" of Dublin Core includes an indication of spatial or temporal information. In the case of learning object temporal information has the greatest importance. Dublin Core defines temporal coverage as "temporal period (a period label, date, or date range)". Using the temporal information will be helpful to take into account the chronology of the publication of learning objects, which will allow both students as users of resource together with the head of the course to relate a specific learning object with lecture timeframe in which this learning object was used. However, another alternative, which is not considered in this project to fill the element "Coverage" may be an indication of the historical period, to which the content of learning object corresponds (e.g. a historical event or memorable date).

Element "Title" in Dublin Core is specified as "the name given to the resource. Typically, a Title will be a name by which the resource is formally known". We propose that the name of the learning object can match with the topic under consideration at the lecture. The structure of a typical syllabus in addition to the date of lecture involves the topic of the lecture.

"Relation" information in Dublin Core specified as "a reference to a related resource". The structure of the syllabus involves an indication of "supplementary reading" materials. Thus, we propose that the reference information about links to electronic resources, research articles, other textbooks, or related learning objects for particular learning object is useful for students. Using such information, students can explore more deeply into topic by reading related materials.

The typical structure of the course syllabus involves indication of lectures schedule. Thus, it is necessary to determine how and in what form the temporal information, lecture titles, and related materials may be present in the schedule area of syllabus and the methods by which this information can be extracted.

4.7.1 Elements Consideration

After the consideration of HTML code of syllabi pages from our data set, we have identified some patterns in representation of temporal information, lecture title, and related materials:

One pattern that have been discovered during the consideration of 60 syllabi represented in form of HTML, is 3 formats of writing of temporal information: "*WEEK #*", "*MONTH*", or date format "*dd/mm/yy*". Also applied reduction such as "Aug. 27". Later these concepts will be used as keywords for regular expressions that will enable us to extract information about chronology of publications from the free text of syllabus (Appendix D.5, Appendix D.6).

As mentioned earlier, temporal information, lecture title, and related materials in most cases form a single structure of representation. The structure can be represented as list, table, or sequence of paragraphs. Figure 15 shows “Temporal information, lecture title, and related materials” representation statistics.

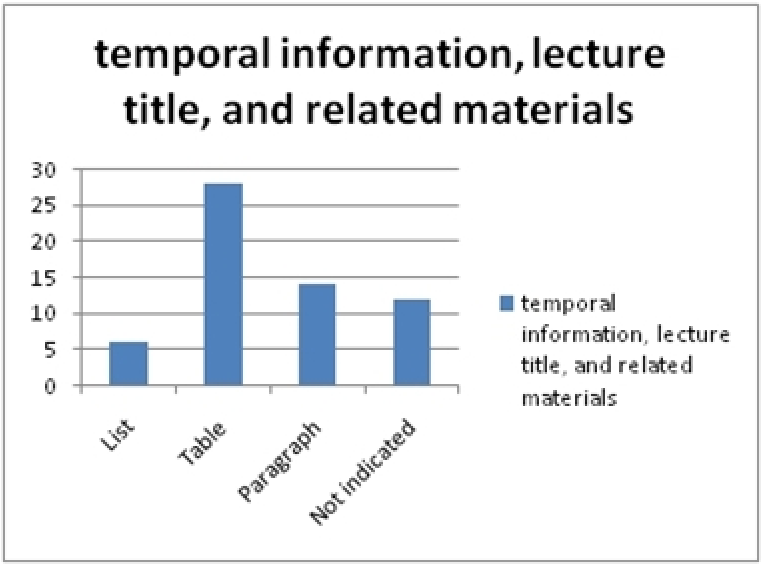
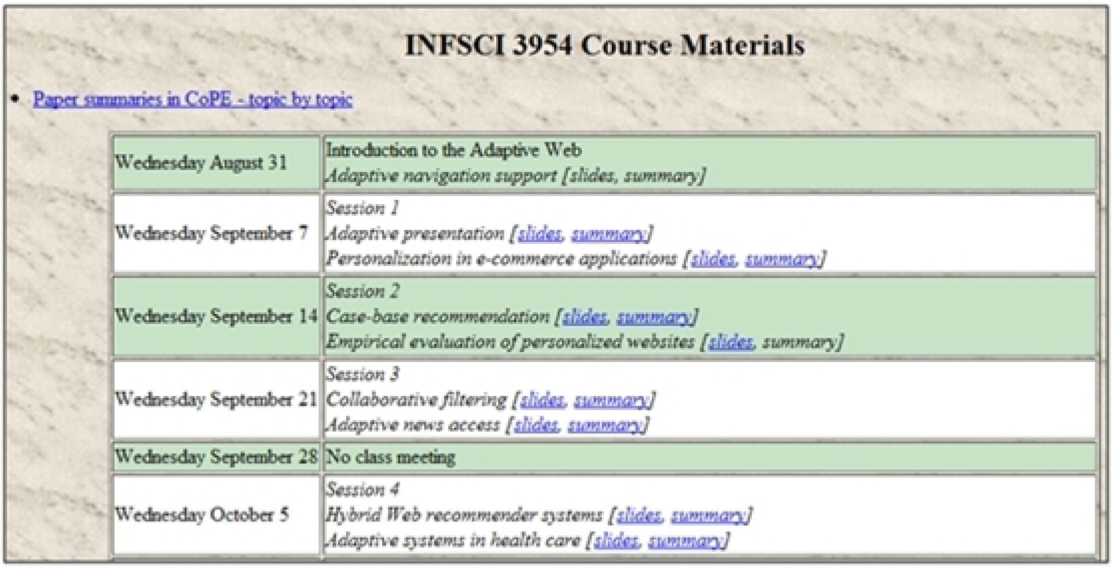


Figure 15: “Temporal information, lecture title, and related materials” representation statistics

During the consideration of course syllabi we found that the tabular presentation of the structure "temporal information, lecture title, and related materials" may have a different format. From the standpoint of analyzing of HTML tags, this structure is often presented in the following ways (Figure 20, 21): A table is divided into rows (with the <tr> tag), and each row is divided into data cells (with the <td> tag). <td> stands for "table data," and holds the content of a data cell [62]. Extraction of information from structure is carried out using DOMDocument object and processing using regular expressions (Appendix D.4).

In the case of the representation of structure "temporal information, lecture title, and related materials" in the list form the following patterns in HTML tags tree have been met (Figure 22, 23): The <ul> tag defines an unordered list and the <ol> tag is used to create an ordered list. The <li> tag defines a list item [62]. Extraction of information from structure is carried out also using DOMDocument object and processing using regular expressions.

Presentation of "temporal information, lecture title, and related materials" in the form of free text causes the greatest difficulty because of its unstructured form. To extract



INFSCI 3954 Course Materials

- Paper summaries in CoPE - topic by topic

Wednesday August 31	Introduction to the Adaptive Web *Adaptive navigation support [slides, summary]*
Wednesday September 7	*Session 1* *Adaptive presentation [slides, summary]* *Personalization in e-commerce applications [slides, summary]*
Wednesday September 14	*Session 2* *Case-base recommendation [slides, summary]* *Empirical evaluation of personalized websites [slides, summary]*
Wednesday September 21	*Session 3* *Collaborative filtering [slides, summary]* *Adaptive news access [slides, summary]*
Wednesday September 28	No class meeting
Wednesday October 5	*Session 4* *Hybrid Web recommender systems [slides, summary]* *Adaptive systems in health care [slides, summary]*

Figure 16: Table structure example

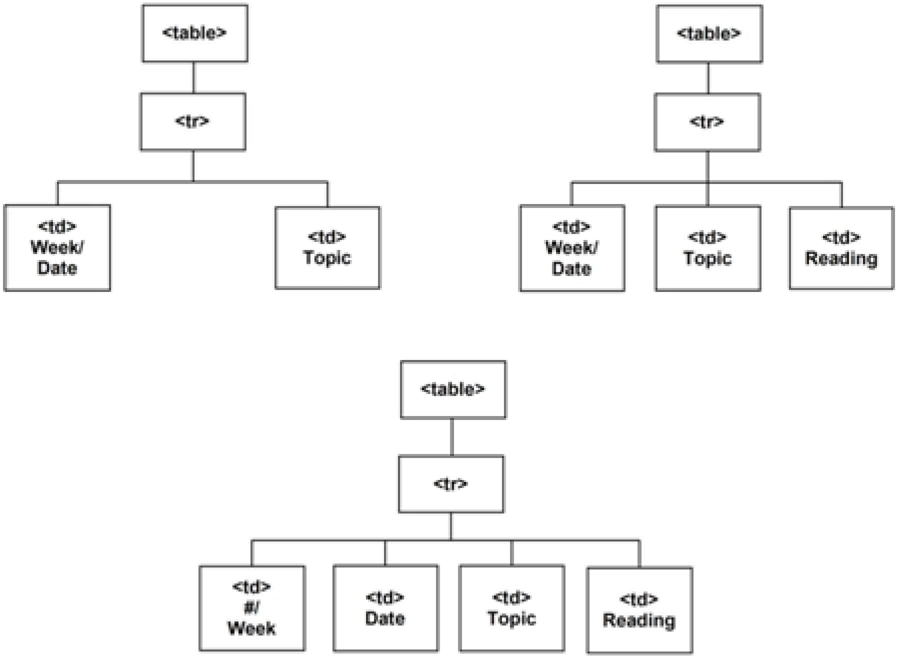
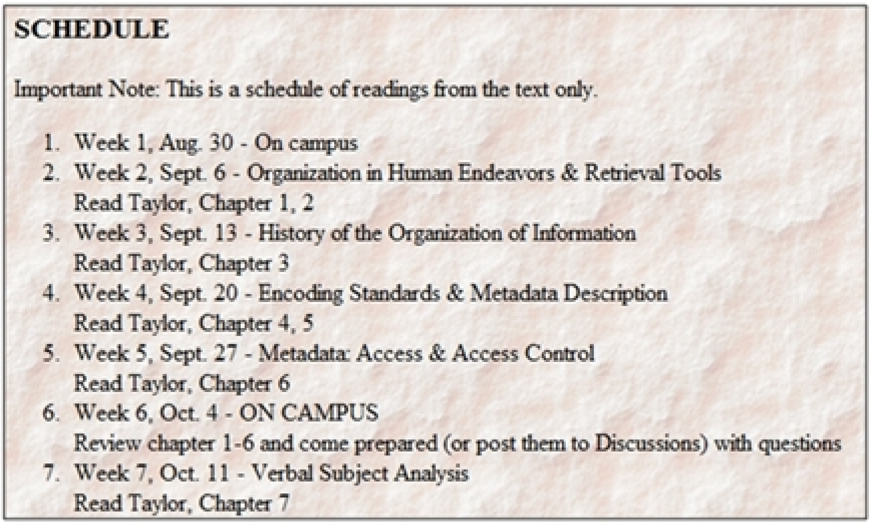


Figure 17: Table tags structure examples



SCHEDULE

Important Note: This is a schedule of readings from the text only.

