Enriched Data Sharing Methodology

Michal Stočes, Eva Kánská, Pavel Šimek, Jiří Vaněk

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

Abstract

The aim of this article is to propose a methodology for improving the sharing of data between applications that support scientific activity, which are focused on agriculture, aquaculture, rural development, etc. The presented methodological approach is referred to as Enriched Data Sharing Methodology (EDSM). The presented methodology is based on two analyzes. The analysis of the data formats used for the metadata description of digital objects and the description of their mutual relations. And analysis of dictionaries of controlled descriptors.

The article presents part of the results of author’s dissertation thesis.

Keywords

Metadata, application profile, social network service for scientists, data sharing, enriched data.

Stočes, M., Kánská, E., Šimek, P. and Vaněk, J. (2018) “Enriched Data Sharing Methodology", *AGRIS on-line Papers in Economics and Informatics*, Vol. 10, No. 2, pp. 85-92. ISSN 1804-1930. DOI 10.7160/aol.2018.100208.

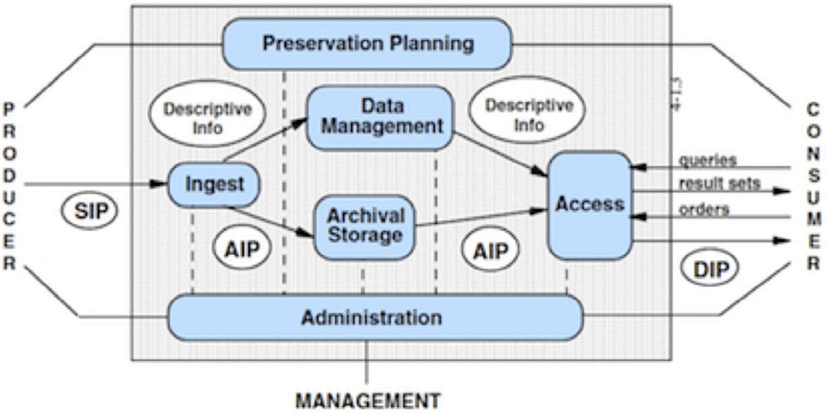
Introduction

Social networks have been the phenomenon of recent times in the scientific community. These sector-specific social networks have significantly changed the form of communication and knowledge exchange. Social networking applications such as Social network service for scientists ResearchGate have been established to support research activities. This new communication platform also affected science and research. Most scientists are forced to work with multiple applications. In each application, the user is prompted to create a profile and upload metadata of scientific publications. Each application creates a specific identifier for the user and his publications. It is difficult for the enriched data created in the system to be exported or transferred from one application to another. (Al-Aufi and Fulton, 2013) (Asmi and Margam, 2018)

The term digital library is very broad, and its definition is inconsistent. In the literature, the following definitions exist: "A managed collection of information along with corresponding services, the information being stored in digital form and accessible through the network. (Arms, 2000) An integrated system, including a set of electronic information resources and services to retrieve, process, search, and use information stored on that system. Digital libraries are accessed through computer networks. The purpose of building a digital library is to give users the opportunity to have unified access to digital or digitized documents, or secondary information about printed primary resources stored in the library. "Organizations providing resources (including dedicated staff), allowing for selection, structuring and accessing digital works collections, further distribute these works, maintain their integrity, and preserve them for the long term - all with regards to the easy and economical use by a particular community or set of user communities" (Van de Sompel and Hochstenbach, 1999).

The most elaborate general architecture of digital libraries is Kahn and Wilenski's architecture (Kahn and Wilensky, 2006). The term digital library is closely related to the term digital repository. According to some authors, there is no difference between these terms, some authors associate the concept of digital repository with specific institutions and the principle of open sharing of these data. The digital repository is addressed by the Open Archival Information System (OAIS), which has been accepted as a standard ISO 14721: 2003. The principle of the OAIP model is illustrated in Figure 1 (Epple et al., 2017).

The aim of this article is to propose a methodology for improving the sharing of data between



Note: Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP)
Source: Epple et al. (2017)

Figure 1: OAIS reference model.

applications that support scientific activity. The presented methodological approach is referred to as Enriched Data Sharing Methodology (EDSM). The article presents part of the results of author's dissertation thesis. It follows the findings of Stočes et al. (2017).

