Agris on-line Papers in Economics and Informatics

Volume IV Number 4, 2012

Using Metadata Description for Agriculture and Aquaculture Papers

P. Šimek, J. Vaněk, V. Očenášek, M. Stočes, T. Vogeltanzová

Faculty of Economics and Management, Czech University of Life Science in Prague, Czech Republic

Anotace

Článek pojednává o nejpoužívanějších metadatových formátech a tezaurech, které jsou vhodné pro popis vědeckých, výzkumných a odborných článků z oblasti zemědělství, potravinářství, vodohospodářství, životního prostředí a venkova. Jedná se o Dublin Core (DC), Metadata Object Description Schema (MODS), Virtual Open Access Agriculture and Aquaculture Repository Metadata Applicaton Profile (VOA3R AP) a AGROVOC. Na základě analýzy metadatových formátů v souladu s životním cyklem vědeckovýzkumného nebo odborného článku autoři doporučují, že každý takový článek by měl být bezprostředně po jeho publikování popsán metadaty, která efektivně charakterizují jeho obsah a vlastnosti. Jedním z nejvhodnějších metadatových formátu je VOA3R AP, vycházející částečně z DC, v kombinaci s tezaurem AGROVOC. Tím bude dosaženo efektivního popisu, zpřístupnění a automatické výměny dat mezi lokálními a centrálními repozitáři.

Klíčová slova

Metadata, element, článek, popis, tezaurus, AGROVOC, Dublin Core, VOA3R AP.

Abstract

The paper deals with the most used metadata formats and thesauri suitable for describing scientific and research papers in the domains agriculture, food industry, aquaculture, environment and rural areas. These include the Dublin Core (DC), Metadata Object Description Schema (MODS), Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile (VOA3R AP) and the AGROVOC thesaurus. Having analyzed the metadata formats and research paper lifecycle, the authors would recommend that each paper should entail metadata description as soon as it is published. The metadata are to describe the content and properties of the paper. One of the most suitable metadata formats is the VOA3R AP that is partially patterned on the DC and combined with the AGROVOC thesaurus. As a result, an effective description, availability and automatic data exchange between and among local and central repositories should be attained.

The knowledge and data presented in the present paper were obtained as a result of the following research programs and grant schemes: the Grant No. 20121044 of the Internal Grant Agency titled „Using Automatic Metadata Generation for Research Papers“, the Grant agreement No. 250525 funded by the European Commission corresponding to the VOA3R Project (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment), http://voa3r.eu and the Research Program titled „Economy of the Czech Agriculture Resources and their Efficient Use within the Framework of the Multifunctional Agrifood Systems“ of the Czech Ministry of Education, Youth and Sport number VZ MSM 6046070906.

Key words

Metadata, element, paper, description, thesaurus, AGROVOC, Dublin Core, VOA3R AP.

Introduction

Nowadays information and knowledge society is characterized by a growing number of information resources in all spheres of human activity. Therefore, the need for a systematic metadata description of the information content and properties has been increasing together with the need for making relevant metadata available. Metadata can be used to describe all electronic objects or database systems.

It means we can provide a description of a book, a picture, a piece of music, SW, a website or a research document. Metadata should describe objects in an unambiguous and appropriate manner (however, in some cases, it is not possible) (Ardo, 2010). Global metadata use is driven by technical or working teams and groups in industry, at universities, research bodies and institutes etc. Agriculture is a good example of application development and integration of the systems requiring structured data (Santos, 2012).

Aggregating metadata from various resources raises practical problems such as incompatibility of different metadata application profiles (AP) or metadata quality (Protonotarios, 2011). Local repositories containing scientific papers on agriculture, food industry, aquaculture, environment and rural development face more or less the same problems. In order to fulfil their mission and maintain a high quality standard, these local repositories have to seek and implement innovations in compliance with the latest technologies and information resources development so that their content can be unequivocally identified and meta-described with a view to content distribution.

Thanks to a dynamic computer and information science development, the field of ontology has been recently gaining popularity in research. In agriculture, the Food and Agriculture Organization of the United Nations (FAO) launched the Agricultural Ontology Service (AOS) in 2001. The AOS strives to serve as a reference initiative in the domain of agriculture (Wei, 2012). Ontology in agriculture should provide both scientists and farmers with the required level of information. The AGROVOC thesaurus serves as a starting point (basic vocabulary) for the creation of domain specific ontologies (Bansal, 2011).