1. Week 1, Aug. 30 - On campus
2. Week 2, Sept. 6 - Organization in Human Endeavors & Retrieval Tools
   Read Taylor, Chapter 1, 2
3. Week 3, Sept. 13 - History of the Organization of Information
   Read Taylor, Chapter 3
4. Week 4, Sept. 20 - Encoding Standards & Metadata Description
   Read Taylor, Chapter 4, 5
5. Week 5, Sept. 27 - Metadata: Access & Access Control
   Read Taylor, Chapter 6
6. Week 6, Oct. 4 - ON CAMPUS
   Review chapter 1-6 and come prepared (or post them to Discussions) with questions
7. Week 7, Oct. 11 - Verbal Subject Analysis
   Read Taylor, Chapter 7

Figure 18: List structure example

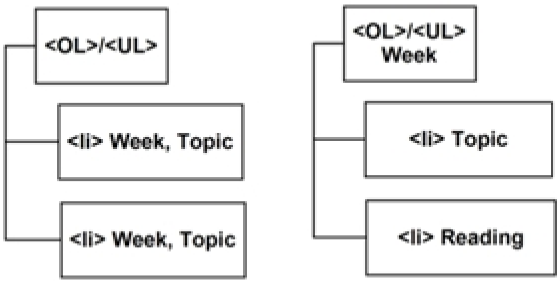


Figure 19: List tags structure examples

Syllabus	Extracted Data
1	Week Topic
2	Date Topic
3	Date Topic
4	Date Topic
5	Week Topic Reading
6	Date Topic Reading
7	Week Topic Reading
8	# Date Topic Reading
9	Week Date Topic Reading
10	Date Day Topic Reading
11	Week Date Topic Reading
12	Week Date Topic Reading
13	Date Day Topic Reading
14	Date Topic Reading
15	# Date Topic Reading
16	# Date Topic Reading
17	Date Topic Reading
18	Week Date Topic Reading

Table 2: Information extracted from the table view of the course schedule

information from free text paragraphs regular expressions are used.

On the other hand it is desirable to note that the Portable Document Format (PDF) contains a key field for recording information about the title of the document. Using the harvesting method, this information can be automatically extracted from the file properties of the document.

4.7.2 Results

The structure "temporal information, lecture title, and related materials" consistent with elements «Coverage», «Title», and «Relation» of Dublin Core metadata standard, respectively. As described earlier, this structure in the course syllabus may be presented in three forms: table, list, or sequence of paragraphs. Table 2 displays the data that was extracted from the table view of the course schedule.

The results of extracting data from the table view schedule of the course show that in all cases it is possible to extract temporal information and topic of the lecture.

In the case of list format representation extraction of structure "temporal information, lecture title, and related materials" is failed. This result is due to the complicated presentation of the structure in the list format.

Only 3 syllabi of 14 potentially suitable for the processing allowes extracting the entire structure "temporal information, lecture title, and related materials" from free text paragraph. The results of all 14 syllabi are presented on Table 3.

However, the use of regular expressions to extract data about temporal information, lecture title, and related materials from free text paragraph gives a good result for the extraction of each element of structure separately. In particular, information about the date of the class have been obtained in 92% of cases, information about lecture topic in 78% of cases. The greatest difficulty causes obtaining information about related materials.

The result of extraction of the structural elements from the 60 examined syllabi can be summarized in Table 4.

Syllabus	Extracted Data
1	Date Topic
2	Date
3	Date Topic Reading
4	Date Reading
5	Date Topic
6	Date Topic Reading
7	Topic
8	Date Topic Reading
9	Date Topic
10	Date Topic
11	Date Topic
12	Date Reading
13	Date Topic
14	Date Topic

Table 3: Information extracted from the free text paragraphs of the course schedule

	Temporal information	Topic	Reading
Table	18	18	14
List	0	0	0
Paragraph	13	11	5
	31	29	19
	52%	**48%**	**32%**

Table 4: Summed results

The use of harvesting methods to extract information about the title of the document from the 30 PDF lecture notes gave the results shown in the Figure 20. Information prescribed in the title property of the PDF document have been found in 20 analyzed learning objects. In some cases, the title information have been listed incorrectly or have not corresponded to the actual topic of learning object. For example, have been met such values as "Ingen lysbildetittel" ("no slide title").

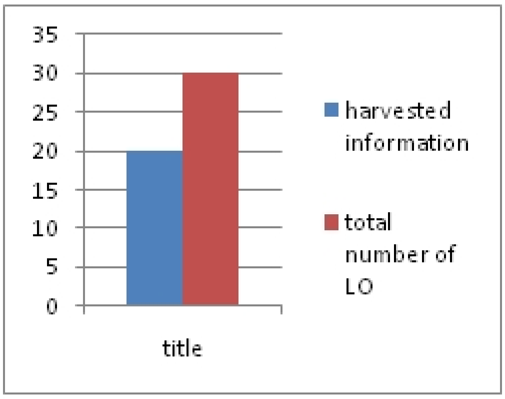


Figure 20: Harvesting results of field Title

4.7.3 Evaluation

Considering the course syllabus, we concluded that the information for filling the "Coverage", "Title", and "Relation" Dublin Core elements, in most cases is represented as a single structure. We determined common formats for these structures. Considering the results of extraction of relevant information for these elements, we can conclude that the elements "Coverage" and "Title" have an average result. For element "Relation" result was quite low.

Some difficulties to correctly extract the data have been met. Firstly, the main problem that preventes the extraction of data is nonuniformity of HTML pages in sense that one page can include more than one table. Moreover, the tables themselves do not have uniform structure. For example, the table may contain a different number of columns or the table has a complicated structure. Also, the extraction of information about related resources for specific lecture shows some potential. However, consideration of the structure "temporal information, lecture title, and related materials" as a single entity does not always give a positive result. The absence of one element of the structure may effects negatively on the extraction results.

Also some difficulties have been met during processing of information in the list form of representation. In particular, the format of the lists has nonuniform presentation. In some cases, the list can be represented as a sequence of elements (1 / 2 / 3), in some cases, the list can be represented as a sequence of elements, together with sub elements (1 / 1.1 / 1.2 / 2 / 2.1 / 2.2 / 3 / 4).

To extract the structure "temporal information, lecture title, and related materials" from the sequence of paragraphs has been applied text processing. Our experiment showes that the extraction of required information from free text paragraph does not have confident results.

To some extent it is possible to use the information provided in the syllabus for the automatic generation of elements "Coverage", "Title", and "Relation" of Dublin Core. However, since this information is often presented in a large volume without linking to a specific learning object, it requires a manual indication of one from the elements on which the system can generate and suggest the remaining elements. For example, a user can manually select the corresponding week of the chronology of learning object. Based on input system will extract the relevant information from syllabus about the topic of learning object and related reading materials.

On the other hand harvesting approach for generation of Dublin Core element "Title" for learning object have been considered. It was observed that the information recorded in the file properties does not always correspond to reality, and correctly displays the contents of the resource. In this case, the information is not actual. Since the system can not automatically check the correctness of the information harvested, it is impossible to use this method in an automatic mode.

4.8 Element "Date"

Element “Date” in Dublin Core is defined as "a point or period of time associated with an event in the lifecycle of the resource".

4.8.1 Element Consideration

According to the description of PDF format date information is described as "The date the document was created or most recently modified in human-readable form". Moreover, depending on the version of the system information about last modification is mandatory. The system can automatically fill field "Date".

Syllabus of the course is not intended to indicate information about the date of last modification of learning object. Therefore, for extraction of information about the date of last modification of the electronic document harvesting approach have been used, which gives 100% of correct result. Thus, the filling of element "Date" of Dublin Core element set can be fully automated.

4.9 Element "Format"

“Format” element in Dublin Core is defined as "a file format, physical medium, or dimensions of the resource" and as well as the date is often stated in the file properties.

4.9.1 Element Consideration

Essentially format is a file structure that determines how to store it and display on screen or in print. The file format is usually indicated in its name as part separated by point [63]. Moreover, for the description of the file format in the implicit form is a widely used method that is common for UNIX-like operating systems. The sense of this method is to save in the file itself a kind of "magic number" (signature) - the character sequence by which can be identified the file format [64]. In some cases, during the publication in the course syllabus for the learning object is indicated the file format. For example "PDF/PPT". However, the use of harvesting approach for extracting information about the format of the electronic document does not require analytics and will be spent fewer computational resources. Moreover harvesting approach gives 100% correct results in extracting information stored by the system during file saving. Thus, the generation of Dublin Core element "Format" can be completely automated.

4.10 Element "Type"

Element “Type” in Dublin Core is defined as "the nature or genre of the content of the resource”.

4.10.1 Element Consideration

The concept of learning object does not imply the use of any one type of resource. Wiley in [3] defines a learning object as "digital resource that can be reused to support learning be it large or small". Thus, the type of learning object can be images, live or prerecorded audio or video, text files, and small web applications.

The DCMI Type Vocabulary provides a list of terms to specify the value of the genre of electronic resource [65]. According to this document with the concept of learning object the following terms can be assigned:

- Dataset - that includes lists, tables, and databases;
- Image - that is a visual representation of images and photographs of physical objects, paintings, prints, drawings, other images and graphics, animations and moving pictures, film, diagrams, maps, musical notation.

- Interactive Resource - which requires direct user interaction for a more detailed study (Web pages, applets).
- Sound – that is a resource primarily intended to be heard including music and recorded speech or sounds.
- Text – that is a resource consisting primarily of words for reading.

However, Dublin Core does not limit the indication the type of document by terms presented above and allows the use of free text. In this project we restrict ourselves by consideration of one of the most popular types of learning object - lecture notes in a slide presentation. Lecture notes can be represented as a combination of different types of objects. Therefore, it is quite difficult to determine the components of the learning object. Thus, it is logically to offer the user specify the type of learning object manually.

4.11 Element "Identifier"

Element “Identifier” in Dublin Core is defined as “an unambiguous reference to the resource within a given context. Example of formal identification systems include the Uniform Resource Identifier (URI) including the Uniform Resource Locator (URL)”.

4.11.1 Element Consideration

If to refer this description to learning object, then the URL of published resource will be generated only after its placement in a physical store. The process of assigning of URL for electronic resource is completely automated and performs directly by Learning Object Management System. Specifying the physical address of the system gives 100% of unique resource identification.

4.12 Element "Source"

Element “Source” in Dublin Core is defined as “a Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or part”.

4.12.1 Element Consideration

In the case of learning object is quite easy to confuse Sourse and Relation information. It is known that most teachers base their lecture on some relevant textbooks. During planning of the classes, in some cases, the head of the course relates the topic of the lecture with chapters or pages of basic literature. The structure of a typical syllabus is intended to include information about literature, on which a course based. Thus, we need to determine in what form such information is represented in the structure of the syllabus and what method can be used for its extraction.

In syllabus Basic literature is marked with the following concepts: Textbook (s), Required text, Reading, Required materials, Reading list, Course book, Course materials, References, Reference material, Teaching materials, etc. In some cases, the list of basic literature indicates International Standard Book Number (ISBN), which serves as an identifier of the book.

Thus, having keywords "Textbook", "Text", "Book", "ISBN", and "Reading" will be extracted information about course basic literature with using regular expressions (Appendix D.7).

4.12.2 Results

Review of the syllabus from our data set shows that 86% from 90 analyzed syllabi contains the basic course materials information.

The process of information extraction about the basic literature from 60 syllabi of the course presented in HTML form gives the results shown in the Figure 21.

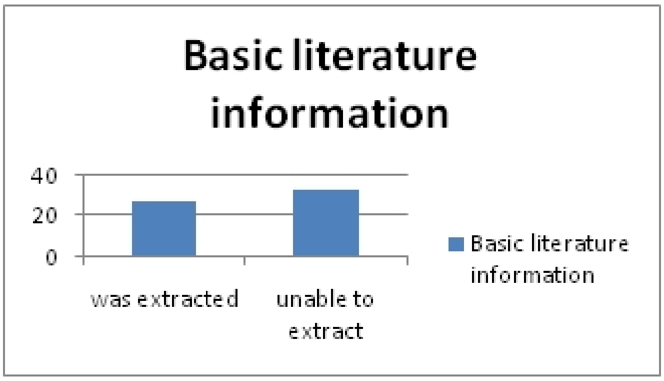


Figure 21: Extraction of basic literature information results

Failure in 55% of cases related with providing information about basic textbooks with complicated structures or no indication.

