# The Use of the Omeka Platform for Digital Libraries in the Field of Mining

UDC 004.738.5:027.7]:004.42OMEKA

DOI 10.18485/infotheca.2017.17.2.2

Aleksandra Tomašević
aleksandra.tomasevic@ rgf.bg.ac.rs

Biljana Lazić
biljana.lazic@rgf.bg.ac.rs

Dalibor Vorkapić
dalibor.vorkapic@rgf.bg.ac.rs

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs

Ljiljana Kolonja
ljiljana.kolonja@rgf.bg.ac.rs

*University of Belgrade*
*Faculty of Mining and Geology*

*Translated by:* Marko Vitas

**ABSTRACT:** This paper will introduce Omeka, a platform for presentation of digital collections and a system for the management of their content. We will illustrate its application in the field of technical sciences (more specifically, in the field of mining) on the example of the digital library ROmeka@RGF. We have decided to use Omeka because it is simple, because it possesses comprehensive supporting documentation and because it does not require any attainment in information sciences, which makes it accessible for most users, and especially for mining engineers, to whom this digital library is chiefly intended. Documents assembled and stored in this digital library will serve as a basis for future research, extraction of terminology, tagging, extraction of knowledge etc.

**KEYWORDS:** Omeka, digital library, mining.

**PAPER SUBMITTED:** 2 September 2017
**PAPER ACCEPTED:** 11 October 2017

## 1 Introduction

For the purposes of this paper, a digital library ROmeka@RGF[1] was created in order for the academic texts in the field of mining to be assembled, systematized, processed and stored. This is meant to serve as an important basis not only for various linguistic and terminological research studies, but also for a whole array of tasks related to the knowledge engineering (not

[1] http://romeka.rgf.rs

only the extraction of knowledge). Documents have been assembled, processed and stored in the digital library in the course of the projects financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia. Inspiration for this project came from the need to make the texts in the field of mining more readily accessible to both mining engineers and linguists. In the course of the preliminary research for this project no digital library for the field of mining was discovered, except for the journal „Underground Mining Engineering" (original title in Serbian: „Podzemni radovi"), which is a component of the bilingual library „Bibliša"[2]. Therefore, the digital library ROmeka@RGF can be considered the first of its kind in Serbia. According to our knowledge, the only example of a digital library on the Omeka web platform, in the field of geology (a field narrowly connected to mining) is the digital library „Geological books of Philippe Glangeaud" (*Carnets géologiques de Philippe Glangeaud*[3]). This digital library had been developed on the Clermont university.

Section 2 of this paper contains a description of the developmental environment of the web platform Omeka and a review of the software's most useful plugins. Section 3 contains a description of the most important elements of the digital library and the ways these elements are created. Section 4 contains a description of the digital library ROmeka@RGF itself. Section 5 explains how to search through the stored textual resources with help of tools and resources available for Serbian. Section 6 contains possible implementations of the TEI guidelines[4]. Section 7 contains conclusions and ideas for future research.

## 2 The Web Platform Omeka

Omeka, the web platform for displaying digital collections and the system for managing their contents[5], has been used for the creation of the digital library ROmeka@RGF. Omeka was developed in the Roy Rosenzweig Center for History and New Media at the George Mason university in Virginia[6]. It belongs to the Open Source Software group and has the General Public Li-

[2] http://jerteh.rs/biblisha/ListaDokumenata.aspx?JCID=2&lng=en

[3] http://bibliotheque.clermont-universite.fr/glangeaud/

[4] TEI: Text Encoding Initiative http://www.tei-c.org/index.xml

[5] Content Management System (CMS)

[6] Roy Rosenzweig Center for History and New Media (RRCHNM), https://rrchnm.org/

cense (GPL v3.0)[7]. This means that its source code is publicly available and that each user may upgrade or adapt it to meet his or her own needs. While the platform's original target audience were academic institutions studying cultural heritage, today it is used by researchers in many different fields.

Omeka's flexibility and user-friendliness earned it a top position among the software of its kind. Its main features are: attractive and flexible visual design, simple installation, capacity to spread which enables replacement of the existing functions and the addition of the new ones, flexible approach to meta-data, support for web standards (CSS, HTML, RSS), import and export of the data in the standardized formats (RDF, CSV, XML, JSON) (Kucsma et al., 2010).

Since the software had not been designed for IT-experts and does not require a high level of proficiency in information sciences, it allows the users to focus on the contents of the digital library, as well as on the description and interpretation of its contents, rather than on programming. Furthermore, the platform disposes of functions for catalogization and presentation of digital items. These functions are based on the Dublin Core, which ensures that the description and organization of digital items is standardized.

## 2.1 The developmental environment of Omeka

There are two basic versions of the web platform Omeka:

- Omeka.neta, a version which does not require its own server. Its file storage capacity is limited to 500MB, the size of a single file is limited to 64MB and the number of available plugins is limited to fifteen. It does not allow for functionality adjustments and options for the platform's appearance adaptation are severely curbed. The lite version can be upgraded, but this requires the user to pay between 35 and 1000 dollars a year, depending on the package.
- Omeka.org, a complete version which can be installed on a local disk or as a virtual machine and allows all functionality adjustments.

Certain adjustments are needed on the server before Omeka can be installed and they include the installment of: the web (HTTP) server Apache, the database management system MySQL (either 5.0 or a newer version) and the program language interpreter PHP (either 5.3.2 or a newer version). The following distributions of the operating system Linux allow Omeka to work

[7] https://www.gnu.org/licenses/gpl-3.0.en.html

steadily: Fedora, OpenSuse and Ubuntu. ROmeka@RGF has been installed on a virtual machine with the operating system Ubuntu 15.10.

The installation of the web platform is initiated by creation of the MySQL base with administrator privileges. The newest version of Omeka is downloaded from the platform's official web page and unpacked. The resulting folder (hereinafter: *omeka-root*) is stored either in the root folder of the web server or in one of its subfolders. Certain changes have to be made in the file *omeka-root/db.ini* (the following field values: database host, username, password and database name) in order for the database management system MySQL to become operational. All folders containing Omeka are required to have a permit. After the installation had been completed it is necessary to adjust privileges. The installation is launched through the web-reader by entering the IP address or domain of the digital library, where the administrator account and the site name have to be defined.

The web platform is multilingual and translated, either fully or partially, into 50 languages. The Serbian version is among the ten that have been fully translated.

### 2.2 The Omeka software plugins

The appearance of the web platform can be altered and adapted: a user is free to either choose among fifteen offered themes or create his or her own theme. Meanwhile, functionality is increased through plugins. Ninety plugins have been developed and adapted to different versions of Omeka. The users of Omeka have developed over 300 plugins of their own, but since they were mainly designed for the older versions of Omeka, further adjustment is needed for the newest version of the platform. Depending on their purpose, plugins can be classified into several categories: plugins for mass creation of collections and items, plugins for content management, plugins for file inspection, plugins for the community contribution descriptions, plugins for geospatial processing and map navigation. Here we will review several plugins, both such that were used for the creation of our digital library and such that we deemed important and useful for all and any users of Omeka.

