Agris on-line Papers in Economics and Informatics

Volume V Number 4, 2013

Using of Automatic Metadata Providing

P. Šimek, M. Stočes, J. Vaněk, J. Jarolímek, J. Masner, I. Hrbek

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

Anotace

Příspěvek prezentuje nezbytnost systémového řešení pro poskytování metadat lokálními archívy do centrálních repozitařů a jeho následnou realizaci Katedrou informačních technologií Provozně ekonomické fakulty České zemědělské univerzity v Praze pro potřeby agrárního WWW portálu AGRIS. Systém podporuje OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting), více metadatových formátů a tezaurů a vyhovuje jakostním požadavkům v podobě funkčnosti, bezporuchovosti, použitelnosti, udržitelnosti a přenositelnosti. SW aplikace pro obsluhu žádostí OAI-PMH je provozována v prostředí WWW serveru Apache s využitím výkonného PHP frameworku Nette a databázové vrstvy dibi.

Klíčová slova

Metadata, OAI-PMH, archiv, agrární portál, distribuce metadat.

Abstract

The paper deals with the necessity of systemic solution for metadata providing by local archives into central repositories and its subsequent implementatiton by the Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences in Prague, for the needs of the agrarian WWW AGRIS portal. The system supports the OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) protocol, several metadata formats and thesauri and meets the quality requirements: functionality, high level of reliability, applicability, sustainability and transferability. The SW application for the OAI-PMH requests' servicing is run in the setting of the WWW Apache server using an efficient PHP framework Nette and database dibi layer.

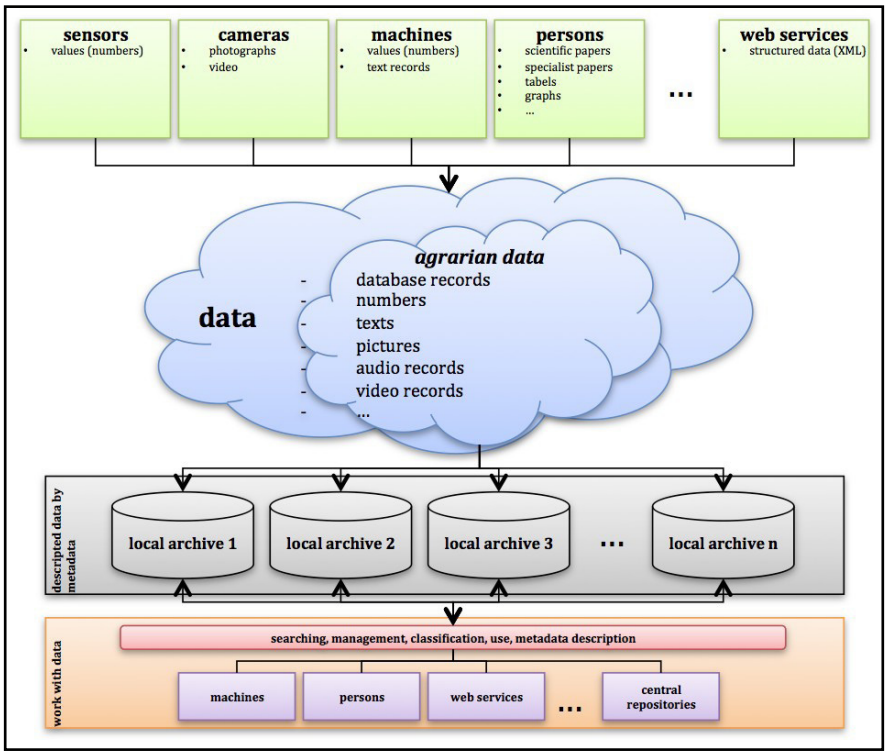
Key words

Metadata, OAI-PMH, archive, agrarian portal, metadata distribution.