4.12.3 Evaluation

Considering the syllabus of the course has been found information about the basic textbooks in 86% of cases. But because of the unstructured and non-standardized format of the syllabus, and also because of the lack of this information, automatic generation of Dublin Core element "Source" has a fairly low rate of 45%. Also the problem of accurately determining the source for a specific learning object have been met. For example, not all of indicated in the syllabus sources can be the basis for particular learning object. In this case, specifying of all sources would be redundant. However, the rate of 45% shows that there is a potential to extract information about the sources. Information extracted from syllabus may be offered to the user for correction or confirmation.

4.13 Element "Rights"

In Dublin Core element “Rights” is described as “information about rights held in and over the resource. Typically a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights”.

4.13.1 Discussion

Using a variety of "copyleft" license is a flexible way of specifying the rights to use intellectual property. In most cases, this process requires manually specifying the conditions under which the learning object is available publicly. Only Creative Commons imply encoding information about licenses in machine readable form. However, during the work we have met only one syllabus defines the Creative Commons license. This issue was excluded from further consideration due to the rare use. The use of copyleft licenses may be subject to further work.

Previously described structure of a typical syllabus is not intended to indicate rights for ownership and distribution of learning objects. In related works (Section 2.10) we have concluded that in the case of learning object in copyrights information an identificator of a direct author or group of authors of learning object can be written. On the other hand if the learning object is created as part of employment, then the authorship belongs to the employing organization. In this case the copyrights information may contain information about the organization. However, automatic extraction of information about the author of object from course syllabus or from properties of the electronic document in order to avoid errors requires control from a human side. Also found that information about the publishing organization has a good potential for automatic extraction from course syllabus. Thus, indication of the owner rights to the use and distribution of learning object can be realized automatically with an indication of the responsible organization.

Must be said that currently there is no definitive mechanism for copyright management. Protecting the rights of authorship in e-learning is a sensitive issue especially in cases of commercial and monetisation use of resources.

4.14 Element "Subject"

In Dublin Core element “Subjects” is defined as “the topic of the content of the resource. Typically, a Subject will be expressed as keywords or key phrases or classification codes that describe the topic of the resource”. Dublin Core practice recommends the "use unique words for keywords". Thus there is the need to find a technique to detect the most relevant keywords for display the essence of the content of electronic documents.

4.14.1 Discussion

During consideration of syllabus of courses has not been met a source for the extraction of keywords related to a specific learning object. The use of harvesting approach allows to extract keywords predefined to e-learning object on the phase of creation. PDF contains a set of key fields for storing metadata. In the case of specifying keywords for the document this field will be "Keywords". However, this process involves human participation and as a consequence the subjective opinions of the person.

On the other hand Section2.6.5 of this project describes the most popular basic approaches for extracting the keywords from the text of electronic document. Described methods include preprocessing of the text document to remove the noise words, and further mathematical processing. The method for comparing words from the text electronic document with WordNet dictionary have been also described. These approaches have been proved and can achieve high accuracy in determining the keywords from the text of document that allows to fully automate these processes. Improving the accuracy of keywords extraction from the text of document can be viewed as a future work for this project.

5 Methods Testing

In this part of experiment we test the application of methods defined in the first part of the experiment. The meaning of the experiment is to check how many metadata records of Dublin Core can be automatically filled by using proposed methods per learning object. After the experiment based on the results we will be able to discuss the quality of generated information and discuss the applicability of the methods for automatic generation of metadata for learning objects.

5.1 Data Set

In this part of the experiment as data set have been used *10 learning objects* belonging to different courses in 10 different schools. Chosen course syllabi have a structure and content that is most similar to the typical syllabus specified in Section 2.6.1. We have set these limitations because of the timeframe of the project. Links to learning objects have been attached inside the body of syllabus. It is important to emphasize that the study is not bound to a single course or school.

5.2 Limitations

Limitation to have only 10 learning objects is due to fact that during the compiling of our data set, we met some difficulties to find a freely published syllabus, which contain direct links to learning objects in its body and access to these learning objects is open.

As determined in the previous part of the experiment, a small number of course syllabi previously considered inherent the presence of indications of language of course instructions. But existing approaches allow detecting the language of document with high accuracy. In this regard, we are not considering the language detection in our study. Also previously found that harvesting approach involves the use of human generated information while extracting the key words in the document. Our goal is to eliminate the subjective influence of the person on the metadata generation. Previously have been described approaches to automate the process of extracting the keywords from the text of document. We do not consider the automatic generation of the element "Subject" of Dublin Core.

5.3 Prototype

To obtain results on how much metadata records of Dublin Core can be automatically filled per learning object, have been applied rules and methods set out in the first part of the experiment. Table 5 shows used methods. In this table A - fully automatic, SA - semiautomatic, M - manul.

For the experiment has been implemented software prototype combining all of established rules, and which allows to simulate the publication of learning object and the automatic generation of metadata. Since for the automatic generation of metadata is used the course syllabus and learning object, then the prototype should allow obtaining of learning object and syllabus as input, as well as an analysis of loaded information.

Dublin Core Element	**Resource for Generation**	**Extent of Automatisation**
Description	Extraction from HTML tag	A
Publisher	Extraction from HTML tag	A
Creator	-	M
Contributor	Text processing using defined keywords and extraction from HTML tag	A
Coverage	Extraction from HTML tag	SA
Title	Extraction from HTML tag	SA
Relation	Extraction from HTML tag	SA
Date	Harvesting	A
Type	-	M
Identifier	Using physical location of LO	A
Source	Text processing using defined keywords	SA
Rights	Extraction from HTML tag	SA
Format	Harvesting	A

Table 5: Applied methods

As we have explained in the first part of the experiment, some metadata elements may require manual input or control and correction from the person. Figure 22 shows the concept of generation process.

In the first part of the experiment, we have found that these elements metadata Dublin Core as Description, Publisher, Contributor, Date, Identifier, and Format can be generated in an automatic manner and do not require manual input and control. In this part of the experiment we will verify this assumption.

Elements Coverage, Title, Relation require the control of the person and the choice of one of the elements of the triple require confirmation from the person on the basis of the provided information. The process of filling the fields of metadata can be implemented through the following steps:

- Step1 : The system allocates the structure in the course syllabus, which contains information corresponding to the structure “temporal information, lecture title, and related materials”.
- Step 2: The system offers the user to select the extracted data about temporal information.
- Step 3: The user selects the temporal information value.
- Step 4: Based on the selected data the system finds the relevant information on the lecture title and related materials and offers to user.
- Step 5: User accepts offered information or corrects it.

In the case of the element Source information can be extracted from the syllabus, but as shows the results of the first part of the experiment, it is difficult to identify automatically with high accuracy the source to build a learning object. Thus, extracted information about the basic course literature may be offered to user for confirmation or correction.

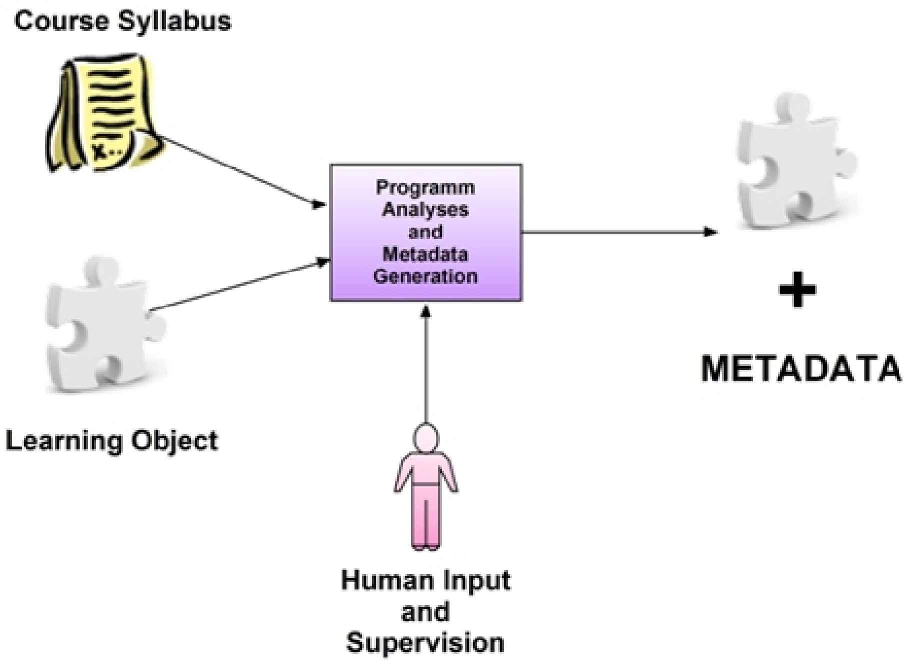
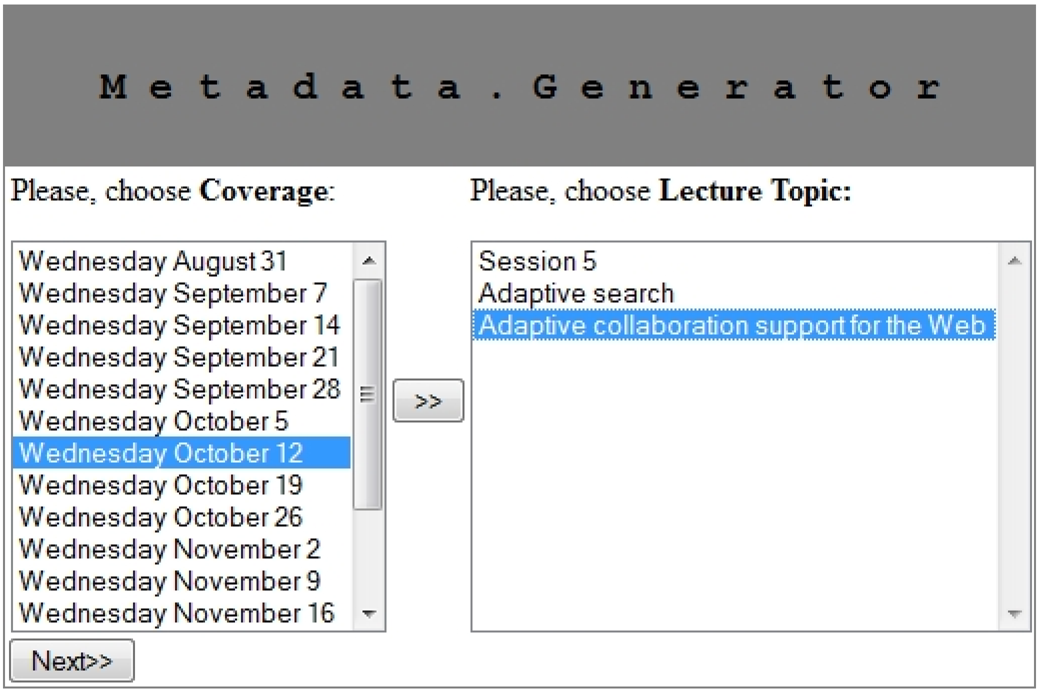

Course Syllabus
Programm Analyses and Metadata Generation
Learning Object
+
METADATA
Human Input and Supervision

Figure 22: The concept of generation process



Please, choose **Coverage**:

Wednesday August 31
Wednesday September 7
Wednesday September 14
Wednesday September 21
Wednesday September 28
Wednesday October 5
Wednesday October 12
Wednesday October 19
Wednesday October 26
Wednesday November 2
Wednesday November 9
Wednesday November 16

Please, choose **Lecture Topic:**

Session 5
Adaptive search
Adaptive collaboration support for the Web

Figure 23: User interface example

Regarding information to fill the Rights element have been proposed to use the data about course instructor or organization responsible for the publication of learning object. Uncertainty with this element of metadata can be solved by offering information to the user for confirmation or correction.

Remaining fields of metadata Creator and Type require manual input. Our prototype provides an opportunity to fill these fields.

