Using metadata formats and AGROVOC thesaurus for data description in the agrarian sector

P. Šimek, J. Vaněk, J. Jarolímek, M. Stočes, T. Vogeltanzová

Faculty of Economics and Management, Czech University of Life Sciences Prague, Prague, Czech Republic

ABSTRACT

The paper deals with a general solution of semantic description related to various electronic data formats in the domains of agriculture, aquaculture, forestry, food industry, environment, horticulture and rural areas. The solution presented was developed on the basis of metadata formats analysis and then complemented with its own software superstructure. It is based on the VOA^{3}R metadata application profile (AP) that was developed within the framework of the Virtual Open Access Agriculture and Aquaculture Repository project for the sake of describing research papers and scientific publications. However, thanks to its complexity and comprehensiveness, it is also suitable for different kinds of data and while combining it with the AGROVOC thesaurus, which is elaborated by the Food and Agriculture Organization of the United Nations, it can become a robust and universal tool for data characteristics and semantic description. The potential of the proposed solution is illustrated by means of two examples. While the first example brings a metadata description of a photograph depicting anthocyanin pigmentation of barley straw caused by phosphorus deficiency, the second example is related to a research paper description. The developed system of metadata description is broadly applicable in agriculture such as in precision agriculture, in plant production or in ground cover monitoring and evaluation based on sensor or visual data.

Keywords: Dublin Core; VOA^{3}R; agrarian data description; open access; electronic data formats

Due to an incessant increase in data volume, it is necessary to describe data in an efficient manner (i.e. to provide primarily a quality content description) and to dispose of tools for their classification, search, sharing, administration, or, as the case may be, also for their automated distribution. Currently, this trend can be observed in all spheres of human activity, including the agrarian sector, where new and new data have been acquired not only by humans but more and more also automatically, e.g. by means of various sensors. According to the study carried out by the EMC corporation, the data volume is expected to reach 3.3 ZB (zetta bytes) in 2013. Therefore, the need for data characterization and semantic description has been increasing together with the need for making relevant metadata available.

While talking about the agrarian sector, it is obvious that very heterogeneous data from very heterogeneous fields are collected. These can be both structured and unstructured, integrating database entries, texts, charts, figures, photographs, audio and video files, records from measuring devices and sensors, geolocation data, text messages, websites etc. For instance, the precision agriculture itself generates a vast amount of data that originate from on-the-go land and crop sensoring (Peets at al. 2012). Metadata (data about data content) represent a sophisticated data description that can be used for all electronic objects or database entries. Nevertheless, metadata should characterize and describe objects in a relevant manner, which is sometimes (websites in particular) not the case (Ardo 2010). In case of web resources, the data are

also very often unstructured and heterogeneous (Perez-Catalan et al. 2013).

Metadata incorporate a certain information value of the data they describe. Thus, metadata provide efficient data characteristics and subsequently facilitate data processing, classification, search etc. Creation and exchange of meaningful metadata is therefore crucial for enhancing the interoperability between and among repositories and providing value added services (Subirats et al. 2012). Data users themselves also need to work with different data centres and metadata models (Wang et al. 2012).

Nowadays, the importance of Open Access has been stressed. The Council of the European Union accentuated in COM 2007 (Commission of the European Communities 2007) that accessing, sharing and saving scientific data and information represents key elements of developing the European Research Area. Both European and international consortia emphasize open access (Aguillo 2012). CERN (European Organization for Nuclear Research), for instance, strives to provide open access publishing environment (Pepe and Yeomans 2007) and this trend is set to continue and spread in other domains of human activity, too.

MATERIAL AND METHODS

The main objective of the paper is to determine a general procedure of semantic description related to various electronic data formats in the domains of agriculture, aquaculture, forestry, food industry, environment, horticulture and rural areas. Analysis of the existing metadata formats, their applicability, implementation and software (SW) realization of metadata formats administration system account for a secondary objective of the present paper.