Introduction

Current on information and knowledge based society (economy) is characterized by an increasing number of information sources in all spheres of human activity, in various shapes and rather different quality and accessibility. New data are created not only by humans but more and more also automatically, e.g. by means of various sensors, cameras etc. There is need for effective semantic data description, their storing, administration and processing or their automated distribution in standardized formats. Another problem is that in agriculture, aquaculture, food industry, environment and rural development very heterogeneous data are collected. These can be both structured and unstructured data integrating database entries, texts, charts, figures, photographs, audio and video files, records from measuring devices and sensors, geolocation data, text messages, websites, presentations, animations etc. For efficient and brief data characterization metadata (data about data content) are being used. Via metadata we can describe all electronic objects or database entries. Thus, metadata provide efficient data characteristics and subsequently facilitate data processing, classification, search etc. According to T. Berners-Lee metadata are in fact machines of meaningful information (Berners-Lee, 1997).

There have been dynamic changes not only in the number of sources but mainly in their form and structure. It has led to stricter demands on local archives, e.g. independent data stores such



Source: own processing

Figure 1: The principle of data and metadata creation and distribution.

as archives of scientific and scholarly research journals, the agrarian www portal etc. Metadata help to aggregate local archives into thematically and technically oriented central repositories. Nevertheless, the aggregation of metadata from various sources often leads to problems such as incompatibility between and among various metadata APs (application profile) or metadata quality. The existence of significantly different metadata APs is the main source of problems with the stores interconnecting (Protonatorios, 2011). In order for the local archives to be utilizable and competitive in the future, their content must be described by metadata and it must provide support for automatic metadata harvesting. However, metadata harvesting itself requires the timely provision of up-to-date records via global networks with minimal demands on both the metadata providing machine and the harvesting machine (Adly, 2009). For the users themselves it is much more comfortable and efficient if they can run their search in one central repository than in many independent local archives and digital libraries (Kadury, 2007).

Materials and methods

Modern systems in libraries and archives have at their disposal various devices for automatic content providing and sharing. Even local archives of technical or scientific and scholarly research journals, web portals, news agencies'servers etc. have been more and more equipped with these systems which enable metadata creation, administration and subsequently automatic distribution. Key requirements for these systems are the following:

- support of several metadata formats, but at least the DC (Dublin Core) and the VOA3R Metadata AP (Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile) plus the support of the AGROVOC thesaurus
- the OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) support
- quality requirements for IS (Information System)

Metadata formats and AGROVOC thesaurus support

Nowadays, there has been more and more pressure on the metadata content and the content characteristics to be described by thesauri or dictionaries not only because of easier access and faster search by users but mainly because

of the automatic accessibility of the content to various machines, web services and databases. Based on analyzing metadata formats for the needs of the agrarian sector the following were chosen (Šimek, 2013):

- metadata format VOA3R Metadata AP,
- international DC standard,
- AGROVOC thesaurus.

Dublin Core (DC)

The Dublin Core (Dublin Core Metadata Initiative) is one of the most universal metadata formats for data description. It consists of 15 basic (recommended) elements that are suitable for describing almost any kind of object. The elements are following:

contributor	publisher
Coverage	relation
Creator	rights
Date	source
description	subject
Format	title
Identifier	type
Language	

Source: The Dublin Core Metadata Initiative, 2010

Table 1: The list of 15 DC elements.

The authors of the DC metadata specification (semantics) didn't just modify the existing MARC format but proposed a completely new data file to describe digital documents (Grandmann, 1998).

Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile (VOA3R Metadata AP)

The VOA3R Metadata AP Format was developed with a view to improve data description and sharing in the domains of agriculture, aquaculture, environment and rural development within the framework of the Virtual Open Access Agriculture and Aquaculture Repository project (Sgouropoulou, 2011).

A complex data or object description can be acquired by means of compulsory and highly recommended elements. In order to create more detailed characteristics it is appropriate to include recommended or even optional elements, too. The VOA3R Metadata AP was based on the methodology and elements of Singapore Framework for the DC AP such as function demands, domain model, description set profile, data format and use instructions (N. Diamantopoulos, 2011). This metadata format is primarily based on the DC standard (Šimek, 2012).