5.4 Results

In the current part of the experiment, we have tested whether the previously proposed methods will be weighty for the generation of Dublin Core metadata elements for learning objects. During the experiment, have been obtained results showing the number of metadata records that can be generated automatically per learning object. Table 9 from appendices displays the results. Further with these results, we can conclude the quality of value of generated metadata records and the applicability of the proposed methods and sources for the automatic generation of metadata for learning objects.

5.5 Metadata Quality Evaluation

In chapter Related Work (Section 2.8) were considered metrics assess the quality of automatically generated metadata. We apply these metrics to evaluate the quality of obtained results. Assessment of quality will allow making conclusion about how well the previously proposed methods work. It also will help set the nuances that affect on one or other quality parameter.

Completeness estimates the degree of fullness of all elements from metadata scheme. Thus will be applied binary value indicating whether the scheme is "complete" or "not complete".

The proposed methods for the automatic extraction of information from the syllabus and the file properties confirm their applicability as shown good results with respect to elements Description, Publisher, Contributor, Date, Identifier, and Format. Elements requiring control of the user Coverage and Title also demonstrate certain potential for extraction. In rare cases, the generation of items Relation and Source was succeeding. This is due to the fact that all the additional information about relater resources in the structure of the syllabus is presented in a complex form and is not presented in a unified form, or may often be absent. As information to fill the Rights element we propose to use information about a service or organization responsible for the publication of learning object. There is good potential to fill Rights element by this information, but control of the user must be present. However, the results show that the entire set of these metadata elements in this part of the experiment failed to generate. On average, for each learning object cannot be generated three metadata records. We were unable to gain access to certain learning objects due to need to undergo the procedure of authorization. This also affected the results. Thus we can conclude that the proposed methods for metadata generation for learning objects did not give in result a complete metadata schema.

Correctness reflects the degree of correctness of metadata fields values used to describe the learning object. To assess the correctness of obtained metadata records, we introduce a scale, which includes three values. "Not correct values" involve the empty field values and the values do not correspond to the desired value. "Requires manual correction" include information that contains redundant or insufficient information. "Acceptably correct" includes information that contains minimum of information to fill the field. For example for these values, we include "INSL 582_001" indicating only the course code, or only the week number value "WEEK 4" for the temporal coverage.

Figure 24 shows the results of correctness criteria evaluation of generated values.

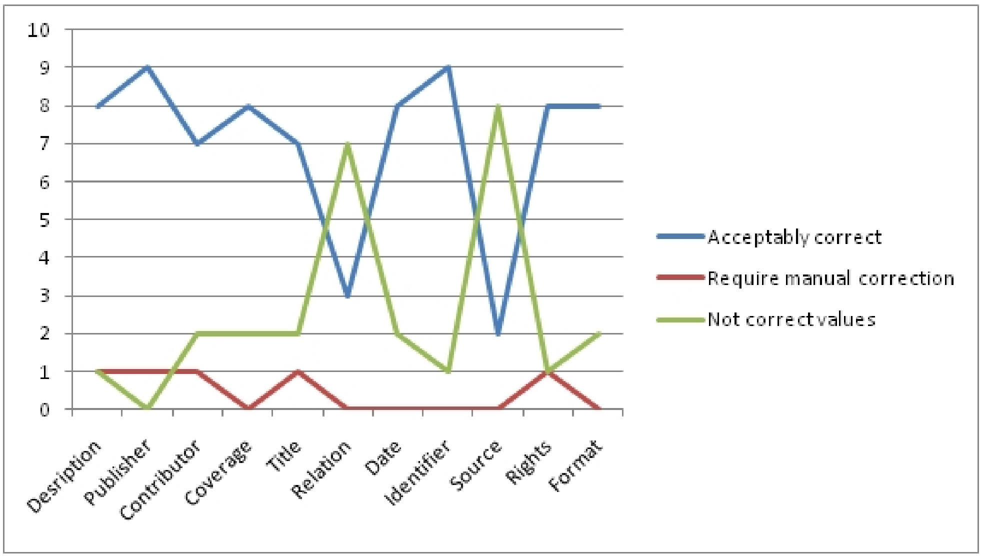


Figure 24: Correctness criteria evaluation

The results of assessing the quality of metadata generated using the proposed methods show good potential in almost all cases, except elements Relation and Source. Information extracted from the syllabus of courses and harvested from the file properties, in most cases is correct and can be used to describe the learning object. However, non-structured syllabus structure makes it difficult to extract the target information.

Logical Consistency defines extent to which the generated values recorded in the metadata fields consistent with the purpose of these fields. In our case as metadata schema was used Dublin Core element set. We based our argumentation on the applicability of particular information source for generation based on definition of elements of the scheme prescribed in the specification Dublin Core. Thus, we suppose that generated metadata values are consistent with their purpose and logically appropriate for describing of learning object.

Provenance reflects reputation of resources for the automatic generation of metadata. In our case, we used two sources for generation - the course syllabus and harvestable information. As described earlier, the course syllabus does not have a unified structure. But a number of rules for information extraction from syllabus have been identified. However, there may exist cases where these rules do not work. Thus we can conclude that the information extracted from the syllabus has not the highest reputation. On the other hand, harvestable information specified automatically by the system and the system is clearly fixes all changes. Thus harvestable information will have a higher reputation than the information extracted from the syllabus. But do not exclude the fact that some elements of the scheme are generated under the control of the person and allow manual correction. In this case view at the reputation of the source may vary and depend on user preferences.

Conformance to Expectations defines the level of significance of the terms used to describe the resource and reflects the applicability of these terms to the search and integration of electronic resources. Since this project does not use a survey and expert opinion to establish a correspondence between the retrieved information and elements of Dublin Core, then all the judgments are based only on a description of the elements in the Dublin Core specification.Further considerations are based on the classification of Dublin Core elements which presented in Table 1 Section 2.4.2.

Elements of Content class allow providing of convenient and accurate search of relevant information among a large volume of accumulated data. Moreover, there is potential for connection of several learning objects together or with other relevant materials.

- To fill field Coverage have been proposed to use information from the course syllabus, which corresponds to the chronology of the classes. This temporal information facilitates search for correlating the requested resource with timeframes.
- To fill field Description was proposed to use the information about the code and course title. This information can relate learning objects with those courses of study in which they were used.
- To fill field Relation have been proposed to use information about relevant resources for a particular learning object. The presence of such information will allow users of learning objects to deepen their knowledge in the study subject, as well authors of learning objects to expand range of offered educational materials. To increase the level of conformity of generated information have been proposed to use human supervision.
- Element Source was proposed to fill by information about the basic course literature. The presence of such information will allow users and creators of learning objects navigate in the sources of the information provided in the content of learning object. To increase the level of conformity of generated information have been proposed to use human supervision.
- Element Title was proposed to fill by information about the topic of ongoing classes. The presence of such information will allow improve the findability of learning object.

Elements of the class Intellectual Property allows you to specify persons who participated in the creation of learning object, specify the persons responsible for the content distribution, and determine the conditions under which learning object can be used.

- To fill the field Contributor have been proposed to use the information about persons who, to some extent involved in the creation of learning object. No doubt these personas can be leaders of the courses. The presence of such information allows setting of those people whose contribution is present in an learning object.
- To fill the field Publisher have been proposed to use information about the organization or service responsible for the publication of learning object. This type of information will allow determine within which organization issued certain learning object and who has the right to use it. It is also proposed to use such information for

fields Rights. But this information can only be offered to the user for confirmation and correction, or also may be supplemented with additional information about the license under which the learning content is distributed.

Element class Instantiation allows to indicate the physical properties of learning object.

- Element Date was proposed to fill by the last modification date of object. The presence of such information allows to navigate in the versions and editions of learning object.
- Element Identifier indicates the physical location of the resource in the repository.

6 Discussion

6.1 Syllabus Benefits

We have considered the course syllabus in this master thesis. Syllabus is examined from the standpoint of its use as a source of information for the automatic generation of metadata for learning objects.

In section 2.6.1 it has been stated that the construction of syllabus is the first step in the course planning. Thus the syllabus is an essential part of the course and any course is necessarily represented by the syllabus, regardless of format, from paper format or publishing the electronic version in network.

Syllabus is designed by educator and must fully represent the important details of the course. To some extent, the course syllabus motivates learners to study the proposed course pointing out what skills he / she must have and can get on successful completion of the study. Moreover during learning, syllabus is a guide for the student. Syllabus designer puts a lot of effort to build a high-quality syllabus and present accurate information about the course.

During the work on the project, it has been found that syllabus includes a wide range of information to describe the course from different angles. We have found that a typical syllabus includes an indication of course information, such as the title of the course, on the basis of which school course is taught; Instructor information, such as identifiers of persons responsible for the course. Other important part is the Course calendar, i.e. presentation of activities plan with dates and correct topic of lectures. Moreover, it has been described that syllabus is closely linked with educational materials provided as part of the course. For specific lecture topic the relevant educational materials are selected and published. This may be direct learning objects or other reading materials.

In this project, the course syllabus primarily is examined in terms of its applicability for the automatic generation of metadata for learning objects. Dublin Core Metadata Element Set has been selected as a metadata schema for the current project. We have drawn an analogy between the information from the fields of the course syllabus and the elements of Dublin Core, based on the definition of the purpose of the element from the vocabulary. After a series of experiments to extract information from the syllabus, we have obtained results which show that the syllabus of the course has the potential to be a source for filling in the Dublin Core metadata schema. Thus, for example based on the classification in Table.5, Course information, Information about reading materials, and Course calendar may be relevant to fill Content elements of Dublin Core. Instructor information from syllabus can be used to fill Intellectual property fields of Dublin Core Metadata Element Set.

6.2 Syllabus Challenges

Considering the course syllabus regarding the automatic generation of metadata, we have encountered several problems, which may affect on the successful retrieval of information from the body of syllabus.

It has been found that the course syllabus may be published in various electronic formats such as text-based formats, for instance PDF, DOC, etc. In other cases, the syllabus is published in HTML format. Ambiguity publication requires different approaches and tools for processing of information. For example, the text presentation format requires text processing, and syllabus in HTML format may require a combination of text processing and tag parsing approaches.

In Section 2.6.1 the typical components of the syllabus are presented. However, there is no single standard of syllabus representation. As has been figured out, various universities and even different courses offered in these universities may have differently designed syllabus. Denoted earlier components of syllabus are not obligatory and their presence depends on the preferences of the designer. From the perspective of using the course syllabus for the automatic generation of metadata, the nonuniformity of information in the syllabus makes it impossible to fill completely all elements of metadata scheme.

While working on the current project, another significant drawback of syllabus has been revealed, namely the unstructured representation of information. Separate parts of content for the syllabus of different universities and even different courses of one university are presented in various forms. For example, as we had defined previously, information in course calendar may be represented by three different formats: table, list, and free text. Moreover, the representation of one format, such as tables, can have a mixed structure, namely a different number of columns and merged cells. This problem complicates the extraction of information from the body of syllabus and may lead to incorrect filling of metadata scheme.

Each of the identified problems can significantly complicate the extraction of information, but usually these problems are combined in the syllabus.

6.3 How to Avoid or Improve Challenges

The presence of problems of ambiguity in the structure, form and publication of curriculum information, require the most rational solutions. Described in Section 2.6.2 Course Description Metadata (CDM) is intended to solve these problems. According to the goals of CDM project, metadata make it easier to describe the education course, set a certain standard for describing the education course, and allow the exchange of descriptive information about the education course for diverse applications.

As mentioned earlier, CDM provides a rich set of elements to describe the wide and narrow parts of the course, defines the semantics and structure of these elements. Any course can be presented in a structured form with elements of CDM. In terms of extraction of information from the course syllabus, the structuredness is a huge plus and makes the processing and determination of the required elements of the syllabus in a straightforward process. However, CDM is not a standard for the direct description of learning objects. Thus, for the realization of the idea of using a structured syllabus as a source of metadata for learning objects it must be installed some kind of analogy between the elements of CDM and Dublin Core in our case. Next, we discuss the possibility of mapping the elements of Course Description Metadata in Dublin Core Element Set. The analogy is based on the description of each element in the specifications.