In short, the methodology is based on analyzing the metadata formats, synthesis of theoretical findings and subsequent realization of a general semantic data description procedure. Key metadata format requirements were the following:

(1) complexity (data description extent),
(2) detail (data description detail),
(3) universality (electronic data formats applicability in particular),
(4) relative simplicity of implementation,
(5) standardization and international use,
(6) Open Access support.

There exist a lot of metadata formats describing various kinds of objects by specific elements. These formats were developed within the framework of the research projects, by communities or standardising bodies. The following metadata formats and thesauri were shortlisted for an in-depth analysis:

– general

DC (Dublin Core) is one of the most universal metadata formats for data description, which consists of 15 basic (recommended) metadata elements (Dublin Core Metadata Initiative 2010).

– specific

MODS (Metadata Object Description Schema) – format designed primarily for library applications entails 20 top-level elements with optional attributes (The Library of Congress 2011).

Exif (Exchangeable image file format) – format designed for images in JPEG or TIFF formats.

– domain-specific focused on the agriculture, food and aquaculture sectors

VOA^{3}R Metadata AP (Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile).

AGROVOC – thesaurus development and maintenance is coordinated by the FAO (Food and Agriculture Organization of the United Nations).

AGROTERM – Czech agricultural thesaurus for document cataloguing (Agricultural and Food Library).

VOA^{3}R Metadata AP format was developed with a view to improve data description and sharing in the domains of agriculture, aquaculture, environment and rural development. The VOA^{3}R Metadata AP was created within the framework of the Virtual Open Access Agriculture and Aquaculture Repository project. The latter format is partially based on the DC standard (Šimek et al. 2012). VOA^{3}R Metadata AP can be used to integrate different metadata APs (Application Profiles) from different repositories, comprising also the research content related to agriculture (Protonotarios et al. 2011). VOA^{3}R Metadata AP basic scheme is shown below (Figure 1).

A complex data or object description can be attained by means of the compulsory and highly recommended elements. In order to create more detailed characteristics, it would be appropriate to include also the recommended or even optional elements.

AGROVOC is a thesaurus containing more than 32,000 entries in 22 languages (as at March 14, 2013) and covering topics related to food, nutrition, agriculture, fishery, forestry, environment and other related domains. It serves to indexing documents in agricultural information systems, primarily in the international AGRIS system.