Material and methods

There exist a lot of metadata formats and domain-specific thesauri describing various objects by specific elements. These have been developed within the framework of research projects, by communities or standardising bodies themselves. The following metadata formats and thesauri are the most used: Dublin Core (DC), Metadata Object Description Schema (MODS), Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile (VOA3R AP) and AGROVOC.

Dublin Core

The Dublin Core is a metadata format that was primarily created for the sake of simple and general web resources description by authors themselves. The original set of 15 metadata elements was extended and refined within the Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH) (Open Archive Initiative, 2008). It was then ratified as IETF RFC 5013, ANSI/NISO Z39.85-2007 standard and ISO 15836:2009 standard. The DC elements describe the most important data and properties of the document (Dublin Core Metadata Initiative, 2010).

Term name: Contributor	
Label:	Contributor
Definition:	An entity responsible for making contributions to the resource.
Comment:	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Term name: Coverage	
Label:	Coverage
Definition:	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Comment:	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Term name: Creator	
Label:	Creator
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.

Figure 1: Overview of 15 DC elements metadata set (source: DCMI).

Term name: Date	
Label:	Date
Definition:	A point or period of time associated with an event in the lifecycle of the resource.
Comment:	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
Term name: Description	
Label:	Description
Definition:	An account of the resource.
Comment:	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.
Term name: Format	
Label:	Format
Definition:	The file format, physical medium, or dimensions of the resource.
Comment:	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].
Term name: Identifier	
Label:	Identifier
Definition:	An unambiguous reference to the resource within a given context.
Comment:	Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
Term name: Language	
Label:	Language
Definition:	A language of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646].
Term name: Publisher	
Label:	Publisher
Definition:	An entity responsible for making the resource available.
Comment:	Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Term name: Relation	
Label:	Relation
Definition:	A related resource.
Comment:	Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Term name: Rights	
Label:	Rights
Definition:	Information about rights held in and over the resource.
Comment:	Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.
Term name: Source	
Label:	Source
Definition:	A related resource from which the described resource is derived.
Comment:	The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Term name: Subject	
Label:	Subject
Definition:	The topic of the resource.
Comment:	Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary.

Figure 1: Overview of 15 DC elements metadata set (source: DCMI) - continuation.

Term name: Title	
Label:	Title
Definition:	A name given to the resource.
Comment:	Typically, a Title will be a name by which the resource is formally known.
Term name: Type	
Label:	Type
Definition:	The nature or genre of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.

Figure 1: Overview of 15 DC elements metadata set (source: DCMI) - end.

Apart from the original 15-element metadata set, a few more elements (also called qualifiers) can be employed. These include:

- Accrual Method,
- Accrual Periodicity,
- Accrual Policy,
- Audience,
- Mediator,
- Instructional Method,
- Provenance
- Rights Holder.

Since 2000, the DC community has been aiming at the Application Profiles (AP) so that metadata records could employ the DC together with other specialized vocabularies. At the same time, the World Wide Web consortium (W3C) has been finalizing the generic metadata data model - Resource Description Framework (RDF). The DC has become one of the most spread and popular data vocabularies used with the RDF.

Metadata Object Description Schema

Metadata Object Description Schema (MODS) is a metadata schema developed and maintained by the specialists of the Library of Congress and MARC Standards Office. It entails a bibliographic element set that is designed primarily for library applications but may be also used for other different purposes.

The schema creation was incited by digital libraries and other communities that required a rich XML description, maintaining complex digital objects and integrating digital libraries metadata databases using MARC with different schemas. Firstly, the schema was intended as a kind of MARC subset using just different element names. In the end, an independent schema was born, carrying key elements from the MARC record but not entailing all MARC fields. On the other hand, it comprises some new elements.

MODS 3.4 entails 20 top level elements with optional attributes. These include (The Library of Congress):

titleInfo	note
name	subject
typeOfResource	classification
genre	relatedItem
originInfo	identifier
language	location
physicalDescription	accessCondition
abstract	part
tableOfContents	extension
targetAudience	recordInfo

Source: the Library of Congress)
Figure 2: Overview of 20 MODS 3.4 elements.

Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile

Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile (VOA3R Metadata AP) format is a European research project, based partially on the DC, striving to improve the description, spread, sharing and application of agriculture and aquaculture open access research results (N. Diamantopoulos, 2011). It comprises 31 elements in 9 categories. The VOA3R AP elements can be compulsory, strongly recommended, recommended or optional.