6.3.1 Analogy between Course Description Metadata and Dublin Core

The CDM specification contains an element *orgUnitName* to describe the organization that produces educational activity. Moreover, the elements *webLink* and *contacts* are able to extend information about the educational institution. Relevant element of Dublin Core is *Publisher*, which also allows specifying the organization responsible for the publication of learning objects.

The CDM specification includes elements *courseName* and *courseCode* for describing information about the course title and to indicate the course code, respectively. In this case the relevant element of Dublin Core *Description* can be filled by sequence courseCode/courseName and allows the use of free text to fill this element.

Elements of CDM specification *name*, *title*, and *role* describe person responsible for course maintaining. Relevant elements of Dublin Core *Contributor* may be filled with a sequence “Professor / Name / Surname”. Having information about the role of the person in the educational process (for example Course leader or Course Assistant) is possible to make a conclusion about its authorship or contribution to the process of learning object creation.

Element *instructionLanguage* of CDM specification describes the language of the course and relevant to element Language from Dublin Core.

The CDM specification contains element *syllabus* as "Information on syllabus, ex books / literature prescribed for study" [36]. This element contains information about literature on which the course is based. In this case, the Dublin Core element *Source* is relevant to *syllabus* element of CDM. In case when *syllabu*s element is listed as sub element of element *timetable* (described below) of CDM scheme then it may contain information about literature appropriate to the described lectures and will be relevant to element *Relation* of Dublin Core.

timetableElement of CDM specification contains information on the ongoing lecture, its topic and description. In our case timetableElement can be used to fill element *Title* of Dublin Core.

Indicate the chronology of teaching activities in the CDM specification is possible by elements containing the actual date value: *CDMdate*. The specification includes an indication of start date and completion date of activity. Relevant element of Dublin Core is *Coverage*.

The Table 6presents results of an analogy between the elements of CDM specifications and elements of the Dublin Core Metadata Element Set.

It is desirable to note the fact that the majority of Dublin Core elements may be represented by combinations of CDM elements. For example element Contributor can be represented by a sequence "name/title/role”.

At the current stage of the project we can bring a direct example of how the course information can be presented in a structured manner using elements of CDM specification listed above. As example we used the syllabus of “IMT4931 Semantic Web” course, which is taught at the Gjøvik University College (Høgskolen i Gjøvik). The course syllabus consists of two parts: The first is the so-called "Student handbook" contains general information about the course, such as Expected learning outcomes, Form of Assessment, Language of instruction, Course responsibility, Teaching Materials (Textbook), and etc. The second part is a direct lecture schedule indicating topics and additional related literature. We joined these two parts and transform them into XML tree format using elements

CDM Element	Dublin Core Element
Organisation Unit: orgUnitName webLink contacts	Publisher
Course Unit: courseName courseCode	Description
Course Unit: instructionLanguage	Language
Course Unit: timetableElement	Title
Course Unit: syllabus	Source/Relation
Person: name title role	Contributor
singleEvent: start CDMdate end CDMdate	Coverage

Table 6: Analogy between CDM and DC

of the CDM specification.

Figure 25 shows only those fields of syllabus, which may have contribution to the automation of metadata generation for learning objects. Without doubt, this scheme can be extended by other necessary elements of syllabus.

One the one hand, latest release of CDM scheme has a well-combined set of elements describing the university, course and learning activities. Moreover, Course Description Metadata has a structured form and thanks to representation in interoperable format XML allows to access individual nodes of XML tree, to share information between different applications, and is understandable to humans. Complete filling of elements of CDM scheme describing the course, the strict relationship between these elements and standardized presentation of results in XML format could lead to a minimal human involvement in the information extraction process from a structured syllabus and improve the accuracy of the extracted information.

But on the other hand, the small numbers of institutions are aware of the existence and benefits of a structured syllabus, and less of implementing this approach in practice. The introduction of new technologies in any process, including education, there is a painful process that requires rethinking of the vision on the problem and presence of special skills.

6.4 Harvesting Issues

In the course of the project it has been found that not all elements of Dublin Core can be filled with information extracted from the course syllabus. For some elements it is suggested to apply the harvestable information, directly related to the file. The results of the

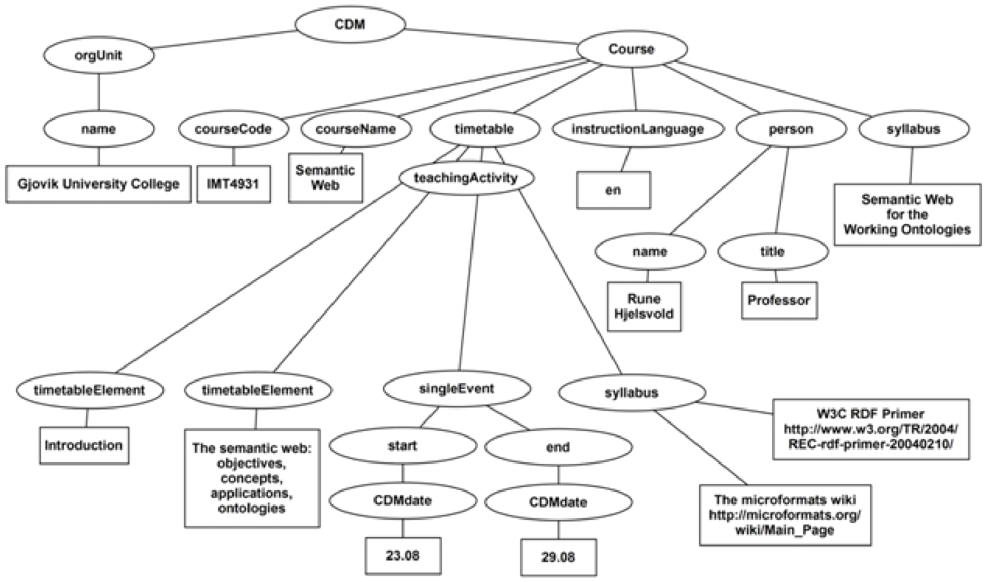

CDM
orgUnit
Course
name
courseCode
courseName
timetable
instructionLanguage
person
syllabus
Gjovik University College
IMT4931
Semantic Web
teachingActivity
en
Semantic Web for the Working Ontologies
name
title
Rune Hjelsvold
Professor
timetableElement
timetableElement
singleEvent
syllabus
Introduction
The semantic web: objectives, concepts, applications, ontologies
start
end
CDMdate
CDMdate
23.08
29.08
W3C RDF Primer http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
The microformats wiki http://microformats.org/wiki/Main_Page

Figure 25: IMT4931 Semantic Web in CDM format (Partly)

experiments shows that harvested information with confidence can be used to describe the physical properties of the learning object, which will allow tracking modifications over the life cycle.

However, in the case of information suitable for describing the content or the contribution of a person in the creation of learning object harvestable information causes some difficulties. For example, fields such as Title or Keywords cannot be filled automatically by the system and require the participation of the person, which in turn implies the subjectivity of information. Moreover, for example, the chosen for consideration the specification of PDF specifies that all fields storing metadata are optional. In this case, there is no strict mechanism for filling the fields, and each author of learning object itself defines the required set of fields. Thus, completeness of scheme also depends on the individual's participation in the process of metadata generation. Another problem discovered during work with harvestable information is that important field, adequate for searching, filled with values, which in fact are not correct and is not relevant. Also may present some confusing values containing inappropriate characters. Another problem associated with metadata harvesting is an inconsistency of schemes. For example the metadata scheme of PDF format and Dublin Core require the establishment of certain rules in the transition from one scheme to another. But these rules can vary from scheme to scheme and from a specific element to element. This ambiguity may complicate interoperability in the transmission and storage of information.

Thus, we note that the harvestable information is suitable and can be used only to describe Instantiation element category of Dublin Core and is of little use to specify the Content and Intellectual Property fields. While on the other hand, in an ideal situation, while respecting the rules harvestable information should be sufficient for a minimum description of learning objects that will enhance searchability of resource in the repository.

6.5 Whether Automatic Metadata Generation Fixes Human Errors

Based on the article [1], in which Cory Doctorow identify barriers to the creation and use of reliable metadata, we can discuss to what extent proposed in this project methods can minimize the negative effects of human intervention.

In describing the learning object human participation can cause a committing of grammatical and other kinds of mistakes that reduce the quality of descriptive information that leads to a reduction of searchability and reusability of the object. Automatic generation requires a high-quality and accurate specifying of information stored in resources for the generation of metadata. We cannot say with absolute certainty that the resources for the generation would not be exposed to human error, but a minimal user interaction and the choice of the provided information will reduce the potential harm.

There is no single way to describe an object, i.e. the subjective opinion and following to personal habits also affects when creating metadata. Two persons may have different visions of the same subject, respectively; the metadata created by one person may not correspond to request entered into the search bar the other person. Such resource for generating metadata as course syllabus is created once and does not change over time; it must conform to the actual description and objectives of the course that is assumed in education regularities. Moreover, harvestable information specified for describing the physical properties of the learning object does not involve human intervention and objectively generated by the system. Thus, automatic generation of metadata using the proposed resources can significantly reduce the ambiguity in describing the learning object.

Intentionally giving false information to promote the content can also be solved by using the proposed methods. In cases where the user does not have an opportunity to correct the information, but only selected from the set of values, the negative impact can be reduced to a minimum.

Another barrier is human laziness, forcing producer to keep some metadata schema fields blank. In this case, the automatic generation takes much of the work itself. User does not need to make much effort to choose, to accept, or not agree with proposed variants for the values of metadata fields.

However, knowing about the existing opportunities to use metadata to provide wider and easier access to published resources, people do not enjoy it. It is necessary to find mechanisms capable of motivating the person to fill at least a minimum set of metadata elements for later reuse of the object. In the current project was proposed to use the Dublin Core Metadata Element Set consisting of fifteen values of the vocabulary. Comparing Dublin Core to describe the objects in any form, whether electronic or not electronic, with specially developed for the description of learning objects LOM Standard which represent the vocabulary of more than 50 fields, it can be said that a minimum set of Dublin Core is more beneficial and less frightening for the author of metadata, and provides the most adequate field corresponding for the search of object.

6.6 Generalization of Proposed Methods

Basically, this project was focused on the metadata generation for learning objects provided in the form of PDF lecture slides. However, as shown by the experimental results and conclusions obtained on the basis of these results, the proposed methods can be confidently used to describe learning objects in other formats. The proposed methods do not involve the direct use of learning object content and use only the context in which it appears. We used information extracted from the course syllabus to generate metadata elements from "Content" and "Intellectual Property Rights" classes. Harvestable metadata is used only to describe the physical properties of the learning object. Thus, there are no serious barriers to the implementation of the proposed methods to describe for example video or audio recorded during the lectures.

Moreover, proposed in this project methods can be applied not only to generate the metadata. These methods do not assume the obligatory binding to the certain learning object. In particular, the identified elements of syllabus and techniques for their extraction can be used for further processing of syllabus and automatic representation it in a hierarchical fashion.

6.7 Criticism of Chosen Methodology

The efforts in this project can be viewed with a critical position. In this project as a resource for metadata generation for learning objects is proposed to use curriculum information provided in the form of the course syllabus, to which the described learning object relates, and harvestable metadata embed directly in file. Extracted information from these resources, we propose to use to fill elements of Dublin Core metadata set. However, although the analogy and correspondence between the generated information and metadata elements is based on the Dublin Core specification, subjective judgments of authors of current project cannot be excluded. Our views was based only on own experience and knowledge, due to absence of similar work in the relevant scientific literature.