*Archive Repertory* enables the user to keep the imported files on the server with their source names. It also allows files to be gathered in a hierarchical structure, based either on items or on collections containing these items. The ability to keep source names makes it easier to read URL-addresses and to manage files.

*Bulk Metadata Editor* enables the user to search and update the meta-data for a large number of items simultaneously and in a fast and simple way. The first step is to make a selection of items based on various criteria (for instance, a selection can encompass: all items of a digital library, items located in a certain collection and items incorporating meta-data which meet one or several criteria). The second step requires the user to choose the piece of meta-data which should be altered. In the third and concluding step, the meta-data is finally altered (for instance, one can alter the following: search and replacement of text, addition of new meta-data in the selected field, addition of text to the existing meta-data, removal of duplicates and empty fields in the selected description of an item, removal of duplicated files in the selected items or removal of all existing meta-data in the selected fields).

*Catalog Search* makes it possible to search through other catalogues by using the field dc: subject. These catalogues include: Archive Grid, Digital Public Library of America, Google Books, Google Scholar, Hathi Trust, JSTOR, Library of Congress, WorldCat. It is equally possible to add links to the catalogues of other institutions.

*COinS*[8] arranges for the meta-data about citations for every item to be embedded into web-pages of the digital libraries. Once activated, *COinS* provides visibility for objects on on-line platforms such as Zotero[9] by automatically embedding meta-data about citations into other web-pages. Additionally, individual items of any Omeka site can be added to the library of the platform Zotero, while the simultaneous addition of several items can be implemented through the scripts operating in the background. *COinS* plug-in makes the research and the interoperability with other systems much easier.

*Collection Tree* secures a visual representation of the hierarchical structure in the digital library, consequently making the browsing through them easier.

*CSV Import* enables mass import of meta-data, labels and files, which are represented by a table in the CSV format. If the titles of the columns are in accordance with the Dublin Core, then the data are mapped automatically; if they are not, then it is necessary to map the data manually.

*Drop Box makes* makes the import of data located on the server (in the catalogue */plugins/Dropbox/files*) simpler for the administrator. The admin-

[8] ContextObjects in Spans (COinS), http://omeka.org/codex/Plugins/Coins_2.0, accessed May 28, 2017.

[9] https://www.zotero.org/

istrative interface allows the data to be imported either individually or on a large scale through selection from a list.

*Dublin Core Extended* expands the list of meta-data of the Dublin Core. This secures a total annotation of items. To 15 original elements (title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights) additional 40 are added: abstract, access rights, accrual method, accrual periodicity, accrual policy, alternative title, audience, date available, bibliographic citation, conforms to, date created, date accepted, date copyrighted, date submitted, audience education level, extent, has format, has part, has version, instructional method, is format of, is part of, is referenced by, is replaced by, is required by, date issued, is version of, license, mediator, medium, date modified, provenance, references, replaces, requires, rights holder, spatial coverage, table of contents, temporal coverage, date valid.

*Geolocation* arranges for the information about locations relative to the digital items to be added to the maps and enables the user to search through them.

*Item Relations* enables the creation of the relation between the digital objects. This plug-in follows the RDF model for defining relations between items. This model is represented as an RDF graph made out of RDF triplets: subject-predicate-object. For example, if one digital item (RDF subject) is a part of another digital item (RDF object), then a correlation *isPartOf* (RDF predicate) is established between them. Similarly, if one digital item is a version of a document, then between these digital items a correlation *isVersionOf* is established. This is how the RDF triplets are formed. These triplets enable the text to be searched through at a later point by using the techniques of the semantic web.

*Hide Elements* enables the user to choose meta-data which will be hidden on the import form, on the web-page of the administrator and/or on the publicly available web-page, and also on the form for the browsing of the meta-data.

*METS Export*[10] enables the export of digital items such as METS XML files, individual files, collections or whole digital libraries. It is supported by the Initiative of the Digital Library Federation[11], which suggests an XML scheme of meta-data for the management of items in a digital library and

[10] The Metadata Encoding and Transmission Standard,
https://www.loc.gov/standards/mets/METSOverview.v2.html

[11] Digital Library Federation, https://www.diglib.org/?s=mets

for their exchange between repositories or between a repository and a user. It is most notably used for gathering and keeping of documents contained in a digital item, given their number and variety. *METS Export* links several digital documents and enables navigation between them. Additionally, it contains technical information necessary for managing digital items: formats, technological characteristics, ways of scanning, digital transformations. *METS Export* does not require a specific group of meta-data to be entered for a digital document, but instead allows the creator of the meta-data to decide which meta-data he or she would like to enter for that purpose. Descriptive meta-data for METS can easily be downloaded from the Dublin Core (Трговац, 2016).

*Neatline*, *NeatlineFeatures*, *NeatlineSmile*, *NeatlineText*, *NeatlineTime* and *NeatlineWaypoints* represent a series of plug-ins which allow the spatial and temporal points on the items map in the digital library to be interconnected. It also allows the documents to be connected with the *Neatline* exhibition. These plug-ins have not been activated in the digital library ROmeka@RGF because they are incompatible with the plug-in *Geolocation*, which is far more useful for engineers.

*OAI-PMH Harvester* gathers meta-data from an *OAI-PMH*[12] data supplier, maps them in a local database and imports them. It can be used both on a short-term and on a permanent basis for updating, synchronization and distribution of the systems. At the present time, the formats it is able to import are the Dublin Core and *METS*.

*OAI-PMH Repository* prepares items for exchange and is functionally in an inverse relationship to the plug-in discussed previously *(OAI-PMH Harvester*). It supports the Dublin Core, MODS and METS.

*PDF Text* enables the optical recognition of the text characters[13], the extraction of the text from the PDF format and the search through the text. If the text is not read satisfactorily it is possible to correlate it or to import a new textual document to the provided field.

*Reference* adds pages equipped with alphabetical index of elements, which, defined in advance, in turn allow browsing the meta-data which were also set in advance.

*Search By Metadata* allows the administrator to define metadata for advanced searches on a HTML page.

[12] Open Archives Initiative Protocol for Metadata Harvesting,
https://www.openarchives.org/pmh/

[13] Optical character recognition (OCR)

*SimplePages* enables the administrator to create dynamical PHP pages without requiring a specific informatic attainment.

*SimpleVocab* and *SimpleVocabPlus* enable the creation of controlled dictionaries and their synchronization on a cloud. In the digital library ROmeka@RGF a controlled authorial dictionary had been created. This secured a consistent export and made searching easier.