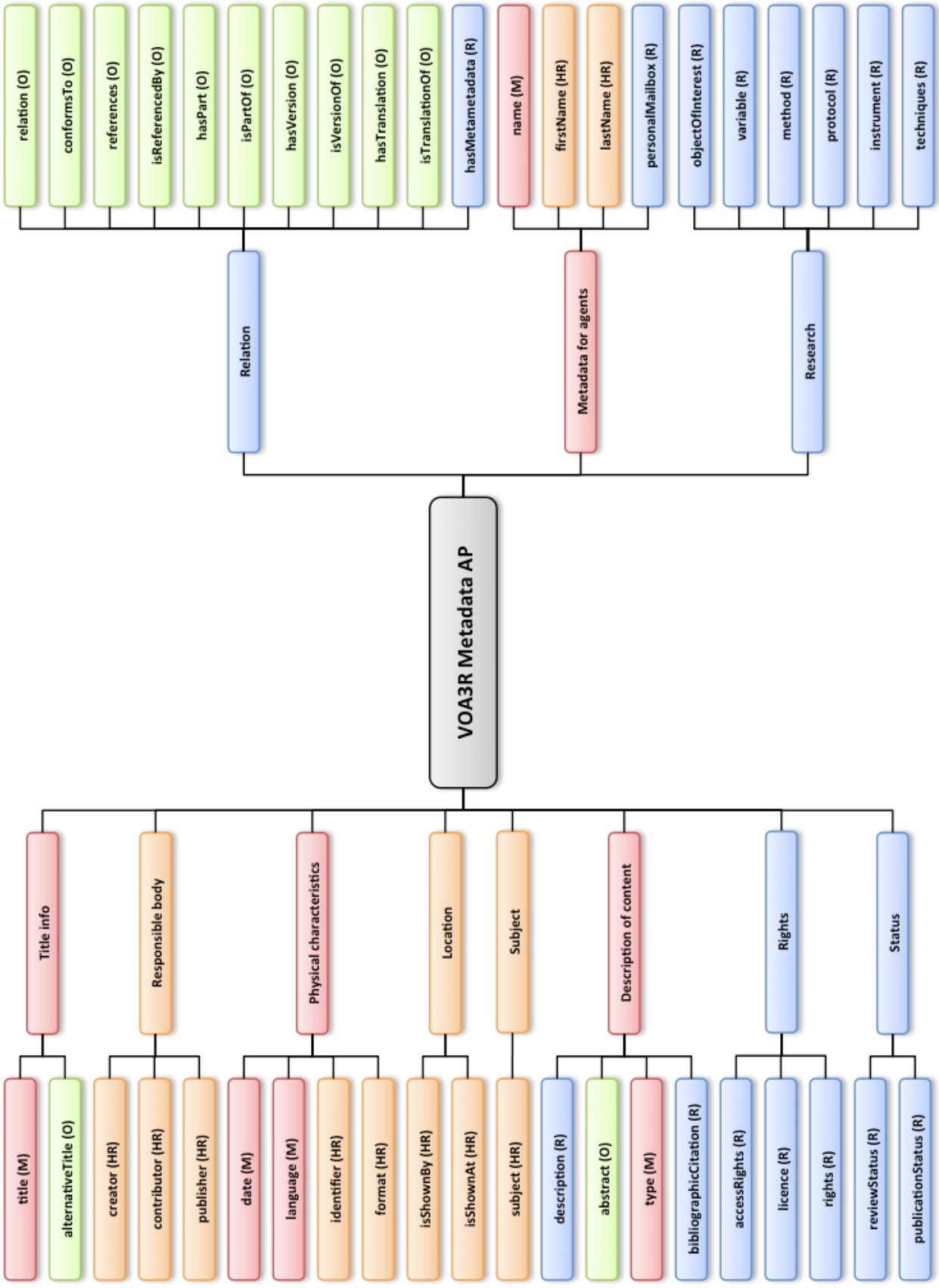

Figure 1. VOA^{3}R Metadata application profile (AP) structure, source: VOA^{3}R. M – compulsory element (red); R – recommended element (orange); HR – highly recommended element (blue); O – optional element (green)

RESULTS AND DISCUSSION

It clearly stems from the metadata formats applicability analysis that the VOA^{3}R Metadata AP is the most suitable and appropriate Open Access platform for describing data in the agrarian sector and rural areas. Based on the analysis, a general semantic data description procedure was developed, resulting not only in the primary use of the VOA^{3}R Metadata AP but also in using the AGROVOC thesaurus for some elements and in complementing semantic description automatically by the DC or, as the case may be, by the Exif for visual data.

A general procedure concerning semantic description of various electronic data formats can be outlined as follows:

The main steps of electronic data semantic description are as follows:

– object identification (data format),
– editing or filling-in the Exif (only for the JPEG, TIFF or RIFF WAVE graphic formats),
– content and characteristics description (i.e. filling in the respective VOA^{3}R Metadata AP elements) using the AGROVOC thesaurus,
– automated generation of the DC,
– metadata revision.

Based on the object identification (electronic data format), the Exif standard using JPEG, TIFF or RIFF WAVE graphic forms is either edited or created. The Exif mostly stores interchange information on the image file such as e.g. date and time when the image was taken, technique used, exposure time, aperture etc.