Also for consideration, published course syllabi have been involved, which are similar in their structure with the most common ideal syllabus described in the literature. Most often, English was the language of syllabus writing. This implies the exclusion from considering some special cases. Syllabi presented in the text formats such as PDF, DOC, etc. have been also excluded from consideration. However, a deeper text processing could give different results.

In the case of harvestable metadata, we have not considered the Extensible Metadata Platform (XMP) [66], which is an extension to describe the content of files in various formats.

Another drawback of the project can be considered the brief review of techniques for the selection of keywords from content of textual files, and mechanisms for determining the language of the text.

It must be noted that in the second part of the experiment have been used only 10 learning objects, which linked with courses syllabi. We understand that this set may be insufficient for the generalization of the results. However, we tested the proposed approaches and the performance of our prototype, which allowed us to draw some conclusions.

We cannot ignore the fact that no ideal programming skills of authors could also affect the success of the developed prototype and its application in the project. Without doubt this lack influenced the results of the experiments.

7 Conclusion and Future Work

All efforts of this project focused on improving searchability and reusability of learning object. Also this project aims to rationalize the process of metadata creation for leaning objects, reduce cost of creation, and to exclude human participation from this process. For this purpose, we proposed to use an automatic approach. As the resources for generation, we used the course syllabus in the context of which the learning object is used and harvestable metadata stored directly in the file. We have studied the role that the automatic generation can play in the production of metadata and what contribution the proposed resources can make to this process.

As a metadata schema was used Dublin Core Metadata Element Set. Our arguments allowed establishing a correspondence between elements of information sources for the generation and metadata elements of Dublin Core. We also proposed a method that can be used to extract required information from the sources. The experiments were allowed to test and verify the applicability of the proposed methods. Based on these results, we can draw final conclusions and to answer the Research Questions and Sub-questions.

7.1 Sub-question1: What contribution textual resources in form of course syllabus that related to a given learning object can make to the automatic generation of Dublin Core metadata elements?

Basically, to fill the Dublin Core elements of class "Content" as the resource was used the course syllabus, to which this learning object belongs. The results of our experiments have shown that in most cases, the generation of the elements, which directly adequate for search of learning object, has confident potential. Extracted information conformed to criteria of correctness. However, in the case of extracting information about resources related for learning object, which is described, the results were low. Evaluation revealed the poor quality of the generated values.

In the case of the Dublin Core element from the class “Intellectual Property” for generation has also been used the course syllabus. Generated results are also acceptably correct and show confident potential. We managed to retrieve the identifiers of persons who have contributed to the creation, publication and maintenance of learning object, as well as the names of organizations and services in which this learning object is published.

7.2 Sub-question2: What contribution the metadata harvesting can make to the automatic generation of Dublin Core metadata elements?

Harvestable metadata is a good resource of information for the description of the physical properties of the learning object. In this case, information retrieval has yielded good results, and generated values are consistent with high quality indices. In order to avoid controversial issues and exclusion of human involvement, harvestable metadata can be used only to describe the elements of class "Instantiation".

7.3 The main Research Question: To what extent the automatic generation of metadata for learning object using proposed resources can be utilized as a supplement to the manual input?

Unfortunately not all elements of the Dublin Core have the potential to be generated automatically using proposed resources. This is due to the fact that the proposed resources have some drawbacks. In particular, ambiguity publication of the course syllabus requires different approaches and tools for processing of information. The nonuniformity of information in the syllabus makes it impossible to fill completely all elements of metadata scheme. The problem of unstructured representation of information complicates the extraction from the body of syllabus and may lead to incorrect filling of metadata scheme. In some cases, the minimum human intervention is necessary to select, accept or reject offered values generated by the system. Some elements require manual input due to the lack of a clear way to determine the appropriate resources for its generation.

Nevertheless, automatic generation of metadata for learning objects using the suggested resources is a significant supplement to manual input. First of all, automatic method for metadata generation allows filling those fields that can be filled without human intervention using the computing power. Some elements require a minimum of intervention and control by humans. In terms of prevention and correction of human error, the automatic generation of metadata acts as assistant, fixes grammatical and other kinds of mistakes that reduce the quality of descriptive information that leads to a reduction of searchability and reusability of the object. Such resource for generating metadata as course syllabus is created once and does not change over time; it must conform to the actual description and objectives of the course that is assumed in education regularities. Thus, automatic generation of metadata using the proposed resources can significantly reduce the ambiguity in describing the learning object.

7.4 Future Work

- First of all it is necessary to know the opinion of experts in the field of metadata about the correctness of the established correspondence between the elements of the proposed resources for the automatic generation of metadata and schema elements. This will increase the potential and contributions of the proposed resources.
- Another continuation of the work may be focusing on the use of a structured syllabus for a detailed description of the course and use this information as a resource for the automatic generation of metadata.
- From a technical point of view, it is possible to improve the methods for information extraction to obtain more accurate values.
- This project examined only the HTML version of the syllabi. However, it is possible to study the contribution of text version and text-based approaches to the generation of metadata.
- The study of Extensible Metadata Platform (XMP) as extension to describe the content of files in various formats can be a further step for this project.

Bibliography

[1] Doctorow, C. 2001. Metacrap: Putting the torch to seven straw-men of the meta-utopia. *http://www.well.com/ doctorow/metacrap.htm*.

[2] IEEE Learning Technology Standards Committee. 2002. Ieee lom specification. *http://ltsc.ieee.org/wg12*.

[3] Wiley, D. A. "*Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy.*", volume 2830, 1–35. University of Utah, 2000. Last visited: June, 2011.

[4] Polsani, P. 2003. Use and abuse of reusable learning objects. *Journal of Digital Information, 3(4)*.

[5] Motelet, O., B. N. & Pino, J. A. MAY 2007. Hybrid system for generating learning object metadata. *JOURNAL OF COMPUTERS*, VOL. 2(NO. 3).

[6] Ochoa, X., Cardinaels, K., Meire, M., & Duval, E. June 2005. Frameworks for the automatic indexation of learning management systems content into learning object repositories. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005*, Kommers, P. & Richards, G., eds, 1407–1414, Montreal, Canada. AACE.

[7] Cardinaels, K., Meire, M., & Duval, E. 2005. Automating metadata generation: the simple indexing interface. In *Proceedings of the 14th international conference on World Wide Web*, WWW ’05, 548–556, New York, NY, USA. ACM.

[8] Yu, X., Tungare, M., Fan, W., Pérez-Quiñones, M., Fox, E. A., Cameron, W., & Cassel, L. 2007. Using automatic metadata extraction to build a structured syllabus repository. In *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*, ICADL’07, 337–346, Berlin, Heidelberg. Springer-Verlag.

[9] Tungare, M., Yu, X., Cameron, W., Teng, G., Pérez-Quiñones, M., Fox, E., Fan, W., & Cassel, L. 2006. Towards a standardized representation of syllabi to facilitate sharing and personalization of digital library content. In *Proceedings of the 4th International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL)*.

[10] Bauer, M., Maier, R., & Thalmann, S. 2010. Metadata generation for learning objects: An experimental comparison of automatic and collaborative solutions. In *E-Learning 2010*, Breitner, M. H., Lehner, F., Staff, J., & Winand, U., eds, 181–195. Physica-Verlag HD. 10.1007/978-3-7908-2355-4_13.

[11] Motelet, O., Baloian, N., & Pino, J. A. 2006. Learning object metadata and automatic processes: Issues and perspectives. *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.3778*.

[12] Ochoa, X. & Duval, E. June 2006. Quality metrics for learning object metadata. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006*, Pearson, E. & Bohman, P., eds, 1004–1011, Chesapeake, VA. AACE.

[13] Wikipedia. Last visited: June, 2011. Learning object. *http://en.wikipedia.org/wiki/Learning_object*.

[14] IEEE Learning Technology Standards Committee. 2001. Draft standard for learning object metadata version 6.1. *http://itsc.ieee.org/doc/*.

[15] Wiley, D. "*Connecting learning objects to instructional design theory: a definition, a metaphor, and a taxonomy.*", chapter (Bloomington, IN: Agency for Instructional Methodology). The Instructional Use of Learning Objects, 2002.

[16] Cisco Systems, Inc. June 25 1999. Cisco systems, reusable information object strategy. *http://www.cisco.com/warp/public/779/ibs/solutions/learning/whitepapers/el_cisco_rio.pdf*.

[17] Wiley, D. 2001. Acessed (June, 2011). The reusability paradox. *http://cnx.org/content/m11898/1.18/?format=pdf*.

[18] Neven, F. & Duval, E. 2002. Reusable learning objects: a survey of lom-based repositories. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, 291–294, New York, NY, USA. ACM.

[19] Holzinger, D. A. 2001. Multimedia learning systems based on ieee learning object metadata (lom), presented at ed-media. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 772–777.

[20] Anido, L. E., Fernandez, M. J., Caeiro, M., Santos, J. M., Rodriguez, J. S., & Llamas, M. 2002. Educational metadata and brokerage for learning resources. *Computers & Education*, 38(4), 351 – 374.

[21] Noufal, P. P. Last visited: June, 2011. Metadata: Automatic generation and extraction. *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.1691*.

[22] Dahl, D. & Vossen, G. 2007. Learning object metadata generation in the web 2.0 era. *IADIS International Conference e-Learning*.

[23] Taylor, C. An introduction to metadata. *http://www.library.uq.edu.au/iad/ctmeta4.html*.

[24] IEEE. 2006. Ims meta-data best practice guide for ieee 1484.12.1-2002 standard for learning object metadata. *http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html*.

[25] Dimitrios A. Koutsomitropoulos, Andreas D. Alexopoulos, Georgia D. Solomou, Theodore S. Papatheodorou. 2010. The use of metadata for educational resources in digital repositories: Practices and perspectives. *D-Lib Magazine*, Volume 16(Number 1/2).

[26] DCMI. Last visited: June, 2011. The dublin core metadata initiative. *http://www.dublincore.org/*.

[27] Hillmann, D. Last visited: June, 2011. Using dublin core. *http://dublincore.org/documents/2001/04/12/usageguide/sectb.shtml*.

[28] McClelland, M. nov. 2003. Metadata standards for educational resources. *Computer*, 36(11), 107 – 109.

[29] Soylu, A., Kuru, S., Wild, F., & Modritscher, F. Last visited: June, 2011. E-learning and microformats: A learning object harvesting model and a sample application. *http://oro.open.ac.uk/25240/1/10.1.1.142.6443_(1).pdf*.

[30] Roy, D., Sarkar, S., & Ghose, S. 2008. Automatic extraction of pedagogic metadata from learning content. *Int. J. Artif. Intell. Ed.*, 18(2), 97–118.

[31] Altman, H. B. 1992. Last visited: June, 2011. Writing a syllabus. *http://honolulu.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/writesyl.htm*.

[32] Parkes, J. & Harris, M. B. 2002. The purposes of a syllabus. *College Teaching*, 50(2), 55–61.

[33] Gerb, O. & Raynauld, J. Last visited: June, 2011. An open syllabus model. *https://confluence.sakaiproject.org/download/attachments/23330830/ED-MEDIA2009Boston.pdf*.

[34] cdm.utdanning.no. Last visited: June, 2011. example of cdm xml document. *http://cdm.utdanning.no/example_of_cdm_xml_document*.

[35] Pezeril, M. 2006, Last visited: June, 2011. Course description metadata (cdm) : A relevant and challenging standard for universities. *http://www.e-quality-eu.org/pdf/seminar/e-Quality_WS2_MPezeril_article.pdf*.

[36] USIT. 2004. Last visited: June, 2011. A specification of course description metadata. *http://www.usit.uio.no/prosjekter/eSU/eSU-revisjon/CDM/index.html*.