VOA3R platform is represented by XML and own terminological thesauri created in line with the recent semantic standards. One of the main VOA3R assets is a direct determination of abundant data

i.e. bibliographic citations. It also allows users to access a complete list of personal details (e.g. author-related) taking the vCard form.

AGROVOC

AGROVOC is a thesaurus that contains more than 40,000 entries in 22 languages and covers topics related to food, nutrition, agriculture, fishery, forestry, environment and other related domains. The AGROVOC is maintained by a global community of editors comprising librarians, terminologists, information managers and software developers. The AGROVOC is expressed in Simple Knowledge Organization System (SKOS) and published as Linked Data. The whole thesaurus is expressed in the concept system SKOS which is a data model for structured controlled vocabularies. The AGROVOC thesaurus schema employs three levels of representation:

- concepts represent abstract meanings and are often identified by URIs, e.g. corn as a cereal is identified by „Concept12332“,
- terms are language-specific forms e.g. corn, maïs, 玉米, or maize
- terms integrate special variants such as spelling variants, singular or plural form e.g. hen, hens, cow or cows.

This is how the abstract concepts/terms and the concrete meanings are related. The AGROVOC is therefore suitable for the description of research papers, information or news in the agrarian sector - Agricultural Information Management Standards.

Results and discussion

In the domain of research and science, the need for both metadata description and metadata access has been constantly growing. One of the main DC advantages is that it allows digital documents authors to make a semantic description of their documents, websites and other digital objects without being specialist in the field and without mastering other purpose-related methods and standards.

The MODS metadata format is suitable for describing publications and library repositories categorisation. The MODS format has the following advantages over other schemes: it is compatible with other tools, especially with the library system MARC 21 and the Dublin Core. It also allows the conversion of these tools to MODS, which removes potential barriers. It also eliminates (by means of a suitable combination) the inconveniences of MARC 21 (excessive complexity, lack of syntax as numeric tags are used) and at the same time extends the Dublin Core (it entails a range of basic elements and a number of sub elements).

VOA3R Metadata Application Profile with an integrated AGROVOC thesaurus is one of the most suitable and viable metadata formats for the paper description in agriculture, aquaculture, food industry, environment and rural development.

Title Info

Title

The Title element should clearly represent the paper as an electronic resource. This element is compulsory. Obviously, it is a name given to papers, a name by which the resource is formally known. In this element, the name should be introduced in the language of origin, including all the translations. If the paper title includes more languages at the same time, there is an independent element with the language marked introduced. The subtitle should be also saved in the Title element while using a gap hyphen gap format, i.e. „ - „ between the title and subtitle.

Alternative title

The Alternative Title element should be used only in case the paper is also known under a different name, including abbreviations or acronyms. However, this element should not be used for translations of the title or subtitles. This element is optional. Nevertheless, when the content exists, it should be considered compulsory.

Responsible body

The Responsible body category comprises all elements containing information about persons that exercise their influence on the paper content during any phase of its lifecycle. These include the creator (author), contributor and publisher.

Creator

The present element describes the author of paper‘s intellectual content. Therefore, it can be a person, an institution or a service. The Creator element should include author‘s name, or as the case may be organization‘s name and/or author‘s URI and/or a reference to the resource describing the author. In case of concrete persons, we always start with their surname, comma and then the first name (full or initials) or other names, e.g. „Šimek, Pavel“ or „Šimek, P.“.

In case there are more authors, the order of elements should reflect their formal hierarchy, i.e. the first author is considered as the main one and the others as co-authors.

VOA3R AP regards the Creator element as strongly recommended and as compulsory in case of research papers.

Contributor

The Contributor element characterizes the persons, institutions or services that contributed to the paper content. In case there are more contributors, the element is repeated and can include e.g. students' tutors, readers, reviewers etc. This element is recommended for research papers.

Publisher

The Publisher element saves information on the person, service or institution that provides access to the paper, respectively published the paper. The element is strongly recommended in order to identify the publishing entity (both commercial and non-commercial), not to identify the author's institution. In case of Publisher research papers, the element should be compulsory.

Physical characteristics

Physical characteristics of a published paper should be described with a view to the date of publishing, identifier, languages and paper resource format.