*TEI Display* renders a TEI file (which had been prepared and joined to an item) into a visually clear shape. The understood XSLT transformation enables two presentation modes: either the whole document or its individual units are presented. While the first mode presupposes the transformation of the whole document into a HTML, the second presents the content of the document (div1 or div2). This latter option is particularly suited for larger documents. Both the presentation modes and XSLT transformations can be adapted, while the metadata from the TEI heading can be automatically mapped in fields of the Dublin Core for items and files.

## 3 The creation of the digital library

According to one of the most frequently cited definitions of digital libraries, conceived by William Y. Arms, digital libraries are controlled and systematically organized collections of information, with adjoined services, which are stored in digital format and accessible through the web. The common feature of all digital libraries is the fact that information is organized on computers and accessible via internet, together with procedures for their selections, organization (so that they can become more readily accessible) and archiving (Arms, 2000).

Digital libraries are collections of digital items which are stored on the web in form of various digital data (e.g. text, image, sound, video, animation) or in form of combinations of digital items (multimedia). They are described by various metadata and connected to other informational services. End users can access and use them without any temporal or spatial limitations. They can also create new digital items and update the old ones (Тртовац, 2016).

Basic elements of a digital library in Omeka are:

- items,
- collections and
- web pages.

The user is free to import an unlimited number of items, documents, signs and collections. The only limitation is that one item can be adjoined to one collection only (Figure 1)[14]. Each of the aforementioned elements of the digital library is liable to a complete visibility control on the web. This control encompasses everything from individual metadata to all of the element in the whole.

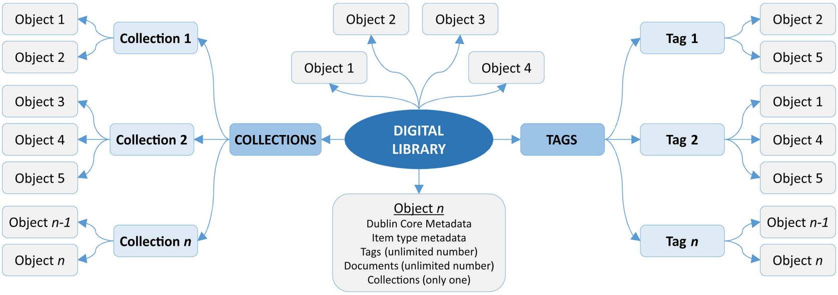


**Figure 1.** A diagram of the elements of a digital library in Omeka

## 3.1 Items

The Omeka web platform has been designed for presentation of objects, as they are the fundamental element of every digital library. Therefore, the creation of a digital library necessarily begins by the creation of items.

Depending on their type, items can vary (and the same goes for units, archives, sources or resources). User is provided with a list of 15 basic types of items. There is also a possibility to add new ones if needed. The following item types are accessible:

- *Moving Image* – video recordings of all sorts: animations, films, television shows;
- *Sound* – audio recordings of all sorts: audio compact disks, recorded speech or sound;
- *Oral History* – information obtained through interviews with persons in possession of first-hand knowledge;

[14] https://omeka.org/codex/Managing_Items, access date 19. May 2017

- *Still Image* – visual presentation of texts, images, drawings, graphic design, plans or maps;
- *Website* – HTML pages with adjoined images, audio and video files etc.;
- *Event* – temporally limited phenomena, for example: an exposition, a web conference, a workshop, a tea party, a fire, a battle, a trial, a wedding;
- *Email* – textual messages with optional attachment(s), sent by one person to the other person or persons;
- *Lesson Plan* – a detailed description of teaching process throughout a course;
- *Person*;
- *Interactive Resources* – web pages, multi-media classroom items, chat services;
- *Dataset* – coded data in a defined structure: lists, tables and databases;
- *Physical Objects* – three-dimensional solid inanimate objects, represented in digital libraries by types such as moving or static pictures et al.
- *Services* – for example: copying services, banking services, interlibrary loans or web servers;
- *Software*;
- *Hyperlinks* – a link or a reference to a different resource on the Internet.

Items are collections of:

- metadata from the Dublin Core describing the digital item itself,
- item type metadata,
- tags,
- documents.

Metadata are structured in such a way that enables them to describe, explain, identify, locate or enables them to make retrieval, use and managing of the information source easier in some other way (Hodge, 2001). They can be:

- *descriptive* – they describe resources that are needed in order to find and identify the source of information. They contain some basic elements such as: title, author, publisher, place, year, language, unique identifier, description, keywords, subject headings, abstract etc.;
- *structural* – they describe the structure of complex resources: types, versions, connections between digital items and other features, connections between the original document and its versions, including the data about changes and other features;

– *administrative* – they offer information on use and managing of resources in connection to the intellectual right. They can be:
  - metadata on copyright (which define the management of rights to access a digital item in accordance with the authorial rights and with the protection of the intellectual property);
  - technical metadata (which contain data on the creation date, technical details of the source, size and type of the file, access to the source, data on all the changes and format of the presentation);
  - metadata on conservation of a digital item;
  - metadata on use (they refer to active tracking of numbers of users who visit and use a particular content, as well as to tracking of use of a digital item content in a new context or in a new version by downloading the metadata and the digital item for a different digital library).

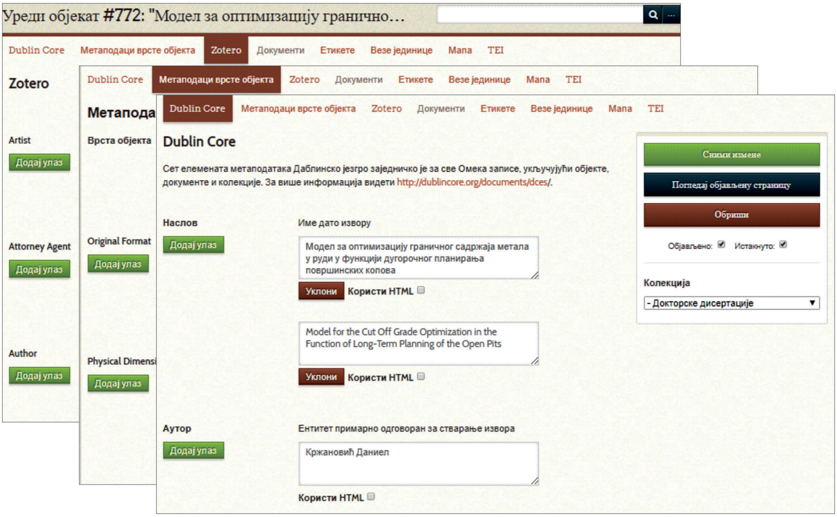
First, items are described by the metadata from the expanded Dublin core. The type of item is also described by the metadata, which are different depending on the type. For example, if the item is „*oral tradition*", then the metadata describing it are: the person conducting the interview, the person being interviewed, the location, the transcript, the duration of the interview etc. If the item is „*web page*", then only URL is entered, while for the item „*person*" corresponding metadata are: the date of birth, the place of birth, the date of death, profession, biography, bibliography etc. It is possible to describe an item through the metadata that will make it more visible in the scope of the Zotero platform (Figure 2).