Materials and methods

Two improvement methods were used in the Enrichment Data Sharing Methodology (EDSM). The first method is described by the PDSA cycle, also known as Deming Wheel. It is based on the English model "plan-do-study-act". It is a method of gradual improvement of many fields, including information technology. It consists of the following phases: (Rao et al., 1996; Deming, 2016):

- Phase 1 - P (plan) - problem identification (intent),
- Phase 2 - D (do) - implementing the plan,
- Phase 3 - S (study) - verification of the result of the implementation compared to the original plan,
- Phase 4 - A (act) - modification of intent and own implementation based on verification and implementation of improvements to practice, implementation of the best solution.

In the context of the PDSA cycle, the second method - the seven-step method - is also often mentioned (Rao, et al., 1996) (Table 1).

PDSA phases	Seven-step method
PLAN	Identification of the problem and its clear definition
	Analysis of current state
	Identification of possible causes of the problem
DO	Planning and implementation of the solution
STUDY	Evaluating results
ACT	Standardizing the solution
	Evaluating the solution and proposing plans and provisions for the future

Source: Rao et al. (1996)

Table 1: Relation between PDSA and seven-step method.

Repositories that focus on long-term storage and access to digital information seeks the status of a trusted long-term repository. ISO 114721:2003, resp. 2012 is the reference model of OAIS, a standard that defines the activities of the long-term repository, its objectives, and introduces the basic terminology and information model. ISO 14721:2003 defines what metadata should be stored by a long-term repository. ISO standard 16363:2012 (follow-up to the Trusted repository audit checklist) is a means of certifying a trusted long-term repository. The repositories that do not store OAIS metadata and do not publish the documentation required by ISO 16363 cannot be considered as trusted repositories in the long-term. (Šimek et al., 2013; Stočes et al., 2018; Planková, 2008; Hodge et al., 2008).

Institutional repositories, including local repositories, collect digital objects that have been created within the institution that established the repository. They are used mainly at universities and research institutions. The content type is limited only by the focus of the founding institution.

Central repositories or subject repositories are focused on a particular science field. They focus on collecting (aggregating) the documents from the institution or even written by independent scientists within the subject frame. Central repositories provide search services over metadata acquired by various local repositories. (Müller and Adelhard, 2002).

Central repositories include repositories that aggregate data according to a particular data type. These repositories include, for example:

- Repository of "Gray Literature" of the National Repository of Gray Literature (NUŠL).
- COnnecting Repositories (CORE) that aggregate hundreds of open-access repositories from different countries.
- Europeana Archive containing scanned artwork, films and books.
- Repository dblp aggregating the metadata of articles and contributions from the field of computer science - http://dblp.uni-trier.de/

Applications using metadata

Metadata of digital artefacts (objects) from digital libraries are used by science support applications. These are primarily web applications that are divided into the following groups:

- Social network services for scientists,
- Reference management software,
- Web search engines of scientific work.

(Thanos et al., 2017).

Social network services for scientists are social networks developed for scientists and serves to support their activities - primarily to promote mutual communication and knowledge sharing. Particularly younger users use social media to communicate and share their knowledge. According to Stočes (2015), it is recommended to integrate some social networking features directly into learning management systems (LMS), which then serve as a communication tool between lecturers and students or even students among themselves. And it will enable students to get the latest knowledge in the area. These networks include, for example, *ResearchGate*, *academia.edu* or *VOA3R (Virtual Open Access Agriculture & Aquaculture Repository) portal* (Gemma and Ángel, 2013).

Reference management software are applications used to manage references or for personal bibliographic management. These software packages usually consist of a database that can provide full bibliographic links and a system for generating selective lists of articles in various formats that are required by publishers and scientific journals. Modern link management packages can usually be integrated with word processors, so a list of references in the appropriate format will automatically be generated when writing an article, thereby reducing the risk that the quoted source will not be included in the list of links. These systems can also link metadata to specific profiles of authors. These are primarily commercial applications. The most important applications of this type include: *Mendeley, EndNote* or *REFWORKS* (Ortega, 2015).