Date

The Date element (compulsory) entails time information related to paper publishing. When this entry is not available, a date when the paper was made accessible should be indicated. The format is to be in line with the W3C Date Time Format (W3CDTM):

- year format YYYY, e.g. 2012,
- month and year format YYYY-MM, e.g. 2012-09,
- a complete date format YYYY-MM-DD, e.g. 2012-09-30,
- a complete date, including the hour and minute format YYYY-MM-DDThh:mmTZD, e.g. 2012-09-30T08:30+1:00,
- a complete date, including the hour, minute and second format YYYY-MM-DDThh:mm:ssTZD, e.g. 2012-09-30T08:30:25+1:00 a
- a complete timestamp format YYYY-MM-DDThh:mm:ss.sTZD, e.g. 2012-09-30T08:30:25.45+1:00

(World Wide Web Consortium, 1997)

Language

The Language element is compulsory and saves information on all languages used in the paper. If there are more languages used, this element is repeated for each and every language. The language is expressed according to ISO639-2, e.g. eng for English or cze for the Czech language.

Identifier

The Identifier element saves a string, an unambiguous reference to the paper resource. For indentifying the resource, formal identification systems can be used, i.e. Uniform Resource Identifier (URI), including Uniform Resource Locater (URL), Digital Object Identifier (DOI) an International Standard Book Number (ISBN). This element is strongly recommended. However, for a research paper, it should be compulsory.

Format

The Format element identifies information on the medium (file format) used to make the paper content available. If a paper is available in multiple formats, a separate Format element is used. The element content is encoded according to Internet Media Types MIME, e.g. application/pdf format for PDF, text/html for an HTML format, text/xml for XML etc. This element is strongly recommended and it should be considered compulsory in case of open access or after publishing the paper.

Location

The resource location is important in order to retrieve the paper for the sake of information exchange.

isShownBy

This element includes an unambiguous URL referring to a paper resource and enabling the user to read it or play it. This element should be independent of the fulltext version location.

isShownAt

The isShownAt element is used to save the unambiguous URL referring to the fulltext paper in a concrete format. Both elements (isShownBy and isShownAt) are strongly recommended and at least one of them should be considered compulsory in case of open access publishing.

The differences between the two elements are the following:

isShownBy: http://www.domena.cz/archiv/2012/01

isShownAt: http://www.domena.cz/archiv/2012/01/clanekXY.pdf

Subject

The Subject category comprises only one element of the same name. This category deals with information connected to the topic of the resource. Typically, the subject is represented using paper topic, classification, keywords or key phrases.

Subject

The Subject element serves to describe the paper topic by means of classification codes, keywords or key phrases. Keywords that are not created or controlled by thesauri are separated using a semicolon character. While writing a research paper related to the domains of agriculture, food industry, aquaculture, environment and rural areas, the AGROVOC thesaurus should be used for keywords. Each keyword from the AGROVOC thesaurus has its own Subject element that is compulsory for papers. URIs to concrete AGROVOC identifiers or keywords corresponding to the AGROVOC thesaurus are inserted in this element.

Description of content

The Description element characterizes the description of two main kinds. These are:

- description related to the content, i.e. description, abstract, references
- description related to the nature or genre of content resources.

Description

The Description element entails a complementary paper description by means of a text describing the paper resource or content or a link to a graphic representation, audiofiles etc. This element is recommended.

Abstract

This element includes the paper abstract and should not be confused with the Description element. From the point of view of a research paper, this element is compulsory.

Type

The Type element is related to the nature/genre of the resource referring to the paper. The present element is compulsory and it can take the following forms:

- Publication Collection
  - Book
  - Journal
  - Conference proceedings
  - Magazine
- Publication Item
  - Book section
  - Journal contribution
    - Article
    - Review
    - Editorial
    - Letter
    - Note
  - Conference contribution
    - Paper
    - Poster
    - Presentation
  - Magazine article
  - Thesis
    - Bachelor thesis
    - Master thesis
    - Doctoral thesis
  - Research report
  - Standard
- Resource
  - Learning resource
  - Multimedia resource
  - Data set
- Event
  - Conference
  - Project
- Other

In case of a paper published in a scientific journal, we deal with the Article Type (Publication item – Journal contribution – Article).

Bibliographic citation

This element is used to encode information concerning bibliographic citations. The recommended best practice is to use the BibTex. Nevertheless, the APA or OpenURL ContextObject can be also used following the Guidelines for Encoding Bibliographic Citation Information

in Dublin Core Metadata DCMI. The VOA3R AP rates the Bibliographic citation element to the recommended ones. However, in case of a research paper, this element is to be compulsory.