Documents which are attached to the items can be imported either individually or on a large scale, either by the use of the DropBox plugin (Figure 3) or by the import from a file in the CSV format. Documents can be found in different formats, and some of the commonest are:

– for text: txt, css, csv, rtf, rtx, doc, docx, pdf, pps, ppt, pptx;
– for tables: xls, xlsx;
– for databases: mdb;
– for images: bmp, gif, jpeg, jpg, tiff, png;
– for video recordings: avi, divx, mpeg, mov, mp4;
– for audio recordings: mp3, mid, midi, wav, wma;
– for executable file: exe, zip.

Every item can receive tags[15], i.e. non-hierarchically ordered keywords or phrases which classify the contents so that it can be found more easily at

[15] https://omeka.org/codex/Managing_Tags_2.0



Уреди објекат #772: "Модел за оптимизацију гранично…

Dublin Core

Сет елемената метаподатака Даблинско језгро заједничко је за све Омека записе, укључујући објекте, документе и колекције. За више информација видети http://dublincore.org/documents/dces/.

Наслов

Име дато извору

Модел за оптимизацију граничног садржаја метала у руди у функцији дугорочног планирања површинских копова

Користи HTML □

Model for the Cut Off Grade Optimization in the Function of Long-Term Planning of the Open Pits

Користи HTML □

Аутор

Ентитет примарно одговоран за стварање извора

Кржановић Даниел

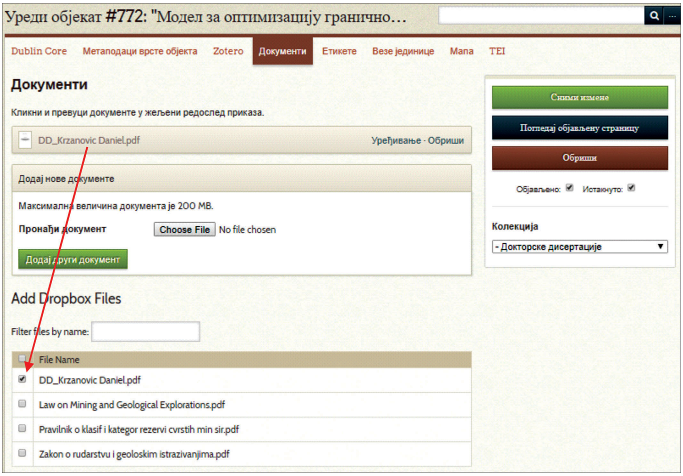
Користи HTML □

Објављено: ☑ Истакнуто: ☑

Колекција

- Докторске дисертације ▼

**Figure 2.** Describing objects with metadata



Уреди објекат #772: "Модел за оптимизацију гранично…

Документи

Кликни и превуци документе у жељени редослед приказа.

DD_Krzanovic Daniel.pdf

Додај нове документе

Максимална величина документа је 200 MB.

Пронађи документ No file chosen

Add Dropbox Files

Filter files by name:

| □ | File Name |
|---|---|
| ☑ | DD_Krzanovic Daniel.pdf |
| □ | Law on Mining and Geological Explorations.pdf |
| □ | Pravilnik o klasif i kategor rezervi cvrstih min sir.pdf |
| □ | Zakon o rudarstvu i geoloskim istrazivanjima.pdf |

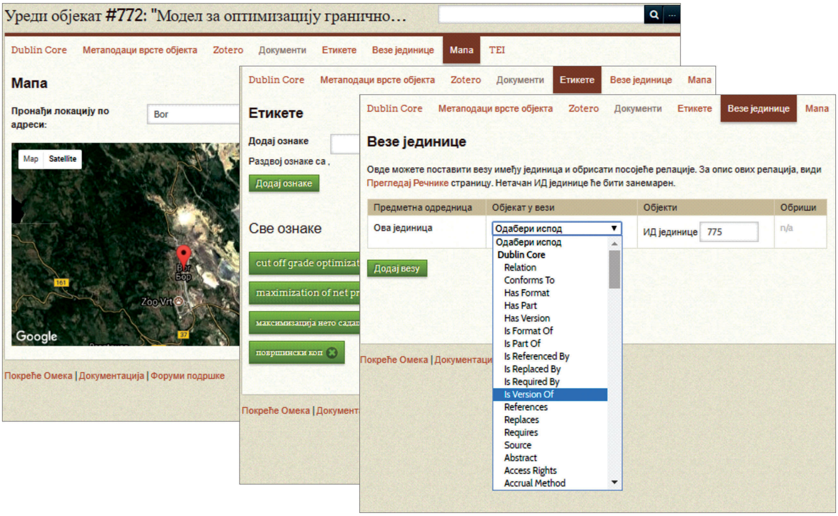
Објављено: ☑ Истакнуто: ☑

Колекција

- Докторске дисертације ▼

**Figure 3.** Panel for the import of documents

a later point. Figure 4 reproduces panels for import of geo-spatial entries, tagging and establishment of relations between items.



Уреди објекат #772: "Модел за оптимизацију гранично…

Dublin Core | Метаподаци врсте објекта | Zotero | Документи | Етикете | Везе јединице | Мапа | TEI

**Мапа**

Пронађи локацију по адреси: Bor

Покреће Омека | Документација | Форуми подршке

Dublin Core | Метаподаци врсте објекта | Zotero | Документи | Етикете | Везе јединице | Мапа

**Етикете**

Додај ознаке

Раздвој ознаке са ,

Додај ознаке

Све ознаке

cut off grade optimizat

maximization of net pr

максимизација нето садаш

површински коп

Покреће Омека | Документ

Dublin Core | Метаподаци врсте објекта | Zotero | Документи | Етикете | Везе јединице | Мапа

**Везе јединице**

Овде можете поставити везу имеђу јединица и обрисати посојеће релацијe. За опис ових релација, види Прегледај Речнике страницу. Нетачан ИД јединице ће бити занемарен.

| Предметна одредница | Објекат у вези | Објекти | Обриши |
|---|---|---|---|
| Ова јединица | Одабери испод | ИД јединице 775 | n/a |

Одабери испод
Dublin Core
Relation
Conforms To
Has Format
Has Part
Has Version
Is Format Of
Is Part Of
Is Referenced By
Is Replaced By
Is Required By
Is Version Of
References
Replaces
Requires
Source
Abstract
Access Rights
Accrual Method