[37] Stubbs, M. & Wilson, S. Last visited: June, 2011. exchanging course-related information: a uk service-oriented approach. *http://dspace.ou.nl/bitstream/1820/837/1/Paper5.pdf*.

[38] JISC. Last visited: June, 2011. Xcri project report. *http://www.elframework.org/projects/xcri*.

[39] Jose A. Pino Janssen, Adriana Berlanga, H. V. R. K. 2008. Towards a learning path specification. *International Journal of Continuing Engineering Education and Life-Long Learning*, 18, 77 – 97.

[40] Islam, M. dec. 2008. An improved keyword extraction method using graph based random walk model. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on*, 225 –229.

[41] Kino High Coursey, Rada Mihalcea, W. E. M. 2008. Automatic keyword extraction for learning object repositories. *http://www.cse.unt.edu/ rada/papers/coursey.asist08.pdf*, 45.

[42] Wikipedia. Last visited: June, 2011. Data pre-processing. *http://en.wikipedia.org/wiki/Data_Pre-processing*.

[43] Manning, C. D., Raghavan, P., & Schtze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[44] Manna, S. & Mendis, B. july 2010. Fuzzy word similarity: A semantic approach using wordnet. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, 1 –8.

[45] van der Plas, L., Pallotta, V., Rajman, M., & Ghorbel, H. 2004. Automatic keyword extraction from spoken text. a comparison of two lexical resources: the edr and wordnet. *CoRR*, cs.CL/0410062. informal publication.

[46] Ljubesic, N., Mikelic, N., & Boras, D. june 2007. Language indentification: How to distinguish similar languages? In *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, 541 –546.

[47] Baldwin, T. 2009. Last visited: June, 2011. Language identification. *http://ww2.cs.mu.oz.au/352/lectures/handout03b.pdf*.

[48] Dunning, T. 1994. Statistical identification of language. *Computing Research Laboratory, New Mexico State University.*

[49] Wikipedia. Last visited: June, 2011. N-gram. *http://en.wikipedia.org/wiki/N-gram*.

[50] Cavnar, W. B. & Trenkle, J. M. 1994. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161–175.

[51] Sibun, P. & Reynar, J. C. 1996. Language determination: Examining the issues. *In Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval*, 125–135.

[52] Mican, D. & Tomai, N. may. 2010. Web 2.0 and collaborative tagging. *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*, 519 –524.

[53] Min, Q. X., Uddin, M., & Jo, G.-S. feb. 2010. The wordnet based semantic relationship between tags in folksonomies. *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, 2, 815 –819.

[54] Limpens, F., Gandon, F., & Buffa, M. sep. 2008. Bridging ontologies and folksonomies to leverage knowledge sharing on the social web: A brief survey. *Automated Software Engineering - Workshops, 2008. ASE Workshops 2008. 23rd IEEE/ACM International Conference on*, 13 –18.

[55] Stuckenschmidt, H. & van Harmelen, F. 2004. Generating and managing metadata for web-based information systems. *Knowledge-Based Systems*, 17(5-6), 201 – 206. Special Issue: Web Intelligence.

[56] Simpson, C. 2010. Last visited: June, 2011. Copyright licensing issues implicated by the learning object repository. preliminary report. *http://thecblor.unt.edu/node/19*.

[57] Casey, J. 2006. Last visited: June, 2011. Intellectual property rights (ipr) in networked e-learning. *http://www.jisclegal.ac.uk/Portals/12/Documents/PDFs/johncasey.pdf*.

[58] Creative Commons. Last visited: June, 2011. Creative commons. *http://creativecommons.org/*.

[59] Free Software Foundation, Inc. Last visited: June, 2011. The gnu general public license v3.0 - gnu project - free software foundation (fsf). *http://www.gnu.org/licenses/gpl.html*.

[60] Stutz, M. Last visited: June, 2011. Design science license. *http://www.gnu.org/licenses/dsl.html*.

[61] Adobe Systems Incorporation. Last visited: June, 2011. Adobe pdf reference version 1.7. *http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf*.

[62] W3Schools. Last visited: June, 2011. Html tutorial. *http://www.w3schools.com/html/*.

[63] Wikipedia. Last visited: June, 2011. File format. *http://en.wikipedia.org/wiki/File_format*.

[64] Kessler, G. Last visited: June, 2011. File signatures table. *http://www.garykessler.net/library/file_sigs.html*.

[65] DCMI. Last visited: June, 2011. Dcmi type vocabulary. *http://dublincore.org/documents/dcmi-type-vocabulary/*.

[66] Adobe Systems Incorporated. Last visited: June, 2011. Adobe xmp: Adding intelligence to media. *http://www.adobe.com/products/xmp/*.

A 60 Corse Syllabi

Course	School	Link
Effective Media Management	Louisiana State University	http://www.lsu.edu/faculty/aclind/syllabus%204050Fall%202001.htm
PUBLIC COMMUNICATION PRACTICES:	Louisiana State University	http://www.lsu.edu/faculty/aclind/syllabus%207209fall%202001.htm
Web Search Engines	New York University	http://www.cs.nyu.edu/courses/fall02/G22.3033-008/
The Adaptive Web	University of Pittsburg	http://www.sis.pitt.edu/~peterb/3954-061/index.html
Statistical Design and Analysis of Experiments	Københavns Universitet	http://www.farma.ku.dk/index.php/Statistical-Design-and-Analysi/7967/0/
Web Programming and Security	Stanford University	http://crypto.stanford.edu/cs142/
Semantic Web	University of Georgia	http://lsdis.cs.uga.edu/SemWebCourse/
Semantic Web Techniques	University of New Brunswick	http://www.cs.unb.ca/~bspencer/cs6795swt/syllabus.html
Web Programming	Marshall University	http://ist.marshall.edu/ist263/syllabus_f03.html
Basic HTML	http://www.simegen.com/	http://www.simegen.com/school/business/webbuilder/html101/index.html
Introduction to HTML Syllabus	Washington University University College	http://www.granneman.com/teaching/washingtonuniversity/webdesign/archives/fall2002syllabus.htm
Information Architecture for the Web	Indiana University at Indianapolis	http://eduscapes.com/arch/course/syllabus.htm
THE CHURCH in the PATRISTIC ERA	Saint John's Abbey School of Theology in Collegeville, Minnesota	http://www.ldysinger.com/CH_583_Patr/webcourse/02_syl-web.htm
The Semantic Web	University of Maryland	http://www.cs.umbc.edu/courses/691m/syllabus.html
Reasoning and Knowledge Representation	University of Illinois	http://reason.cs.uiuc.edu/eyal/classes/f05/cs498ea/
Information Organization	Berkeley School of Information	http://courses.ischool.berkeley.edu/i290-iol/f10/
Introduction to Computing and Programming	University of Georgia	http://www.cs.uga.edu/~bsmith/cs1301/syllabus.html
Information Retrieval and Web Search	The University of Texas and Austin	http://www.cs.utexas.edu/~mooney/ir-course/syllabus.html
MULTICULTURAL COUNSELING	Northern Arizona University	http://jan.ucc.nau.edu/llb3/syl_mult.htm
History 143B: Modern Middle East: 1800-2000	California State University, Sacramento	http://www.csus.edu/indiv/c/chambersh/mmeast/Hs143BSyllS2001.html
Assessing Sustainable Energy Technologies	University of Colorado Boulder	http://leeds-faculty.colorado.edu/lawrence/syst6820/syllabus2.htm
Poetry in an Age of Prose	Purdue University	http://web.ics.purdue.edu/~felluga/eng551/551.html
Artificial intelligence	Stanford University	http://see.stanford.edu/see/courseInfo.aspx?coll=348ca38a-3a6d-4052-937d-cb017338d7b1
Microsystem Engineering	Hamburg University of Technology	http://intranet.tu-harburg.de/kvvz/vorlesung.php3?Lang=en&id=2118
PRINCIPLES OF MANAGEMENT	California State University	http://www.csupomona.edu/~wcweber/301/301syl.html
Principles of Languages	Northern Arizona University	http://www.cefns.nau.edu/~edo/Classes/CS396_WWW/syllabus.html
Object Oriented Programming	Portland State University	http://web.cecs.pdx.edu/~harry/cse509/syllabus.cse509.w2002.html
Network Programming	Johns Hopkins University	http://www.apl.jhu.edu/~jcn/network_programming/fall97/syllabus.html
Dynamic Web Programming	The University of Minnesota	http://cda.morris.umn.edu/~elenam/1101_spring07/index.html
Graphical User Interface Programming	University of Maryland	http://www.cs.umbc.edu/~squire/cs437s_syl.shtml
Systems Analysis	School of Information and Library Science, University of North Carolina	http://www.ils.unc.edu/~stephani/sasp09/home.html
SYSTEMS ANALYSIS	DeSales University	http://www4.desales.edu/~dlm1/it532/class00/syllabus.html
Data Analysis	Southern Methodist University, Dallas	http://faculty.smu.edu/rkemper/anth_7333/anth_7333_SYLLABUS.htm#SCHEDULE
Social network analysis: Sociology 157	University of California, Riverside	http://faculty.ucr.edu/~hanneman/soc157/syllabus.html
Biological Data Analysis	University of Delaware	http://udel.edu/~mcdonald/statsyllabus.html
Fourier Analysis	University of California	http://www.math.ucdavis.edu/~saito/courses/129.s10/syllabus.html
Empirical Political Analysis	West Virginia University	http://www.polsci.wvu.edu/duval/ps100sum/100syl.html
INTELLIGENCE ANALYSIS	http://www.drtomoconnor.com/	http://www.drtomoconnor.com/4125/default.htm
Analysis of Software Artifacts	Carnegie Mellon University	http://www.cs.cmu.edu/~aldrich/courses/654/
Military Operations Research: Cost Analysis	George Mason University	http://classweb.gmu.edu/aloerch/OR651_S03.htm
VIRTUAL SCRIPT ANALYSIS	Washington State University	http://www.wsu.edu/~converse/scriptsyl.html
Multivariate Analysis	http://www.philender.com/	http://www.philender.com/courses/multivariate/mulsyl.html
Microeconomic Analysis	Wesleyan University	http://chogendorn.web.wesleyan.edu/wescourses/2006s/econ301/02/
Corpus-based text analysis	King's College London	http://www.kcl.ac.uk/schools/humanities/depts/cch/pg/madh/avmlan/syllabus.html
Numerical Analysis	University of Illinois	http://www.math.uic.edu/~hanson/mcs471syllabus.html
Consumer Analysis	The State University of New Jersey	http://crab.rutgers.edu/~ckaufman/ConsumerAnalysisSyllabusSpring05.html
QUALITATIVE ORGANIC ANALYSIS	SOUTHWESTERN OKLAHOMA STATE UNIVERSITY	http://faculty.swosu.edu/william.kelly/qorg.htm#classtime
NUMERICAL ANALYSIS	Troy University	http://spectrum.troy.edu/~belyi/num_an/Syll_Num_An.htm
Event History Analysis	UCI School of social Sciences	http://www.socsci.uci.edu/~schofer/2008Soc229EHA/syllabus229EHA.htm
INFORMATION SYSTEMS ANALYSIS	UNIVERSITY OF NEBRASKA AT OMAHA	http://www.isqa.unomaha.edu/pietron/isa/isafl.htm
Organization of Information	School of Information and Library Science, University of North Carolina	http://ils.unc.edu/~stephani/orginfsp05/syllabus.html
Introduction to Information Literacy	Lake Land College	http://webclass.lakeland.cc.il.us/info_lit/syllabus.html
Information	Horace Mann School	http://mail.horacemann.org/~weitz/information/infosyl.html
Information Systems Management	Massey University	http://www.massey.ac.nz/~dviehlan/157700.html
Information Visualization	Georgia Tech College of Computing	http://www.cc.gatech.edu/~stasko/7450/index.html
Information Organization	Simmons College	http://web.simmons.edu/~joudrey/415/syllabus.htm
ORGANIZATION OF INFORMATION AND RESOURCES	University of Washington	http://faculty.washington.edu/acarlyle/lis_530_syllabus.htm
Information Technology	University of Maryland	http://www.umiacs.umd.edu/~jimmylin/LBSC690-2008-Fall/
Organization of Information	University of Arizona	http://timon.sir.arizona.edu/faculty/coleman/501/fall04/
Semantic Web	Gjøvik University College	http://english.hig.no/course_catalouge/student_handbook/2010_2011/courses/avdeling_for_informatikk_og_medieteknikk/imt4931_semantic_web