Filling in the VOA^{3}R Metadata AP elements itself can be a rather demanding process which basically depends on the description extent and depth required by authors or data administrators. The difference in description also arises from the nature, format and use of data. It is clear that a photo saved in a database on one hand, and a published paper with many references on the other, deserve a different extent of metadata description. Editorial boards of specialist journals can define the description extent in a different way than those of scientific journals. The deeper and more detailed the description, the better and clearer the content identification. At least, the compulsory elements i.e. title, date, language, type and name and highly recommended elements creator, contributor, publisher, identifier, format, isShownBy, isShownAt, subject a firstName a lastName in

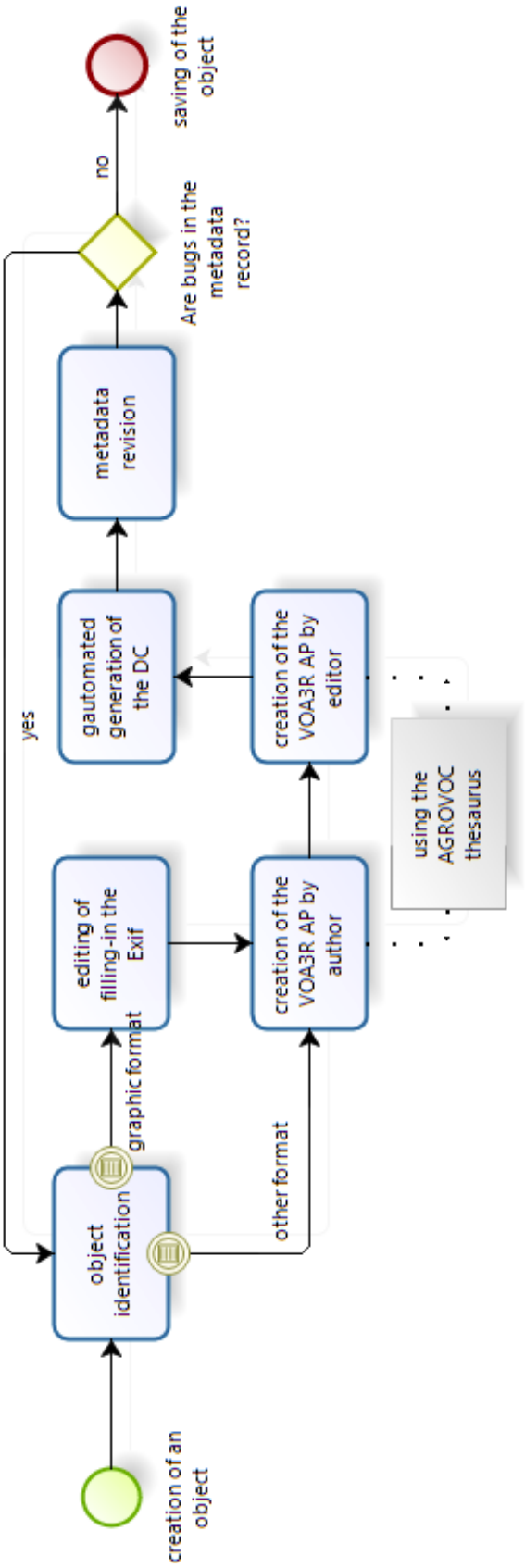

Figure 2. General procedure concerning semantic description of various electronic data formats

the metadata for agents should be introduced. In case of subject/object of interest elements, the AGROVOC thesaurus should be used in order to make an internationally unified record and to avoid at the same time incompatibility or typing errors.

Since the VOA^{3}R Metadata AP is partially patterned on the DC, it is quite easy to generate automatically also this metadata format. Therefore, the metadata record might be processed also by systems endorsing only this basic format. When metadata are checked, they are saved in the database, unequivocally identified and unmistakably matched with the object described (data).

In order to illustrate the above solution, two different metadata records are shown. The first example deals with describing a photograph while the second one is concerned with a research paper description.

Example 1 – description of a photograph depicting anthocyanin pigmentation of barley straw bases caused by phosphorus deficiency.