Додај везу

Покреће Омека | Документаци

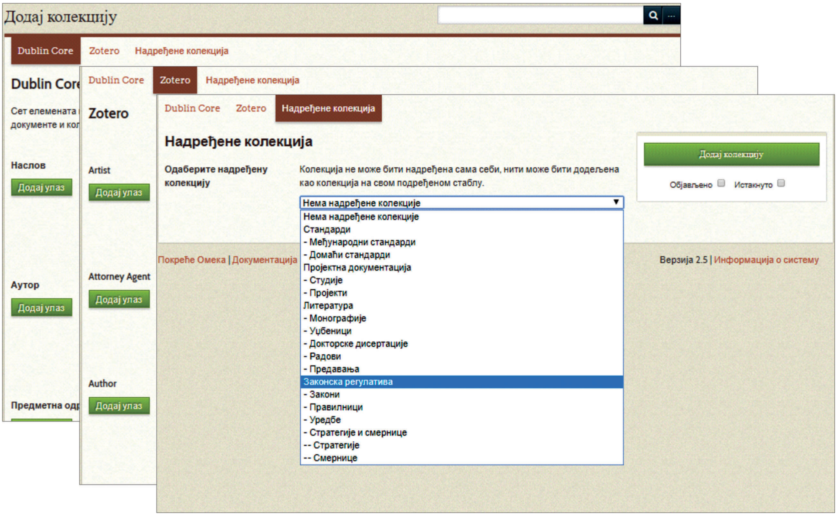
**Figure 4.** Import of geo-spatial entries, tagging and establishment of relations between items

In the *Map* panel added are the information on the location which is in connection with the item, and it is possible to search items with location in view.

It is possible to establish relation between two or more items. For example, The Law on Amendments and Supplements of The Law on Safety and Health is in the *isPartOf* relation with The Law on Safety and Health.

## 3.2 Collections

Collections represent groups of items that are organized in a way that makes it easier to search through them. It is meticulously described by metadata from the expanded set of the Dublin core elements and from the Zotero platform (Figure 5). They can be hierarchically structured, meaning that they can have defined subordinate and superordinate collections. The range of subordinate collections of any one collection spans from none to infinitely many, while there is always a single superordinate collection. For every collection it is possible to define visibility on the web page.



Одаберите надређену колекцију: Нема надређене колекције

Колекција не може бити надређена сама себи, нити може бити додељена као колекција на свом подређеном стаблу.

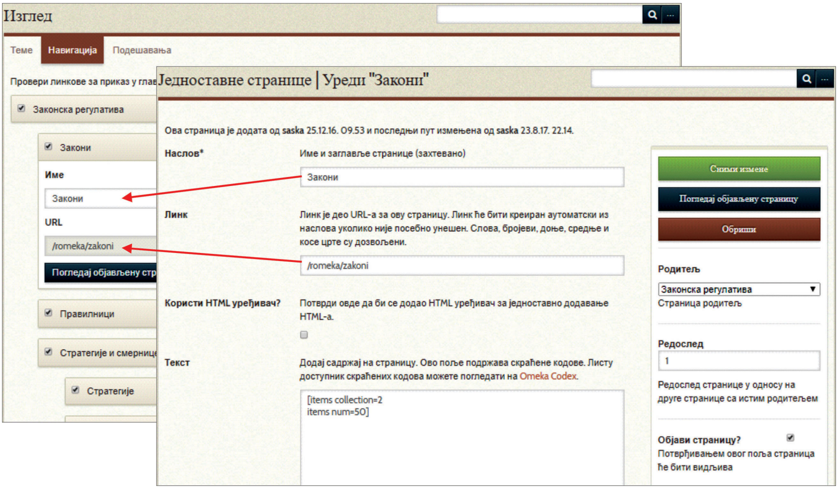
**Figure 5.** Panels for creating collections

## 3.3 Web pages

It is necessary to create web pages in order for items and collections to be visible on the web. Apart from the web page title it is necessary to define a relative path to the web page, as a part of the URL-address of that web page (Figure 6). In the part *Text* it is allowed to enter codes from the list of abridged codes, through which the appearance of the webpage itself is adapted. In the example (Figure 6) a code is given for presentation of the collection identified as ID 2 (*Laws*), with the number of items listed on one page limited to 50.

```
[items collection=2
 items num=50]
```

After the web page is created, in the part *Navigation*, on the panel for managing the appearance of the web site, the basic data (title and URL) are entered for every individual web page, while their order is determined through simple dragging of a field to the desired position. In this way, web pages are hierarchically structured and appear on the web site as drop-down menus.



Провери линкове за приказ у глав

☑ Законска регулатива

☑ Закони

Име

Закони

URL

/romeka/zakoni

☑ Правилници

☑ Стратегије и смерниц

☑ Стратегије

Једноставне странице | Уреди "Закони"

Ова страница је додата од saska 25.12.16. 09.53 и последњи пут измењена од saska 23.8.17. 22.14.

Наслов* Име и заглавље странице (захтевано)

Закони

Линк Линк је део URL-а за ову страницу. Линк ће бити креиран аутоматски из наслова уколико није посебно унешен. Слова, бројеви, доње, средње и косе црте су дозвољени.

/romeka/zakoni

Користи HTML уређивач? Потврди овде да би се додао HTML уређивач за једноставно додавање HTML-а.

□

Текст Додај садржај на страницу. Ово поље подржава скраћене кодове. Листу доступник скраћених кодова можете погледати на Omeka Codex.

```
[items collection=2
items num=50]
```

Родитељ

Законска регулатива ▼

Страница родитељ

Редослед

1

Редослед странице у односу на друге странице са истим родитељем

Објави страницу? ☑

Потврђивањем овог поља страница ће бити видљива

**Figure 6.** Creation of a web page

## 4 The digital library ROmeka@RGF

The digital library ROmeka@RGF[16] contains 209 texts primarily in the field of mining, but also in the fields of security, occupational safety, risk assessment, as closely related fields. The selection of documents of the digital library was based on accessible digital resources and for purposes of this digital library no further scanning was arranged for documents previously available in paper format only.

[16] The name ROmeka@RGF was coined by combining the abbreviated name of the Mining department (Rudarski odsek – RO) of the Faculty of Mining and Geology (Rudarsko-geološki fakultet – RGF) and the name of the Omeka web platform.

Items are categorized into 4 main, superordinate collections and 15 subordinate collections. The hierarchical structure of these collections is presented in the Table 1.

| **Collection** | **Subordinate Collection** | |
|---|---|---|
| Legislative | Laws | |
| | Rulebooks | |
| | Statutes | |
| | Strategies and directives | Strategies |
| | | Directives |
| Project documentation | Studies | |
| | Projects | |
| Literature | Monographies | |
| | Textbooks | |
| | Doctoral dissertations | |
| | Papers | |
| | Talks | |
| Standards | International standards | |
| | Domestic standards | |

**Table 1.** Hierarchical structure of collections

All items are meticulously described by all types of metadata that have been listed in the section 3.1, except for the metadata on use, which we plan to set in motion in the nearest future.