Rights

The Rights category comprises information about intellectual property rights held in and over the resource, including the resource use and access.

Access rights

This element contains information concerning access rights after the paper was published (open access, closed access, paid or restricted access). In order to describe the access rights, the recommended practice is to use the Eprints AccessRights Vocabulary Encoding Scheme. After the paper was published, the Access Rights element should be compulsory, taking the following effects:

- Open Access (the paper/article is freely available on the Internet)
- Restricted Access (the paper/article is available on the Internet but the access to fulltext version is restricted or controlled)
- Closed Access (the paper/article is available on the Internet but the access to fulltext version is restricted or controlled)

License

The License element contains detailed information concerning the terms of use and distribution. The Creative Commons license is considered to be the best practice for the purpose given. It entails the following CC licenses and their combinations:

- Attribution
- Share Alike
- No Derivatives
- Non-Commercial
- Non-Commercial Share Alike
- Non-Commercial No Derivatives

Rights

The Rights element includes the name of a copyright holder, e.g. name of a publisher. The copyright statement is a legal measure concerning the terms of use. While publishing a research paper, this element can be considered compulsory.

Status

This category entails information on article properties related to the review and publishing process.

Review Status

This element – compulsory for research papers - informs users on the review process using the following statuses:

- Non-reviewed (the paper has not been reviewed)
- Peer Reviewed (the paper went through the review process)
  - Accepted (paper accepted for publication in the review process)
  - Rejected (paper rejected in the review process)
- Community Reviewed (paper reviewed by a community of practice)
  - Commented (paper commented by a community of practice)
  - Rated (paper was community rated)

Publication status

The present element gives information on the publication status and should be considered compulsory for publishers. It entails the following statuses:

- Working Draft
- Final (final version)
- Submitted (pre-print version)
- Published

Relation

Data saved in the Relation category are important for resource location and content information retrieval. Properties related to resource location are represented by the isShownBy and isShownAt elements (see above). Other metadata elements regarding various relations are incorporated in the Relation category.

Relation

This element relates articles to other resources by means of relevant references. Related resources are indentified using a string or a number conforming to a formal identification system, e.g. URI, URL, DOI etc. These relations can be used also for identifying different versions, translations etc. The Relation element is optional.

Conforms to

The Conforms element was designed to enhance relations and is used to describe references to

a document informing on the standard used while creating an article or the standard referenced in the content. This element is optional.

References

The References element also strives to enhance or refine relations. It is used for references, in-text citations or other resources used in the article. This element is optional too.

Is referenced by

An optional element used for the sake of relating a published article as a resource for other articles that quote it or draw from its content.

Has part of

The „Has part of" element saves URIs to identify the resource the parts of which (physical or logical) were inserted in the article. This element is optional.

Is part of

The „Is part of" element saves URIs to identify the document that includes a part or parts of the published article (physical or logical). This element is optional.

Has version

The „Has version" optional element saves URIs to identify different versions, modifications, adaptations etc. of the article described.

Is version of

If the article described in VOA3R is a certain version, modification or adaptation of a different document, the URI to identify the latter document should be saved in this element.

Has translation of

This optional element describes URIs identifying different article translations.

Is translation of

If the article, i.e. a research paper, described in VOA3R is a translated version, the URI identifying the original document is saved in the „Is translation of" element.

Has meta-metadata

The „Has meta-metadata" element identifies (by means of URIs) the source of metadata for the article described. Unlike all other elements in this category, it is recommended according to VOA3R AP specification.

Metadata for agents

While talking about research papers or specialist articles, the recommended best practice is to describe also entities that are involved in the paper lifecycle and exercise their influence on it. These are the so-called agents, including concrete persons, institutions, organizations or services.

Name

The „Name" element (compulsory) describes the name of an agent or an organization that is part of agent's description – an author, a contributor or a publisher.