- **title**: Anthocyanin pigmentation of barley straw bases caused by phosporus deficiency
- **creator**: Václav Vaněk
- **contributor**: Jindřich Černý
- **publisher**: Czech Univeristy of Life Sciences in Prague
- **date**: 2011-04-01
- **language**: CZE
- **identifier**: http://www.domena.cz/jecmen/DCF20110456.jpg
- **format**: image/jpeg
- **isShownBy**: http://www.domena.cz/jecmen
- **isShownAt**: http://www.domena.cz/jecmen/DCF20110456.jpg
- **subject**:
  - barley (AGROVOC 823, unequivocal AGROVOC thesaurus identifier)
  - barley straw (AGROVOC 15941)
  - phosphorus (AGROVOC 5803)
  - phosphorus fertilizers (AGROVOC 27907)
- **abstract**: Long-term substantial lack of phoshorus is manifested by these distinctive signs – lower growth, narrower, smaller and upright leaves, thin stems, limited stolons and root growth. Leaves and basal parts bluish green turning often into red or purple tints due to a higher anthocyanin production (Vaněk et al. 2012).
- **description**: The photograph depicts a long-term phosporus lack in the soil and the impact on green barley.
- **type**: Other
- **access rights**: Restricted Access
- **rights**: All rights reserved
- **name**: Department of Information Technologies FEM CULS in Prague
- **firstName**: Pavel
- **lastName**: Šimek
- **personalMailBox**: simek@pef.czu.cz
- **objectOfInterest**
  - barley (AGROVOC 823)
  - phosphorus (AGROVOC 5803)
- instrument
  - Olympus E-M5

As soon as the metadata are filled in, the respective entry is saved in the database and unambiguously linked with the data as such, in this particular case with the photo saved on a hard drive of a server or digital repository.

Example 2 – Description of a research paper

- **title**: Foliar fertilization with molybdenum in sunflower (*Helianthus annuus* L.)
- **creator**: P. Škarpa, E. Kunzová, H. Zukalová
- **contributor**: V. Vaněk
- **publisher**: Czech Academy of Agricultural Sciences
- **date**: 2013
- **language**: EN
- **identifier**: http://www.agriculturejournals.cz/publicFiles/87628.pdf
- **format**: application/pdf
- **isShownBy**: http://www.agriculturejournals.cz/web/pse.htm?volume=59&type=volume
- **isShownAt**: http://www.agriculturejournals.cz/publicFiles/87628.pdf
- **subject**:
  - foliar nutrition (36147)
  - oil content (12910)
  - fatty acids (2818)
  - sunflower (14691)
  - nutrition (49892)
- **description**: The article describes foliar fertilization with molybdenum in sunflower.
- **abstract**: The objective of the vegetation experiment established in 2008–2011 was to explore the effect of the time and dose of foliar molybdenum (Mo) application ...

– **type**: paper
– **bibliographicCitation**:
Ayala J., Castillo A. M., Colinas M. T., Pineda J. Effect of foliar application of calcium, boron and molybdenum in nutrient content of poinsettia plants. HortScience, *40*: 1086.
...
– **accessRight**: Open Access
– **rights**: All rights reserved
– **reviewStatus**: Peer reviewed
– **publicationStatus**: published
– **isPartOf**: Plant, Soil and Environment
– **name**: Department of Information Technologies PEF ČZU in Prague
– **firstName**: Pavel
– **lastName**: Šimek
– **personalMailBox**: simek@pef.czu.cz
– **objectOfInterest**:
  ○ foliar nutrition (36147)
  ○ oil content (12910)
  ○ sunflower (14691)
– **variable**:
  ○ nutritional value (12872)
– **method**:
  ○ Kjeldahl method
  ○ method of Jones
– **instrument**:
  ○ Grinder
  ○ spectrometer
– **technique**: Nutritional analysis

The two examples clearly illustrate that, while using the metadata description introduced in the paper, the data will be very well organised and described in an efficient and structured form. In other words, metadata records can be subsequently processed not only by users but also by various devices, harvesting mechanisms, machines etc. As a result, data classification will become easier as well as the searching efficiency and distribution will be significantly improved. The developed system of metadata description is broadly applicable in agriculture such as in precision agriculture, in plant production or in ground cover monitoring and evaluation based on sensor or visual data.