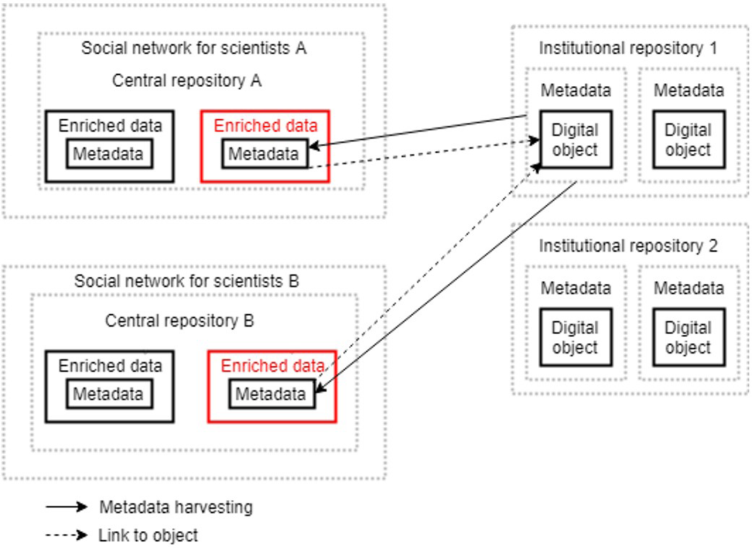
There is a whole range of *Web search engines of scientific work*, most of which are focused on a specific scientific field.

Google Scholar is a freely-accessible web multidisciplinary search engine that indexes full text or metadata of professional literature in a variety of publishing formats. The Google Scholar Index, released in the beta version in November 2004, contains the most reviewed online academic journals and books, conference articles, theses and dissertations, prepress, abstracts, technical reports, and other professional literature, including court testimonials and patents. Google Scholar is the most popular and most comprehensive application in this category. Google Scholar allows you to link your own google profile with indexed articles (Masner et al., 2016).

Results and discussion

Social networking applications for scientists allow users to search for repository objects and add additional data. These data can be called "enriched" data. Enriched data is stored within the social network and can only be accessed through that specific network. The principle is illustrated in Figure 2. Enriched data can be classified into two groups, namely linking and other metadata. The structure and function of other metadata is created by each social network separately. Examples of such data may be comments, ratings, etc. Linking metadata includes the following links:

- Digital artefact - Person (author, co-authors)
- Digital artefact - Digital artefact (citation, reference)



Source: Author

Figure 2: Relation between Institutional repositories and Social networks for scientists.

The new proposed methodological approach is based on the analytical findings. It aims to improve the metadata transfer, which is enriched by the social network services for scientists. The new methodology is referred to as Enriched Data Sharing Methodology (EDSM). The presentation of the design uses the UML (Unified Modeling Language) specification diagram. The issue of metadata descriptions that are used in the methodology is dealt with by Stočes (2017).

EDSM methodology formulation

The presented Enriched Data Sharing Methodology consists from two stages:

1. Identifying metadata describing digital artefacts (objects).
2. Creating an application profile.

Stage 1 - Identifying metadata

This preparatory stage is focused on identifying and categorizing the structure of the data model (metadata structure) of the application, to which elements will be assigned to the next stage. Stage 1 consists of two phases. Identification of the data model (metadata structure) is based on the knowledge gained from the metadata format analysis.

Phase 1 - Identifying primary metadata

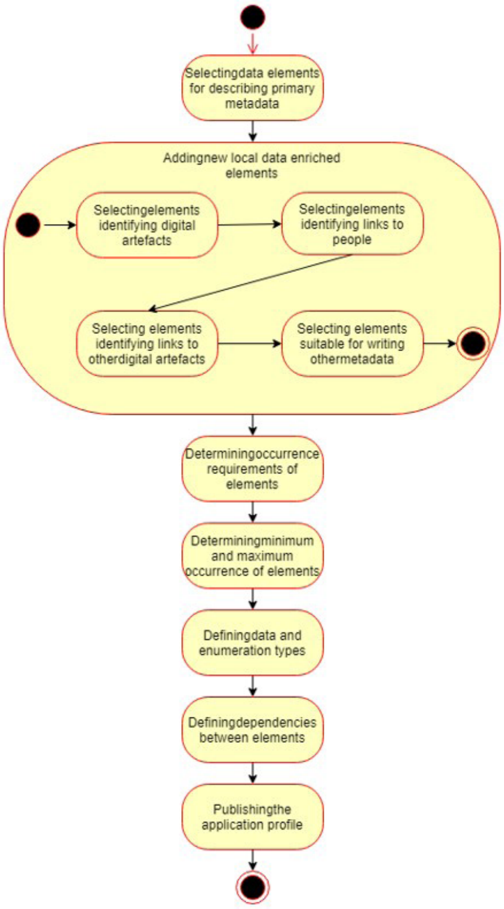
The initial phase identifies metadata that did not result from social networking applications for scientists. These are metadata descriptions made when publishing digital artefacts, e.g.: name, authors, licenses, etc. These elements are also referred to as primary elements. Each local repository (publisher) has a different metadata model, so it is necessary to identify primary elements as an intersection of all metadata models whose digital artefacts are in the application.