Visibility on web is provided for most of the digital items, but limitations are introduced for all items from collections *Project documentation* and *Standards* and for several items from the collection Literature. The reason for this limitation is related to the question of the publication rights, either because of confidentiality or because of author's rights.

All texts that are stored in the digital library will be used for research in terminology. In order to accomplish this, texts have been cleaned and parts in foreign languages have been removed together with tables, images, references and links. In order for these texts to be jointly processed, they have been merged into a single textual file with size of 39 MB, 6200 pages of text of A4 format. The processing of the text yielded 150.365 sentences and 2,719,086 (100,414 different) monomial lexical units. Around 1900 monomial terms, used particularly in the fields of mining, security, occupational safety and risk assessment are in preparation. After that, polynomial terms will be extracted, using methodology described in the paper (Stanković et al., 2012). We also plan to integrate the search through the corpus of mining texts with the Serbian Language Corpus SrpKor.

## 5 Search through the textual resources

Search for information in textual resources (comprising of a set of methods and techniques) takes into account both the resources themselves and/or the metadata describing those resources (Baeza-Yates and Ribeiro-Neto, 1999). Systems intended for information search are based on two concepts: *query* and *object*. Queries are formal requests for necessary information entered by the user into the search system. Objects are entities comprising requested information. User queries are matched with objects which are often stored in databases. Examples of the object data are documents and web pages.

Simple search for information in textual resources is implemented as string matching and does not take into account syntactic and semantic features of the requested word. These queries comprise either of one or of multiple words which can be connected through logical operators „and/or".

When it comes to searching through a digital library, the formulation of more complex queries is enabled through expanded regular expressions. Depending on resources one is searching through, answers can range from documents, to metadata or lists of web pages.

In the course of the search, three basic measures have to be taken into account: *recall*, *precision* and *ranking*. Recall expresses the level of completeness of answers received to a particular query and is represented as the ratio of the total number of relevant documents that were found and the total number of all relevant documents. Precision expresses the level of correctness of answers received to a particular query and is represented as the ratio of the total number of the total number of relevant documents that were found and the total number of answers that were found. Recall indicates the level of comprehensiveness of the system in the course of the search for relevant information.

Problems related to searching through the textual resources can be classified in two categories:

- general problems which are not related to language and

- problems that are specific to a particular language or a group of languages.

The problem encountered in the course of searching through the texts written in Serbian are different code schemes, as well as the existence of two alphabets (the Cyrillic and the Latin alphabet). This is the case with the digital library ROmeka@RGF because documents have been imported in their original form and in both scripts. Search without expansion of the query allows the contents to be searched through on the bases of one alphabet alone, while the expansion of the query automatically involves both alphabets. Searching documents written in Serbian is a complex process because of the very rich morphological system of the Serbian language. Most often, queries are looking for words in their canonical shape (Nominative singular for nouns, infinitive for verbs) (Lazić et al., 2016). However, documents can contain any flection of a flectional word. This problem becomes even more complex if compound words and synonyms are also taken into account (Stanković, 2009).

Elements of a digital library that can be included in the process of searching are: metadata, documents, tags, reports, expositions, web pages. Originally, the search is performed through: key words, bull operators and complete matching. The digital library ROmeka@RGF is upgraded through implementation of expanded queries. Web services (Stanković et al., 2012) and morphological electronic dictionaries for Serbian (Krstev et al., 2008; Stanković et al., 2016) have been used:

- for a morphological expansion of a query::
  http://hlt.rgf.bg.ac.rs/vebran/api/delafs/ključna_reč
- for a semantic and morphological expansion of a query:
  http://hlt.rgf.bg.ac.rs/vebran/api/sinonimi/ključna_reč

Searching with and without morphological and semantic-morphological expansions of queries will be presented on the example of a search for a lexeme *хомогенизација*.

An unexpanded query is taking only the Nominative singular into account:

*хомогенизација*

A query with a morphological expansion allows the inclusion of all flectional forms of the requested lexeme, on both alphabets (Latin and Cyrillic). Therefore, in this case, the following flections are searched through:

*homogenizacija, homogenizacijama, homogenizacije, homogenizaciji, homogenizacijo, homogenizacijom, homogenizaciju, хомогенизација, хомогенизацијама, хомогенизације, хомогенизацији, хомогенизацијо, хомогенизацијом, хомогенизацију.*

A query containing both semantic and morphological expansions allows the inclusion of synonyms and lexical metonymies, as well as of their flectional forms:

*homogenizacija, homogenizacijama, homogenizacije, homogenizaciji, homogenizacijo, homogenizacijom, homogenizaciju, ujednačavanje kvaliteta uglja, homogenizacija kvaliteta uglja, homogenizacijama kvaliteta uglja, homogenizacije kvaliteta uglja, homogenizaciji kvaliteta uglja, homogenizacijo kvaliteta uglja, homogenizacijom kvaliteta uglja, homogenizaciju kvaliteta uglja, homogenizacija uglja, homogenizacijama uglja, homogenizacije uglja, homogenizaciji uglja, homogenizacijo uglja, homogenizacijom uglja, homogenizaciju uglja, upravljanja kvalitetom uglja, upravljanje kvalitetom uglja, upravljanjem kvalitetom uglja, upravljanjima kvalitetom uglja, upravljanju kvalitetom uglja, upravljanja kvalitetom, upravljanje kvalitetom, upravljanjem kvalitetom, upravljanjima kvalitetom, upravljanju kvalitetom, хомогенизација, хомогенизацијама, хомогенизације, хомогенизацији, хомогенизацијо, хомогенизацијом, хомогенизацију, уједначавање квалитета угља, хомогенизација квалитета угља, хомогенизацијама квалитета угља, хомогенизације квалитета угља, хомогенизацији квалитета угља, хомогенизацијо квалитета угља, хомогенизацијом квалитета угља, хомогенизацију квалитета угља, хомогенизација угља, хомогенизацијама угља, хомогенизације угља, хомогенизацији угља, хомогенизацијо угља, хомогенизацијом угља, хомогенизацију угља, управљања квалитетом угља, управљање квалитетом угља, управљањем квалитетом угља, управљањима квалитетом угља, управљању квалитетом угља, управљања квалитетом, управљање квалитетом, управљањем квалитетом, управљањима квалитетом, управљању квалитетом.*

Results of the search for monomial and polynomial terms (*хомогенизација*, *управљање квалитетом угља*, *површински коп*, *роторни багер*, *експлоатација* и *рударство*) in the digital library Romeka@RGF, figure in the Table 2.