AGROVOC thesaurus support

The AGROVOC is the most comprehensive thesaurus containing more than 32,000 entries in 22 languages[1] covering topics related to food industry, nutrition, agriculture, fishery, forestry, environment and other related domains. It serves to indexing documents in agricultural information

[1] as at September 1, 2013

Mandatory	Highly recommended	Recommended	Optional
Title	creator	description	alternativeTitle
Date	contributor	bibliographicCitation	abstract
language	publisher	accessRights	relation
Type	identifier	Licence	conformsTo
Name	format	Rights	references
	isShownBy	reviewStatus	isReferencedBy
	isShownAt	publicationStatus	hasPart
	subject	hasMetametadata	isPartOf
	firstName	personalMailbox	hasVersion
	lastName	objectOfInterest	isVersionOf
		variable	hasTranslation
		Method	isTranslationOf
		protocol	
		instrument	
		techniques	

Source: Sgouropoulou, 2011

Table 2: The list of VOA3R Metadata AP elements.

systems, primarily in the international AGRIS system. The AGROVOC thesaurus development and maintenance is coordinated by the FAO (Food and Agriculture Organization of the United Nations) within the framework of AIMS (Agricultural Information Management Standards).

The whole thesaurus is formulated as the SKOS conceptual system (Simple Knowledge Organization System) and published as Linked Data (linked, interconnected data) which presents a data model for structured dictionaries. The AGROVOC thesaurus conceptual scheme contains full and extensive KOS (Soergel, 2004) using three levels of depiction:

- terms have abstract meaning and are also often described using the URI address (Uniform Resource Identifier): e.g. for maize in the sense of cereals „Concept12322“ is used,
- terms are specified linguistically, e.g. corn, maize, 玉米, maïs,
- terms have specific options (range) such as spelling variations or singulars and plurals, e.g. hen, hens, cow, cows etc. (Agricultural Information Management Standards).

This system secures terminological relations between and among terms and their specific meanings. The AGROVOC is therefore suitable for the description of scientific and scholarly research papers, technical papers, information and news from the agrarian sector, audiovisual data etc. The AGROVOC has one more advantage: it is accessible via web service which can be called from all clients' applications. When using web service, changes on AGROVOC Concept Server can be accessible immediately after their application (Sini, 2008).

OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting)

There are a lot of tools for providing local archives' content and they were analysed in detail. Based on this analysis one of the most universal ways of the contents' providing - the OAI-PMH protocol (Open Archive Initiative – Protocol for Metadata Harvesting) was chosen. Sets of these tools which provide a coherent information presentation across various standards are important not only for data search, but also for data access itself (Devarakonda, 2011). The OAI-PMH defines the mechanism of the metadata records harvesting from various repositories. It means that the OAI-PMH provides a simple technical means to data providers to make its metadata open to services based on widely-spread standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The OAI-PMH was originally developed as a tool for an easy access to various e-print archives via metadata harvesting and aggregation. The protocol proved its usefulness and potential for a wide range of usability just two years after the publication of its permanent version (2.0) (Shreeves, 2005).

Quality requirements

The key quality requirements for IS are following:

- *Functionality:* the ability to tend functions which secure users' implied or set needs while operating the system.
- *High level of reliability:* the ability to maintain the specified level of performance while using the system.
- *Applicability:* the ability to be comprehensive with easy and user friendly operation, and to be attractive while using the system.
- *Sustainability:* the ability to be modified including errors correction, improvements and adjustments needed because of the changes in environment, requirements and functional specification.
- *Transferability:* the ability to be transferred from one environment to another.

Results and discussion

Before the implementation of the OAI-PMH solution itself for the needs of local repository of the agrarian WWW Agris portal or perhaps other local repositories of the Faculty of Economics and Management, the Czech University of Life Sciences, an analysis of freely accessible and the OAI-PMH supporting software (SW) was carried out - for example DSpace, Drupal, etc. The result of the analysis proved that - because of technical reasons - it is not possible to install freely accessible SW and run it in the environment of the Czech University of Life Sciences. It also showed that the analysed software doesn't provide the service required in adequate scale. Freely accessible SW is unsuitable mainly because it can be run on database platforms different from those which are available at the Czech University of Life Sciences or which are supported there. Freely accessible SW was not satisfactorily transferable. Above mentioned problems and errors often appear in other public and private organizations, too. They are usually solved

by the implementation and operation of another platform, together with the provision of staff and material.

The Department of Information Technologies at the Faculty of Economics and Management developed an intuitive system for the semantic description of objects (e.g. scientific and scholarly photographs, statistical and economic data, text messages etc.) by metadata format VOA3R Metadata APs Level 1 – Level 4 and international DC standard. For some metadata elements the AGROVOC thesaurus is being used.