Person

If this agent is a concrete person, his/her surname, name, or as the case may be also his/her mailbox should be introduced. Name and surname elements are strongly recommended while the mailbox one is recommended. In case of a research paper, the description would take the following form:

```
<dcterms:creator>
    <foaf:Person>
        <foaf:firstName xml:lang="en">Pavel</
          foaf:firstName>
        <foaf:lastName xml:lang="en">Simek</
          foaf:lastName>
    </foaf:Person>
    <foaf:Person>
        <foaf:firstName xml:lang="en">Vanek</
          foaf:firstName>
        <foaf:lastName xml:lang="en">Jiri</
          foaf:lastName>
    </foaf:Person>
</dcterms:creator>
```

MARC Relation Properties

The above-mentioned Metadata for agents can be replaced by edt, rev and trl MARC Relation Properties in Dublin Core Metadata.

MARC Relation Properties

The above-mentioned Metadata for agents can be replaced by edt, rev and trl MARC Relation Properties in Dublin Core Metadata.

Edt

This recommended element describes editor's relation to the intellectual content of an article. The information is expressed by means of a vCard or a URI.

Rev

This recommended element describes reviewer's relation to the intellectual content of an article.

The information is expressed by means of a vCard or a URI.

Trl

This recommended element describes translator's relation to the intellectual content of an article. The information is expressed by means of a vCard or a URI.

Research

In order to both enhance and refine the description of research and scholarly papers, the Research elements characterising in detail the domain, procedures, methods, instruments etc. are the most suitable and recommended.

Object of interest

The „Object of Interest" element is used in order to specify the research domain or field of the paper. This element is recommended. However, it should be considered compulsory for research papers and the best practice is to use the AGROVOC thesaurus. The present element comprises URIs to concrete AGROVOC identifiers or keywords corresponding to the AGROVOC thesaurus.

Variable

The „Variable" element describes research objects (or measurement objects) that constitute the topic of the paper. The information can be text-based or vocabulary or thesaurus-based, e.g. the AGROVOC can be used. The description is similar to the „Object of Interest" element.

Method

The „Method" element (recommended) describes the procedures and methods used in the research the paper deals with. It is recommended to use a free text form e.g. weighted sum approach.

Protocol

„Protocol" is a recommended element that – by means of a free text – provides information on standardised methods used to create and process research data.

Instrument

This recommended element provides information on data mining tools and instruments used in the research described.

Techniques

The present recommended element provides descriptions of the data mining and data processing techniques that were used in the research the paper aims at.

Conclusion

Recently, a number of central repositories, e.g. citation databases, have been implementing various metadata harvesting mechanisms and it can be assumed that these mechanisms will be strictly required in near future. Moreover, it is very likely that local repositories will be forced to employ a quality metadata content description and metadata harvesting system. Most leading citation databases consider metadata, or as the case may be metadata harvesting systems, conditional for integrating or monitoring the repository.

Based on the analysis of metadata formats related to the lifecycle of a research paper in agriculture, food industry, aquaculture, environment or rural development, the authors would recommend to describe each paper by metadata that clearly identify its content and properties. This meta-description should be done as soon as the paper is published. VOA3R AP is one of the most suitable metadata formats in this domain. It is partially patterned on the Dublin Core and combined with the AGROVOC thesaurus. The first metadata set should be created by paper authors themselves (e.g. abstract, keywords etc.) while publishers should be in charge of the second metadata set (information on the review process status, publisher, copyright etc.). An overview of metadata creators, element statuses and AGROVOC thesaurus use is given in Fig. 3 – Fig. 6 below.

Element	Source	Note	Status
abstract	dcterms		comp.
bibliographicCitation	dcterms		comp.
creator	dcterms		comp.
description	dcterms		comp.
objectOfInterest	voa3r	AGROVOC	comp.
subject	dcterms	AGROVOC	comp.
title	dcterms		comp.
variable	voa3r	AGROVOC	comp.
edt	marcrel		recom.
hasMeta-metadata	voa3r		recom.
instrument	voa3r		recom.
method	voa3r		recom.
protocol	voa3r		recom.
techniques	voa3r		recom.

Figure 3: Compulsory, and recommended elements – author.

Element	Source	Note	Status
alternativeTitle	dcterms		opt.
conformsTo	dcterms		opt.
hasPartOf	dcterms		opt.
hasTranslationOf	voa3r		opt.
hasVersion	dcterms		opt.
isPartOf	dcterms		opt.
isTranslationOf	voa3r		opt.
isVersionOf	dcterms		opt.
references	dcterms		opt.

Figure 4: Optional elements – author.