The diagram (Figure 7) illustrates the results of the search for monomial and polynomial terms in the digital library Romeka@RGF based on the data

| **Type of query** | **homogenizacija** | **upravljanje kvalitetom uglja** | **površinski kop** | **rotorni bager** | **eksploatacija** | **rudarstvo** |
|---|---|---|---|---|---|---|
| Without expansion for the Cyrillic script | 6 | 0 | 21 | 0 | 13 | 15 |
| Without expansion for the Latin script | 9 | 14 | 49 | 33 | 50 | 23 |
| Morphological expansion | 33 | 22 | 106 | 42 | 120 | 106 |
| Semantic and morphological expansion | 47 | 47 | 118 | 43 | 123 | 120 |

**Table 2.** Results of the search for monomial and polynomial terms

given in the Table 2. It is worthwhile observing that queries with morphological and semantical expansions yield more results than unexpanded queries.

## 6 TEI (Text Encoding Initiative)

In order to extract information from digital items which are a part of the digital library ROmeka@RGF, we have decided to store items which are in the textual format in the XML format as well in accordance with the TEI[17] guidelines. We have made this decision because TEI is a *de facto* standard for annotation of arbitrary document types, including legal texts and project documentation texts. Annotation in accordance with TEI guidelines should enable the connection between parts of a project documentation text which refer to law articles and legal regulations.

In the course of annotation in accordance with the TEI P5 guidelines (TEI-Consortium, 2017), tags <div1>, <div2>, <div3> and <div4> were used in accordance with the hierarchy between a law, a chapter, a section and an article (Васиљевић, 2015). The tag <p> was used as equivalent to point, subpoint and indent of a law.

Tags with different attribute values were used for tagging of titles and subtitles. In accordance with the connection prerequisites and mindful of

[17] http://www.tei-c.org/

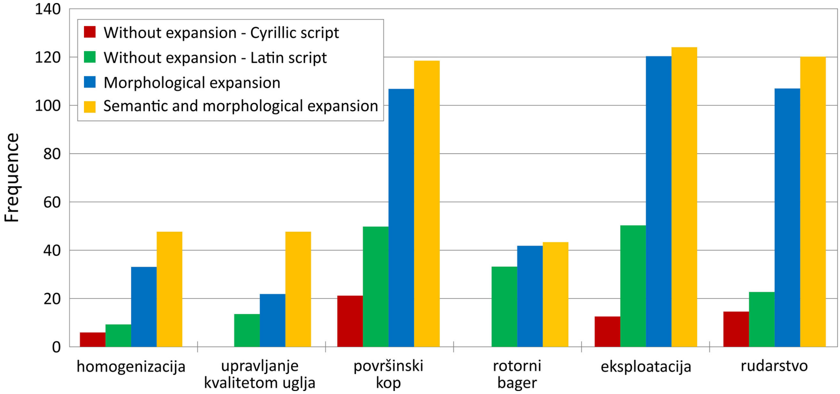


**Figure 7.** Results of the search for monomial and polynomial terms in the digital library ROmeka@RGF

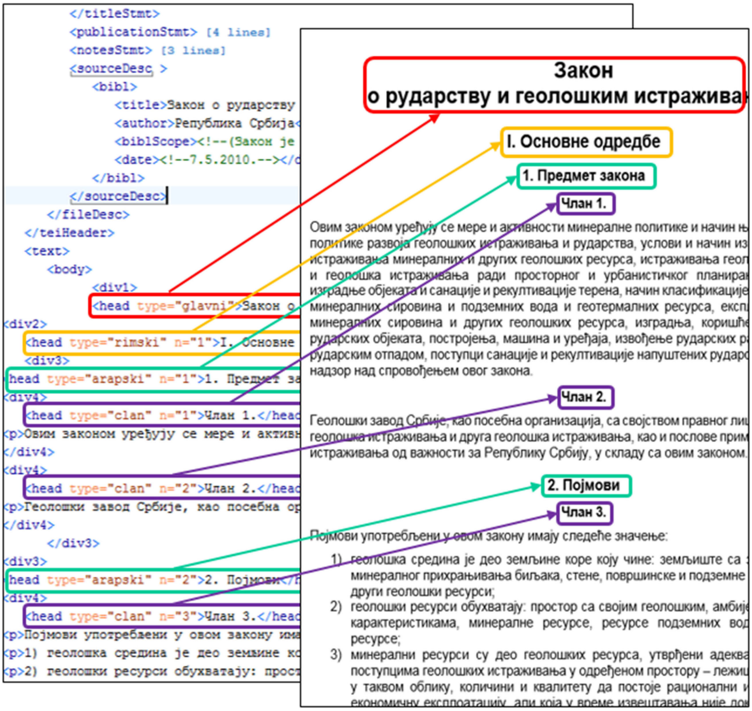
the connections between the law articles, we used the tag <head> to mark the title, chapter, section and article of a law. Within this tag, attributes *type* and *n* were used. The values of the attribute *type* are crucial for the distinction between different parts of a law. There are four possible values of this attribute: *main*, *chapter*, *section* or *article*. The values of the attribute n are the ordinal numbers of the current elements (i.e. of a chapter, section or article).

Figure 8 represents a part of the document *Law on Mining and Geological Explorations* tagged in accordance with the TEI guidelines in the way discussed previously.

Mindful of the mining project presentations, we made an effort to conserve all elements – most notably the tables – as important sources of information, while we were using TEI. Figure 9 illustrates a tabular presentation within the TEI version of the project documentation.

## 7 Conclusion

In this paper both advantages and disadvantages were shown of Omeka as a platform for development of this type of library. Additionally, the extent to which the application of morphological dictionaries affects the quality of a search in the course of a morphological and semantical expansion of a query



```
            </titleStmt>
            <publicationStmt> [4 lines]
            <notesStmt> [3 lines]
            <sourceDesc >
               <bibl>
                  <title>Закон о рударству
                  <author>Република Србија<
                  <biblScope><!--(Закон је
                  <date><!--7.5.2010.--></d
               </bibl>
            </sourceDesc>
         </fileDesc>
      </teiHeader>
      <text>
         <body>
               <div1>
               <head type="glavni">Закон о
<div2>
   <head type="rimski" n="1">I. Основне
   <div3>
<head type="arapski" n="1">1. Предмет за
<div4>
   <head type="clan" n="1">Члан 1.</head
<p>Овим законом уређују се мере и активн
</div4>
<div4>
   <head type="clan" n="2">Члан 2.</head
<p>Геолошки завод Србије, као посебна ор
</div4>
         </div3>
<div3>
<head type="arapski" n="2">2. Појмови</h
<div4>
   <head type="clan" n="3">Члан 3.</head
<p>Појмови употребљени у овом закону има
<p>1) геолошка средина је део земљине ко
<p>2) геолошки ресурси обухватају: прост
```

Закон
о рударству и геолошким истражива

I. Основне одредбе

1. Предмет закона

Члан 1.