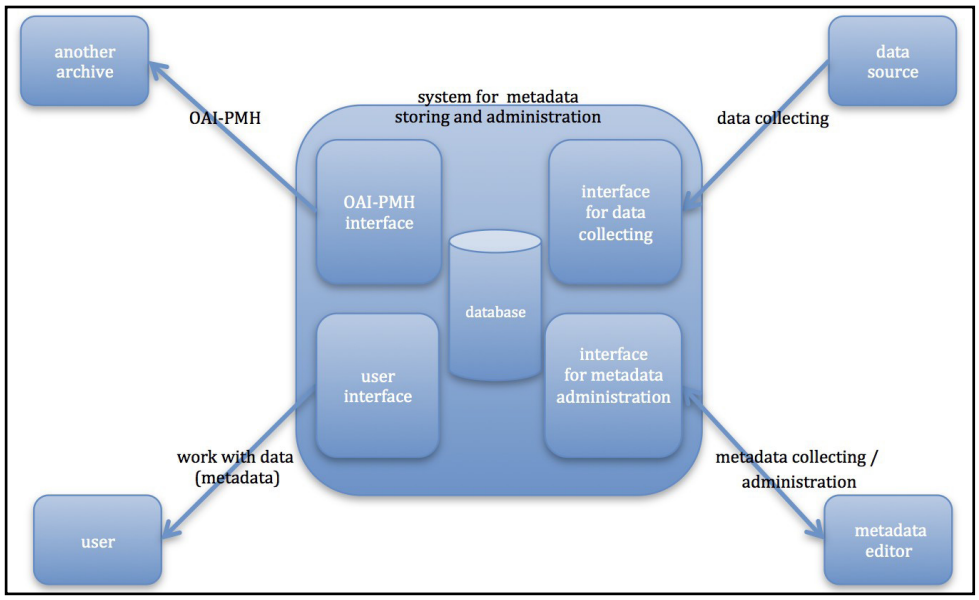
The whole process of metadata collecting, administration and distribution can be characterized by the following diagram (Figure 2).

SW application for the OAI-PMH request operation is run in the setting of WWW Apache server using an efficient PHP framework Nette and a database dibi layer. The Nette Framework was chosen because the final application has the requested quality, and to eliminate security risks. The local metadata repository has been developed and is run on the MySQL database server. The whole SW solution on the World Wide Web platform has been created as a robust modern application with a possibility of further extension and development.

The SW application for the OAI-PMH servicing itself is run separately from the original local archives of the Faculty of Economics and Management. The main reason of this solution has been the possibility of using the new metadata repository by other local archives run by the Czech University of Life Sciences. The web application supporting metadata harvesting provides a simple interface for manual obtaining of concrete objects' metadata stored in local databases and archives.

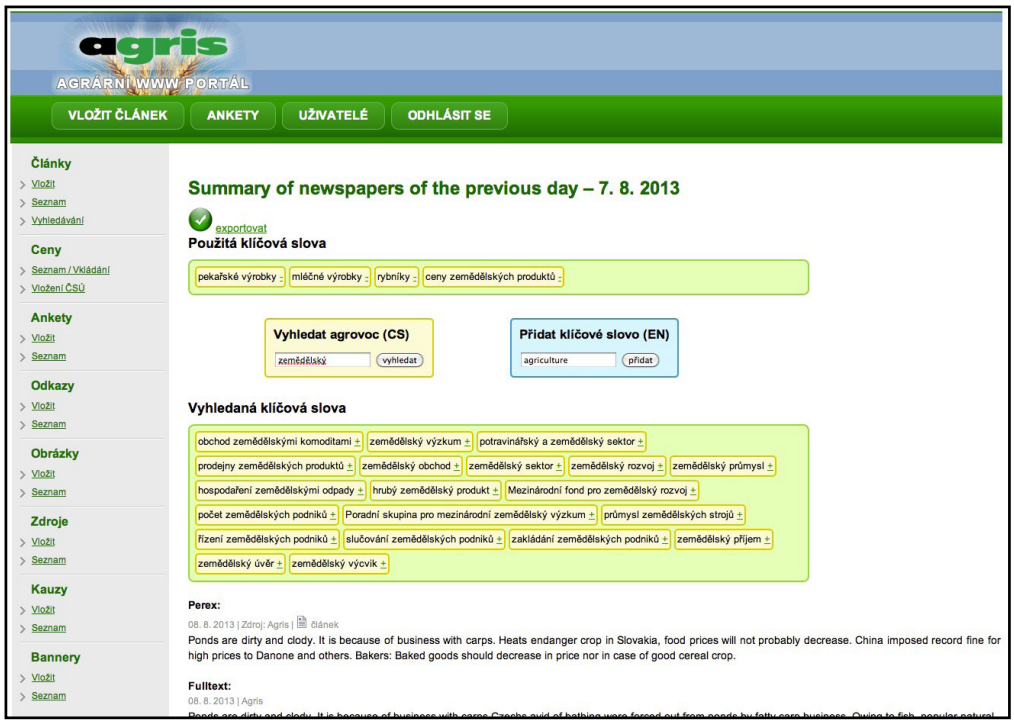
The WWW application for metadata providing has been proposed in such a way that it can - without problems and without huge time demands - send back a valid XML file for all Open Archives Initiative standardized requests including some additional parametres or reports on errors or exceptional conditions.