Figure 3. Anthocyanin pigmentation of barley straw bases caused by phosphorus deficiency (Vaněk et al. 2012)

Acknowledgements

The knowledge and data presented in the paper were obtained as a result of the following research program and grant scheme: Grant agreement No. 250525 funded by the European Commission corresponding to project VOA3R (Virtual Open Access Agriculture and 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 Sports, Project No. MSM 6046070906.

REFERENCES

Aguillo I.F. (2012): Technologies, research and future of the profession. EPI, Barcelona. Profesional de la Informacion, *21*: 5–7.

Ardö A. (2010): Can we trust web page metadata? Journal of Library Metadata, *10*: 58–74.

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

Commission of the European Communities (2007): Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation, Brussels. Available at http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf

Dublin Core Metadata Initiative (2010): Dublin Core Metadata Element Set. Version 1.1, [cit. 2013-01-30]. Available at http://dublincore.org/documents/dces

Library of Congress (2011): Outline of Elements and Attributes in MODS Version 3.4. Available at http://www.loc.gov/standards/mods/mods-outline.html

Peets S., Mouazen A.M., Blackburn K., Kuang B., Wiebensohn J. (2012): Methods and procedures for automatic collection and management of data acquired from on-the-go sensors with ap-

plication to on-the-go soil sensors. Computer and Electronics in Agriculture, *81*: 104–112.

Pepe A., Yeomans J. (2007): Protocols for scholarly communication. Astronomical Soc Pacific, San Francisco. Astronomical Society of the Pacific Conference Series, *377*: 147–154.

Perez-Catalan M., Berlanga R., Sanz I., Aramburu M.J. (2013): A semantic approach for the requirement-driven discovery of web resources in the Life Sciences. Knowledge and Information Systems, *34*: 671–690.

Protonotarios V., Gavrilut L., Athanasiadis I., Hatzakis I., Sicilia M.A. (2011): Introducing a content integration process for a federation of agricultural institutional repositories. Metadata and semantic research. Communication in Computer and Information Science, *240*: 467–477.

Subirats I., Malapela T., Dister S., Zeng M., Gooaverts M., Pesce V., Jaques Y., Anibaldi S., Keizer J. (2012): Reorienting Open Repositories to the Challenges of the Semantic Web: Experiences from FAO's Contribution to the Resource Processing and Discovery Cycle in Repositories in the Agricultural Domain. Metadata and Semantic Research. Communication in Computer and Information Science, *343*: 158–167.

Šimek P., Vaněk J., Otčenášek V., Stočes M., Vogeltanzová T. (2012): Using Metadata Description for Agriculture and Aquaculture Papers. Agris on-line Papers in Economics and Informatics. Czech University of Life Sciences in Prague, Prague, 79–80. Available at http://online.agris.cz/files/2012/agris_online_2012_4_simek_vanek_ocenasek_stoces_vogeltanzova.pdf

Vaněk V., Balík J., Černý J., Pavlík M., Pavlíková D., Tlustoš P., Valtera J. (2012): Garden Plants Nutrition. Academia, Praha. (In Czech)

Wang H.L., Shao Y.Z., Di L.P., Kang L.J. (2012): Discovery of Agriculture Data using Federated Catalogue Service. In: Procedings of the 1st International Conference on Agro-geoinformatics, IEEE, New York, 645–649.

Agricultural and Food Library. Thesauri. Available at http://www.nzpk.cz/tezaury/ (In Czech)

Received on April 15, 2013
Accepted on May 30, 2013

Corresponding author:

Ing. Pavel Šimek, PhD., Česká zemědělská univerzita v Praze, Provozně ekonomická fakulta, Katedra informačních technologií, Kamýcká 129, 165 21 Praha 6-Suchdol, Česká republika
e-mail: simek@pef.czu.cz