Element	Source	Note	Status
accessRights	dcterms		comp.
date	dcterms		comp.
format	dcterms		comp.
identifier	dcterms		comp.
licence	dcterms		comp.
publicationStatus	voa3r		comp.
publisher	dcterms		comp.
reviewStatus	voa3r		comp.
rights	dcterms		comp.
type	dcterms		comp.
isShownBy	ese		recom.
rev	marcrel		recom.

Figure 5: Compulsory and recommended elements – publisher.

Element	Source	Note	Status
language	dcterms		comp.
contributor	dcterms		recom.
name	foaf		recom.
person	foaf		recom.
isReferencedBy	dcterms		opt.
relation	dcterms		opt.
trl	marcrel		opt.

Figure 6: Shared elements - authors and publishers.

If all authors and publishers add at least compulsory (or as the case may be also recommended or optional) elements during the research paper lifecycle, these papers or articles will be very well meta-described from the viewpoint of their content and properties. As a result, the searching efficiency over local repositories and automatic metadata harvesting for the sake of central repositories or citation databases will improve significantly.

Acknowledgements

The knowledge and data presented in the present paper were obtained as a result of the following research programs and grant schemes:

Grant No. 20121044 of the Internal Grant Agency titled „Using Automatic Metadata Generation for Research Papers“.

Grant agreement No. 250525 funded by the European Commission corresponding to the VOA3R Project (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment), http://voa3r.eu.

Research Program titled „Economy of the Czech Agriculture Resources and their Efficient Use within the Framework of the Multifunctional Agrifood Systems“ of the Czech Ministry of Education, Youth and Sport number VZ MSM 6046070906.

Corresponding author:
Ing. Pavel Šimek, Ph.D.
Department of Information Technologies, Faculty of Economics and Management,
Czech University of Life Sciences in Prague, Kamýcká 129, 165 21 Prague 6- Suchdol, Czech Republic
E-mail: simek@pef.czu.cz

References

[1] Agricultural Information Management Standards. AGROVOC. [online]. Available at http://aims.fao.org/standards/agrovoc/about.

[2] Ardö, A. Can We Trust Web Page Metadata? Journal of Library Metadata Volume 10, Issue 1, 2010, pages 58-74. ISSN 19386389.

[3] Bansal, N., Malik, S. K. A framework for agriculture ontology development in semantic web. Proceedings 2011 International Conference on Communication Systems and Network Technologies, CSNT 2011, 3. – 5. June 2011. p. 283 – 286. ISBN 978-0-7695-4437-3.

[4] Diamantopoulos, N., Sgouropoulou, C., Kastrantas, K. Manouselis, N. Developing a metadata application profile for sharing agricultural scientific and scholarly research resources. Communication in Computer and Information Science. Volume 240 CCIS, 2011. p. 453 – 466. ISSN 1865-0929.

[5] The Dublin Core Metadata Initiative. Dublin Core Metadata Element Set [online]. Version 1.1, 11.10.2010 [cit. 2012-09-10]. Available at http://dublincore.org/documents/dces.

[6] The Library of Congress. Outline of Elements and Attributes in MODS Version 3.4 [online]. Version 3.4, 22.9.2011 [cit. 2012-09-11]. Available at http://www.loc.gov/standards/mods/mods-outline.html.

[7] Open Archive Initiative. The Open Archives Initiative Protocol for Metadata Harvesting [online]. Version 2008-12-07T20:42:00Z [cit. 2012-09-10]. Available at http://www.openarchives.org/OAI/openarchivesprotocol.html.

[8] Protonatorios, V., Gavrilut, L., Athanasiadis, I. N., Hatzakis, I. Sicilia, M. – A. Introducing a content integration process for a federation of agricultural institutional repositories. Communication in Computer and Information Science. Volume 240 CCIS, 2011. p. 467 – 477. ISSN 1865-0929.

[9] Santos, C., Riyuiti, A. An overview of the use of metadata in agriculture. IEEE Latin America Transactions, volume 10, issue 1, January 2012. p. 1265 – 1267. ISSN 1548-0992.

[10] Wei, Y. Y., Wang, R. J., Hu, Y. M., Wang, X. From Web Resources to Agricultural Ontology: A Method for Semi-Automatic Construction. Journal of Integrative Agriculture, volume 11, issue 5, May 2012. p. 775 – 783. ISSN 2095-3119.

[11] World Wide Web Consortium. Date and Time Formats. [cit. 2012-09-15]. Available at http://www.w3.org/TR/NOTE-datetime.