In order to have a simple metadata administration, the users have at their disposal an SW layer in the shape of users' superstructure above their own database. The users can therefore search in metadata with the help of simple forms. The same applies to their setting or modifying. To enable more comfortable work with the AGROVOC thesaurus, AJAX (Asynchronous JavaScript and XML) has been used. The advantage of this approach is that it is not necessary to download all the dictionary entries into a web browser; a JavaScript code on the background of the form requests from the server relevant terms which it shows to the user. (based on key words filled in by the user). Then the user with a mere click of his mouse on the appropriate icon either matches the chosen AGROVOC term to the described object or alternatively deletes it (Figure 3).



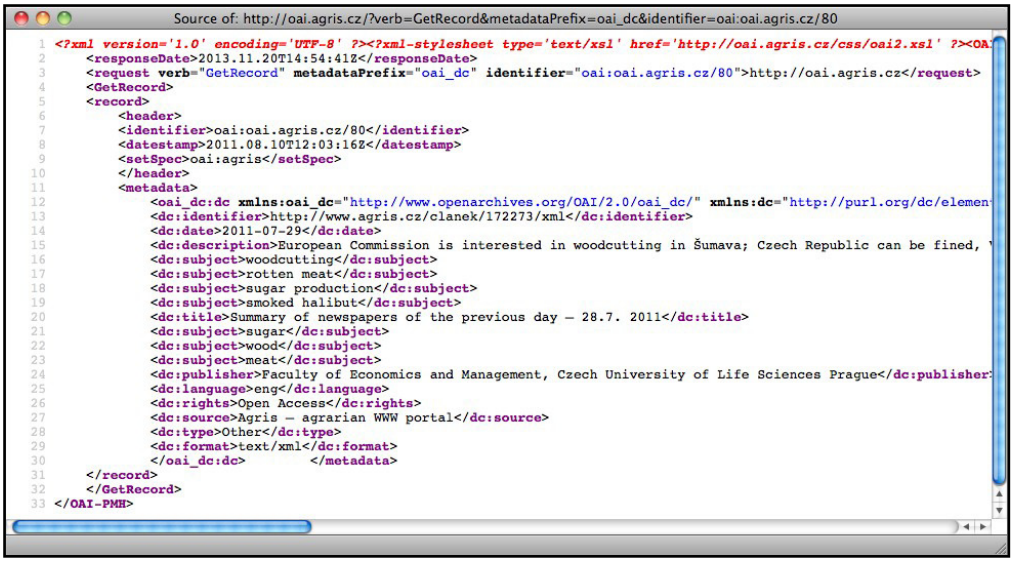
Source: own processing

Figure 2: Metadata administration in the environment of the Faculty of Economics and Management, the Czech University of Life Sciences.