Table 7: 60 Corse Syllabi

B 30 Learning Objects

Course	**School**	**Number of LO**
Coding and compression of media data	Gjøvik University College	5
Semantic Web	Gjøvik University College	4
Media Management and Business Development	Gjøvik University College	1
Web Programming and Security	Stanford University	5
Semantic Web Techniques	University of New Brunswick	2
The Semantic Web	UMBC: An Honors University in Maryland	5
Military Operations Research: Cost Analysis	George Mason University	5
QUALITATIVE ORGANIC ANALYSIS	SOUTHWESTERN OKLAHOMA STATE UNIVERSITY	3

Table 8: 30 Learning Objects

C Results of the Second Part of Experiment

LO	Description	Publisher	Contributor	Coverage	Title	Relation	Date	Identifier	Source	Rights	Format
1	CS142 Web Programming and Security	Network Security at Stanford	John Ousterhout, Dan Boneh, John Mitchell, Collin Jackson	1/30/09	Cookie same origin policy; Basic cross site scripting attacks (XSS)	-	30.01.2009	http://.../10-cookie-security.pdf	-	Copyright Network Security at Stanford	pdf
2	QUALITATIVE ORGANIC ANALYSIS	SWOSU Southwestern Oklahoma State University	Dr. William J. Kelly	WEEK 4	Phenols	-	31.12.1997	http://... /qo8.pdf	The Systematic Identification of Organic Compounds ...	Copyright SWOSU Southwestern Oklahoma State University	pdf
3	CS6795 Semantic Web Techniques	UNB Faculty of Computer Science	Harold Boley, Bruce Spencer, NRC	Week 1 Sep 6	Intro to course Semantic Search Engines	-	07.09.2007	http://... /sw10pass-talk.pdf	-	Copyright UNB Faculty of Computer Science	pdf
4	INFSCI 3954: The Adaptive Web	School of Information Sciences University of Pittsburgh	-	Wednesday October 5	Adaptive systems in health care	-	05.10.2005	http://... /AdaptiveHealthcarePresentation.ppt	-	Copyright School of Information Sciences University of Pittsburgh	ppt
5	Stanford School of Engineering - Stanford Engineering Everywhere	Stanford School of Engineering - Stanford Engineering Everywhere	Ng, Andrew	-	-	-	11.07.2008	http://... /Lecture01.pdf	-	Copyright Stanford School of Engineering - Stanford Engineering Everywhere	pdf
6	UMBC CMSC 491/691m The Semantic Web	Computer Science and Electrical Engineering	-	2/20	RDF/RDFS	SWP 84-106, The shortest path to the future web	20.02.2007	http://... /05rdfs.pdf	-	-	pdf
7	INLS 582_001	UNC School of Information and Library Science	Stephanie W. Haas	Tuesday, 1/13/09	Introductions, business	-	-	http://... /sa-overview.pdf	-	Copyright UNC School of Information and Library Science	-
8	Biological Data Analysis	University of Delaware	John McDonald	Sept. 7	Hypothesis testing	http://... /stathw1.html	-	http://... /stathyp-testing.pdf	-	Copyright University of Delaware	-
9	INFORMATION SYSTEMS ANALYSIS	ISQA	Dr. Leah R. Pietron	September 5	The Systems Development Environment	Chapter 1	17.05.2005	http://... /chapter01.pdf	-	Copyright ISQA	pdf
10	Semantic Web - Gjøvik University College	GUC - Gjøvik University College	Rune Hjelsvold	-	-	-	27.08.2010	-	ISBN-13: 978-0-12-373556-0	Copyright GUC - Gjøvik University College	pdf

Table 9: The second part of experiment results

In some cells "..." stands for actual values in order to save space, "-" stands for empty values.

D Source Code Examples

D.1 Element "Title" generation

```
<?php

$val="http://cda.morris.umn.edu/~elenam/1101_spring07/syllabus.html"; // input URL

$text = file_get_contents($val);
  $doc = new DOMDocument(); //create new DOMDocument

    $doc -> loadHTML( $text ); //load html page
    $titles = $doc->getElementsByTagName('title'); // get data from tag title
    for ($i = 0; $i < $titles->length; $i++) {
    $str = $titles->item($i)->textContent;
    $array[] = $str; //adding in array

echo "</br>";
echo $str;

$str_rep = "-!#%?/.,\'\""; // remove punctuation and side-text

$str = ereg_replace("[^a-zA-Z] ", " ", $str);
$str = str_replace("Syllabus", "", $str);
$str = str_replace("Syllabus for", "", $str);
$str = str_replace("Homepage", "", $str);
$str = str_replace("Course materials", "", $str);

echo "</br>";
echo $str;
}


?>
```

D.2 Element "Contributor" generation (names)

```
<?php

$file = file("C:\websites\test.txt"); // syllabus text as input

foreach($file as $str)
  {

$str = strtolower($str);

// extraction of course instructor information using defined keywords

$f = ereg("instructor", $str);
if ($f == 1)
{
  $str = str_replace(":", "", $str);
  $str = str_replace("instructor", "", $str);
  echo $str. "</br>";
}

$f = ereg("professor", $str);
if ($f == 1)
{
  $str = str_replace(":", "", $str);
  $str = str_replace("professor", "", $str);
  echo $str. "</br>";
}

$f = ereg("dr.", $str);
if ($f == 1)
{

  echo $str. "</br>";
}

$f = ereg("doctor", $str);
if ($f == 1)
{

  echo $str. "</br>";
}
  }
fclose($file);


?>
```

D.3 Element "Contributor" generation (e-mail information)

```
<?php

$val="http://crypto.stanford.edu/cs142/staff.html"; // input URL

$text = file_get_contents($val);
  $doc = new DOMDocument(); //create new DOMDocument

    $doc -> loadHTML( $text ); //load html page
    $titles = $doc->getElementsByTagName('a'); // get data from tag <a>
    for ($i = 0; $i < $titles->length; $i++) {
    $str = $titles->item($i)->textContent;
    $array[] = $str; //adding in array



$f = ereg("@", $str); // extract email information
if ($f==1)
{
  echo "</br>";
  echo $str;
}
}


?>
```

D.4 Table processing

```

<?php
//ini_set('display_errors',1);
//error_reporting(E_ALL);

$val="http://www.cs.uga.edu/~bsmith/cs1301/syllabus.html"; // input URL

$text = file_get_contents($val);
$n= 6;
  $doc = new DOMDocument(); //create new DOMDocument

    $doc -> loadHTML( $text ); //load html page
    $titles = $doc->getElementsByTagName('td'); // get data from tag <td>
    for ($i = 0; $i < $titles->length; $i++) {
    $str = $titles->item($i)->textContent;
    $array[] = $str; //adding in array


}

$doc = new DOMDocument(); //create new DOMDocument

    $doc -> loadHTML( $text ); //load html page
    $titles = $doc->getElementsByTagName('tr'); // get data from tag <tr>
    for ($i = 0; $i < $titles->length; $i++) {
    $str = $titles->item($i)->textContent;
    $arr[] = $str; //adding in array
    }
$tr = count($arr);


$a = count($array);
//echo $a;
$a1 = 0;
$array1[] = $array[$a1];
$a2 = 1;
$array2[] = $array[$a2];
$a3 = 2;
$array3[] = $array[$a3];
$a4 = 3;
$array4[] = $array[$a4];
$a5 = 4;
$array5[] = $array[$a5];
$a6 = 5;
$array6[] = $array[$a6];

for ($i=1; $i<=$tr-1; $i++)
{
  $array2[] = $array[$a2+$n*$i];
  $array3[] = $array[$a3+$n*$i];
  $array4[] = $array[$a4+$n*$i];
}

foreach($array2 as $val)
{
  echo $val;
  echo "</br>";
}

foreach($array3 as $val)
{
  echo $val;
  echo "</br>";
}

foreach($array4 as $val)
{
  echo $val;
  echo "</br>";
}

//echo $array2[3], $array3[3], $array4[3];
?>
```

D.5 Free-text syllabus processing 1

```

<?php
//ini_set('display_errors',1);
//error_reporting(E_ALL);

$file = file("free_text.txt");

foreach($file as $str)
{

  $f = ereg("Topic:", $str);
  if ($f == 1)
  {
    $str = str_replace("Topic:", "", $str);
    echo $str. "</br>";
    $array_topic[] = $str;
  }

  $f = ereg("Date:", $str);
  if ($f == 1)
  {
    $str = str_replace("Date:", "", $str);
    echo $str. "</br>";
    $array_date[] = $str;
  }

  $f = ereg("Book:", $str);
  if ($f == 1)
  {
    $str = str_replace("Book:", "", $str);
    echo $str. "</br>";
    $array_book[] = $str;
  }
}

echo $array_topic[1], $array_date[1], $array_book[1];

?>
```

D.6 Free-text syllabus processing 2

```
<?php
//ini_set('display_errors',1);
//error_reporting(E_ALL);

$file = file("free_text3.txt");

$month = array("January ","February ","March ","April ","May ","June ","Jule ","August ","September ","October ","November ","
      December ");
foreach($file as $str)
{
  foreach($month as $val)
  {
  $f = ereg($val, $str);
  if ($f == 1)
  {
    $tok = explode(":", $str);
    for ($i = 0; $i<=count($tok); $i++)
    {
      $ost = $i % 2;
      if ($ost == 0)
      {
        $date[] = $tok[$i];
      }
      else
      {
        $topic[] = $tok[$i];
      }
    }

  }

  }

}

foreach ($date as $d)
{
  echo $d;
  echo "</br>";
}

foreach ($topic as $d)
{
  echo $d;
  echo "</br>";
}

?>
```

D.7 Textbook information extraction

```
<?php
```

```
ini_set('display_errors',1);
error_reporting(E_ALL);

$file = file("text.txt");

foreach($file as $str)
{

  $f = ereg("Textbook", $str);
  if ($f == 1)
  {
    echo $str;
    $tok =  explode(":", $str);
    for ($i = 0; $i<=count($tok); $i++)
    {
      $ost = $i % 2;
      if ($ost != 0)
      {
        $book[] = $tok[$i];
      }
    }

  }

  $f = ereg("Text", $str);
  if ($f == 1)
  {
    $tok =  explode(":", $str);
    for ($i = 0; $i<=count($tok); $i++)
    {
      $ost = $i % 2;
      if ($ost != 0)
      {
        $book[] = $tok[$i];
      }
    }

  }

  $f = ereg("Book", $str);
  if ($f == 1)
  {
    $tok =  explode(":", $str);
    for ($i = 0; $i<=count($tok); $i++)
    {
      $ost = $i % 2;
      if ($ost != 0)
      {
        $book[] = $tok[$i];
      }
    }

  }

  $f = ereg("Reading", $str);
  if ($f == 1)
  {
    $tok =  explode(":", $str);
    for ($i = 0; $i<=count($tok); $i++)
    {
      $ost = $i % 2;
      if ($ost != 0)
      {
        $book[] = $tok[$i];
      }
    }

  }

  $f = ereg("ISBN", $str);
  if ($f == 1)
  {
    $book[] = $str;
  }

  }

}

foreach ($book as $d)
{
  echo $d;
  echo "</br>";
}

?>
```