Овим законом уређују се мере и активности минералне политике и начин њ политике развоја геолошких истраживања и рударства, услови и начин из истраживања минералних и других геолошких ресурса, истраживања геол и геолошка истраживања ради просторног и урбанистичког планира изградње објеката и санације и рекултивације терена, начин класификације минералних сировина и подземних вода и геотермалних ресурса, експ минералних сировина и других геолошких ресурса, изградња, коришће рударских објеката, постројења, машина и уређаја, извођење рударских р рударским отпадом, поступци санације и рекултивације напуштених рударс надзор над спровођењем овог закона.

Члан 2.

Геолошки завод Србије, као посебна организација, са својством правног лиц геолошка истраживања и друга геолошка истраживања, као и послове прим истраживања од важности за Републику Србију, у складу са овим законом.

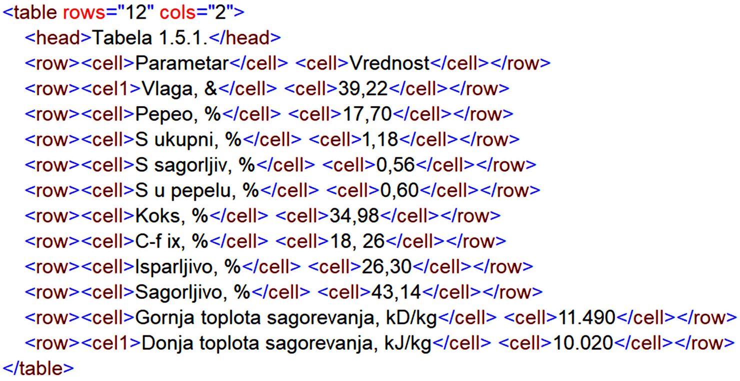
2. Појмови

Члан 3.

Појмови употребљени у овом закону имају следеће значење:

1) геолошка средина је део земљине коре коју чине: земљиште са минералног прихрањивања биљака, стене, површинске и подземне други геолошки ресурси;
2) геолошки ресурси обухватају: простор са својим геолошким, амбије карактеристикама, минералне ресурсе, ресурсе подземних вод ресурсе;
3) минерални ресурси су део геолошких ресурса, утврђени адеква поступцима геолошких истраживања у одређеном простору – лежиш у таквом облику, количини и квалитету да постоје рационални и економичну експлоатацију, али која у време извештавања није лок

**Figure 8.** A part of the document tagged in accordance with the TEI PS guidelines



```
<table rows="12" cols="2">
   <head>Tabela 1.5.1.</head>
   <row><cell>Parametar</cell> <cell>Vrednost</cell></row>
   <row><cel1>Vlaga, &</cell> <cell>39,22</cell></row>
   <row><cell>Pepeo, %</cell> <cell>17,70</cell></row>
   <row><cell>S ukupni, %</cell> <cell>1,18</cell></row>
   <row><cell>S sagorljiv, %</cell> <cell>0,56</cell></row>
   <row><cell>S u pepelu, %</cell> <cell>0,60</cell></row>
   <row><cell>Koks, %</cell> <cell>34,98</cell></row>
   <row><cell>C-f ix, %</cell> <cell>18, 26</cell></row>
   <row><cell>Isparljivo, %</cell> <cell>26,30</cell></row>
   <row><cell>Sagorljivo, %</cell> <cell>43,14</cell></row>
   <row><cell>Gornja toplota sagorevanja, kD/kg</cell> <cell>11.490</cell></row>
   <row><cel1>Donja toplota sagorevanja, kJ/kg</cell> <cell>10.020</cell></row>
</table>
```

**Figure 9.** TEI P5 illustration of the tabular presentation

has also been shown. The conclusion is that combining Omeka and morphological dictionaries is likely to assist better organisation and searchability of items stored in the digital library ROmeka@RGF.

Since it is important that both the collections and the search tools are regularly upgraded, we plan to follow the developments of further plugins and technologies which could assists the connection of the digital library ROmekaRGF with other sources of similar information. Furthermore, we plan to supplement the digital library with various new digital items. We also plan to extract information from digital items themselves.

## Acknowledgements

The research presented in this paper has been financially supported by the Ministry of education, science and technological development of the Republic of Serbia through the projects TP33039 ("Improvement of Lignite Opencast Mining Technology in Order to Increase Energy Efficiency and Occupational Safety").

## References

Arms, William Y. *Digital libraries.* Cambridge, Massachusetts, USA: M.I.T. Press, 2000. http://www.cs.cornell.edu/wya/diglib/ms1999/

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. *Modern information retrieval*, Vol. 463, New York, USA: ACM Press, 1999. http://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf

Hodge, Gail. *Metadata made simpler.* Niso Press, 2001.

Krstev, Cvetana, Ranka Stanković, Dusko Vitas and Ivan Obradović. "The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines". In *LREC*, Paris, France: European Language Resources Association (ELRA), 219–224, 2008. http://lrec-conf.org/proceedings/lrec2008/pdf/67_paper.pdf

Kucsma, Jason, Kevin Reiss and Angela Sidman. "Using Omeka to build digital collections: The METRO case study". *D-Lib magazine* Vol. 16, no. 3/4 (2010). http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/march10/kucsma/03kucsma.html

Lazić, Biljana, Danica Seničić, Aleksandra Tomašević and Bojan Zlatić. "Terminological and Lexical Resources Used to Provide Open Multilingual Educational Resources". Belgrade, Serbia, 2016. http://www.baektel.eu/documents/conferences/eLearning_2016_BL_DS_AT_BZ.pdf

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić. “A tool for enhanced search of multilingual digital libraries of e-journals”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 1710–1717, 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf

Stanković, Ranka, Cvetana Krstev, Biljana Lazić Ivan Obradović and Aleksandra Trtovac. “Rule-based Automatic Multi-Word Term Extraction and Lemmatization”. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC*, 507–514, 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1033_Paper.pdf

Stanković, Ranka M. “Modeli ekspanzije upita nad tekstuelnim resursima”. Doktorska disertacija. Univerzitet u Beogradu, Matematički fakultet, 2009.

TEI-Consortium. “TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0.”, 2017. Accessed August 23, 2017, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Васиљевић, Небојша. “Аутоматска обрада правних текстова на српском језику”. Докторска дисертација. Универзитет у Београду, Филолошки факултет, 2015. https://fedorabg.bg.ac.rs/fedora/get/o:10687/bdef:Content/get

Тртовац, Александра С. “Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама”. Докторска дисертација. Универзитет у Београду, Филолошки факултет, 2016. https://fedorabg.bg.ac.rs/fedora/get/o:12605/bdef:Content/get