Source: own processing

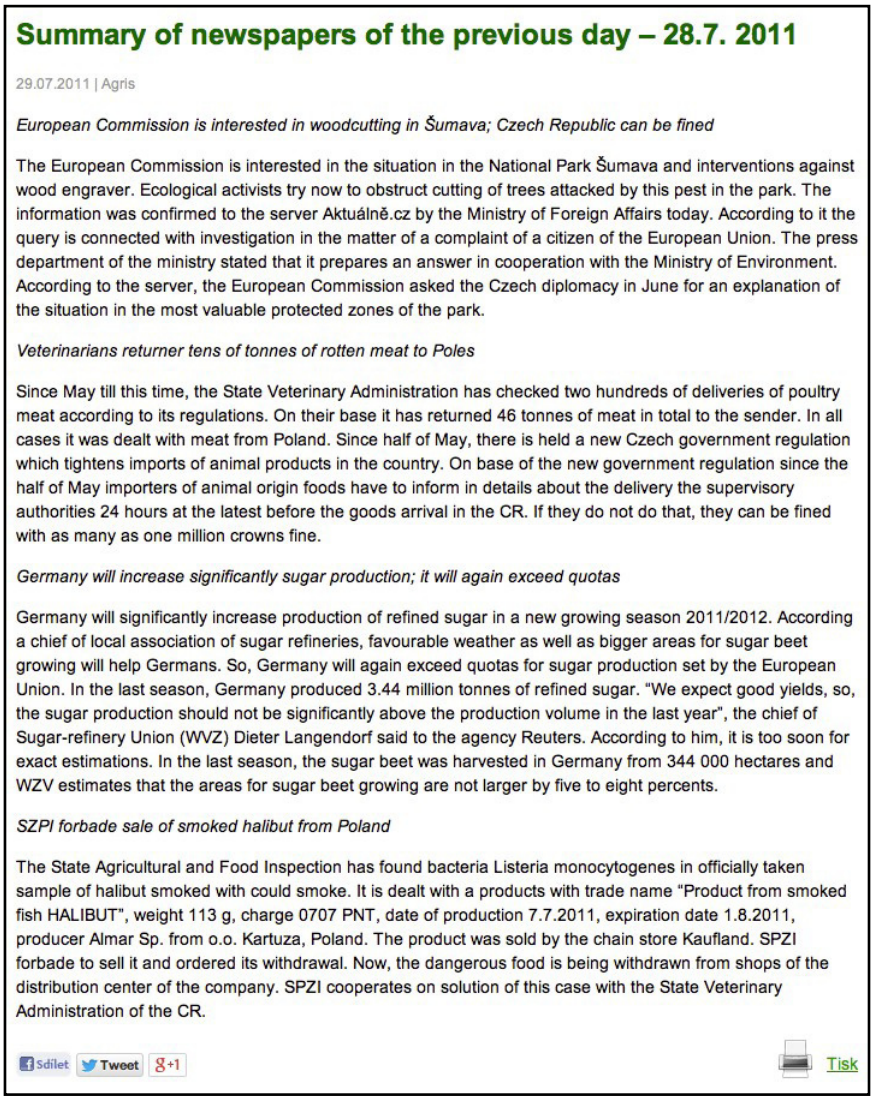
Figure 3: Metadata (the key words of the AGROVOC thesaurus) administration by the user.



Source: own processing

Figure 4: An example of a metadata record of a selected paper from the agrarian WWW AGRIS portal.

The final solution was implanted - apart from others - on the agrarian WWW AGRIS portal which has in the long term held an important position among departmental information sources. The agrarian WWW AGRIS portal provides an integrated on-line platform for information publishing for the domain of the agrarian sector and rural areas. Currently, there are in the agrarian WWW portal database about 100,000 papers and there has been an incessant increase - about several dozens of papers a day. Daily summaries in English are described by metadata from the VOA3R Metadata AP and by the AGROVOC thesaurus (Figure 4).



Summary of newspapers of the previous day – 28.7. 2011

29.07.2011 | Agris

European Commission is interested in woodcutting in Šumava; Czech Republic can be fined

The European Commission is interested in the situation in the National Park Šumava and interventions against wood engraver. Ecological activists try now to obstruct cutting of trees attacked by this pest in the park. The information was confirmed to the server Aktuálně.cz by the Ministry of Foreign Affairs today. According to it the query is connected with investigation in the matter of a complaint of a citizen of the European Union. The press department of the ministry stated that it prepares an answer in cooperation with the Ministry of Environment. According to the server, the European Commission asked the Czech diplomacy in June for an explanation of the situation in the most valuable protected zones of the park.

Veterinarians returner tens of tonnes of rotten meat to Poles

Since May till this time, the State Veterinary Administration has checked two hundreds of deliveries of poultry meat according to its regulations. On their base it has returned 46 tonnes of meat in total to the sender. In all cases it was dealt with meat from Poland. Since half of May, there is held a new Czech government regulation which tightens imports of animal products in the country. On base of the new government regulation since the half of May importers of animal origin foods have to inform in details about the delivery the supervisory authorities 24 hours at the latest before the goods arrival in the CR. If they do not do that, they can be fined with as many as one million crowns fine.

Germany will increase significantly sugar production; it will again exceed quotas

Germany will significantly increase production of refined sugar in a new growing season 2011/2012. According a chief of local association of sugar refineries, favourable weather as well as bigger areas for sugar beet growing will help Germans. So, Germany will again exceed quotas for sugar production set by the European Union. In the last season, Germany produced 3.44 million tonnes of refined sugar. "We expect good yields, so, the sugar production should not be significantly above the production volume in the last year", the chief of Sugar-refinery Union (WVZ) Dieter Langendorf said to the agency Reuters. According to him, it is too soon for exact estimations. In the last season, the sugar beet was harvested in Germany from 344 000 hectares and WZV estimates that the areas for sugar beet growing are not larger by five to eight percents.

SZPI forbade sale of smoked halibut from Poland

The State Agricultural and Food Inspection has found bacteria Listeria monocytogenes in officially taken sample of halibut smoked with could smoke. It is dealt with a products with trade name "Product from smoked fish HALIBUT", weight 113 g, charge 0707 PNT, date of production 7.7.2011, expiration date 1.8.2011, producer Almar Sp. from o.o. Kartuza, Poland. The product was sold by the chain store Kaufland. SPZI forbade to sell it and ordered its withdrawal. Now, the dangerous food is being withdrawn from shops of the distribution center of the company. SPZI cooperates on solution of this case with the State Veterinary Administration of the CR.

Sdílet Tweet 8+1 Tisk

Source: own processing

Figure 5: An example of a final paper from the agrarian WWW AGRIS portal.

Conclusion

An incessant increase in data volume in all spheres of human activity has been registered in recent years. It is necessary to describe these data in an efficient manner and to dispose of tools for their semantic description, storing, administration and processing or as the case may be also for their automated distribution. The key requirements for the metadata administration system in the environment of the Faculty of Economics and Management are its support of the OAI-PMH, the support of several metadata formats and thesauri and quality requirements in the shape of functionality, high level of reliability, applicability, sustainability and transferability. Transferability and in some cases applicability were the most serious problems of the analysed freely accessible SW for metadata administration and distribution. In order to cope with these problems the Department of Information Technologies developed an interactive application for metadata creation and administration with a possibility of their automated distribution via the OAI-PMH protocol.

The whole application construction to support the OAI-PMH has been proposed in a universal but homogenous way. Because local archives and repositories often provide rather different content and functions, the developed metadata repository functions as an independent separate web application with its own database.

The implemented WWW application is able

to respond automatically and flexibly to all six types of the OAI-PMH requests including reports on errors and exceptional conditions. Even the content of the agrarian WWW AGRIS portal was implemented into the system. In the case of the local archive of the agrarian WWW AGRIS portal the metadata are stored in the Dublin Core and VOA3R Metadata AP formats using the AGROVOC thesaurus. Owing to these internationally recognized metadata formats the content of the agrarian WWW AGRIS portal is easily accessible to the users and machines worldwide.

Acknowledgements

The knowledge and data presented in the present paper were obtained as a result of the Grant No. 20121044 of the Internal Grant Agency titled „Using Automatic Metadata Generation for Research Papers“.

Ing. Pavel Šimek, Ph.D.,
Department of Information Technologies, Faculty of Economics and Management,
Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Prague 6, Czech Republic
Phone: +420 2 2438 2050, E-mail: simek@pef.czu.cz

References

[1] Adly, N. Harvesting OAI-PMH repositories using adaptive synchronization. AEJ – Alexandria Engineering Journal, Vol. 48, Issue 1, January 2009, p. 95 – 106, ISSN 1110-0168.

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

[3] Berners-Lee, T. Metadata Architecture [online]. W3C, 199. Last edit 1998-08-27. Available: http://www.w3.org/DesignIssues/Metadata.html.

[4] Devarakonda, R., Palanisamy, G., Green, J. M., Wilson, B. E. Data sharing and retrieval using OAI-PMH. Earth Science Informatics. Vol. 4, Issue 1, March 2011, p. 1 – 5, ISSN 1865-0473.

[5] Diamantopoulos, N., Sgouropoulo, C., Kanstrantas, K., Manouselis, N. Developing a metadata application profile for sharing agricultural scientific and scholarly research resources. Springer-Verlag Berlin, Berlin. Published in Metadata and Semantic Research: Communication in Computer and Information Science, Vol. 240, p. 453 – 455, ISSN 1865-0929.

[6] The Dublin Core Metadata Initiative. Dublin Core Metadata Element Set [online]. Version 1.1, 11.10.2010. Available: http://dublincore.org/documents/dces.

[7] Grandmann, S. Cataloguing vs. Metadata : old wine in new bottles? In 64th IFLA General Conference, Amsterdam, Netherlands, August 16 - August 21, 1998 [online]. Available: http://archive.ifla.org/IV/ifla64/007-126e.htm.

[8] Kadury, A., Frank, A. J. Harvesting and aggregation of digital libraries using the oai framework. Published in 3rd International Conference on Web Information Systems and Technologies Proceedings. Volume WIA, 2007, p. 441 – 446. ISBN 978-972-8865-78-8.

[9] Protonotarios, V., Gavrilut, L., Athanasiadis, I., Hatzakis, I, Sicilia, M. A. Introducing a Content Integration Process for a Federation of Agricultural Institutional Repositories. Springer-Verlag Berlin, Berlin. Published in Metadata and Semantic Research: Communication in Computer and Information Science, 2011, Vol. 240, p. 467 – 477, ISSN 1865-0929.

[10] Sgouropoulou, C., Diamantopoulos, N., Kastrantas, K., Koutoumanos, A., Manouselis, N., Picarella, A. Specification of metadata profiles and mappings to existing technology (Part B: The VOA3R AgRes AP Metadata Terms). Deliverable number D 3.5 of Virtual Open Access Agriculture & Aquaculture Repository. Technological Educational Institute of Athens, 2011.

[11] Shreeves, S. L., Habing, T. G., Hagedorn, K., Young, J. A. Current developments and future trends for the OAI protocol for metadata harvesting. Library Trands. Vol. 53, Issue 4, Published: September 2005, p. 576-589, ISSN 0024-2594.

[12] Šimek, P., Vaněk, J., Očenášek, V., Stočes, M., Vogeltanzová, T. Using Metadata Description for Agriculture and Aquaculture Papers. Published in Agris on-line Papers in Economics and Informatics. Czech University of Life Sciences Prague, Faculty of Economics and Management, Prague, 2012. pp 79 – 80. Available: http://online.agris.cz/files/2012/agris_on-line_2012_4_simek_vanek_ocenasek_stoces_vogeltanzova.pdf, ISSN 1804-1930.

[13] Šimek, P., Vaněk, J., Jarolímek, J., Stočes, M., Vogeltanzová, T. Using metadata formats and AGROVOC thesaurus for data description in the agrarian sector. Published in Plant, Soil and Environment. Czech Academy of Agricultural Sciences, Prague, 2013, p. 378 – 384. Available: http://www.agriculturejournals.cz/publicFiles/97726.pdf, ISSN 1805-9368.

[14] Sini, M., Lauser, B., Salokhe, G., Keizer, J., Katz, S. The AGROVOC Concept Server: Rationale, goals and usage. Emerald Group Publishing Ltd., 2008. Published in Library Review, Vol. 57, p. 200-212, ISSN 0024-2535.

[15] Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S. Reengineering thesauri for new applications: The AGROVOC example. British Computer Society, 2004. Journal of Digital Information, Vol. 4, p. 26, ISSN 1368-7506.