Phase 2 - Identifying enriched data

In the second phase, the metadata created by social networking applications for scientists is identified. Firstly, metadata specifying links to author profiles or other digital artefacts (quotes, references, projects) is identified. Next is the identification of other enriched data specific to given application, examples of which may be various comments, ratings, terms from specific thesaurus (e.g. agriculture thesaurus AGROVOC) etc.

Stage 2 – Creating an application profile

Stage 2 consists of seven phases, based on CEN/ISSS and the Singaporean application profile creation framework. Is uses the DC, MODS, and LOM metadata element names for description (Stočes et al., 2017). The dependency of the individual phases is shown in the UML task diagram (see Figure 3), the term "phase" is used for different activities within the diagram. The resulting application profile can be presented in a table format describing the individual elements and their properties or as XML template written in XSD or RDF format. (Carey et al., 2012; Taheri and Hariri, 2012).


Selectingdata elements for describing primary metadata
Addingnew local data enriched elements
Selectingelements identifying digital artefacts
Selectingelements identifying links to people
Selecting elements identifying links to otherdigital artefacts
Selecting elements suitable for writing othermetadata
Determiningoccurrence requirements of elements
Determiningminimum and maximum occurrence of elements
Definingdata and enumeration types
Definingdependencies between elements
Publishingthe application profile

Source: Author

Figure 3: EDSM – Activity diagram of creating application profile.

Phase 1 - Selecting data elements for describing primary metadata

First, elements will be selected to describe primary (original) metadata from which a new application profile will be created. This phase will use elements from the Dublin Core namespace.

Phase 2 - Adding new local data enriched elements

New local elements (elements created by social networking activities) of enriched data will be added. This phase can be divided into four steps:

Phase 2a - Selecting elements identifying digital artefacts

The data part identifying the digital artefact itself is created using the MODS standard by using the identifier element.

Phase 2b - Selecting elements identifying links to people

The data part identifying the link to a person (link: digital artefact - person) is created by combining elements of the namespaces of LOM and MODS. These are the elements: lom:lifeCycle, lom:contribute, lom:role, lom:source, lom:value, and mods:identifier. The most important of these links is the identification of the author of the digital artefact.

Phase 2c - Selecting elements identifying links to other digital artefacts

The data part identifying the links to other digital artefacts (link: digital artefact - other digital artefacts) is created using the MODS schema. The following elements are used: relatedItem and identifier. Links to other digital artefacts include, above all, the identification of citations and references.

Phase 2d - Selecting elements suitable for writing other metadata

The structure of the description of other metadata is designed based on their nature, depending on given social network, by using selected elements from LOM, DC, MODS or their combinations, or by defining new elements. An example might be an extension of the keywords for a dictionary item. Or, if the social network enriches some elements of education, the relevant specific elements of the LOM standard can be used.

Phase 3 - Determining occurrence requirements of elements

This step determines the requirements for occurrence of individual elements:

- mandatory
- recommended
- conditional
- optional

Mandatory elements must include at least the name of the digital artefact (<dc: title>) and the type (<dc: type>) whenever describing text, visualization, sound, etc.

Phase 4 - Determining minimum and maximum occurrence of elements

In addition, it is necessary to determine the minimum and maximum number of occurrences of individual elements. For elements identifying links to other artefacts or individuals, it is recommended not to restrict the upper limit of occurrence.

Phase 5 - Defining data and enumeration types

The definition of data types and enumerations of values is based on chosen metadata standard, or a new data enumeration type can be created.

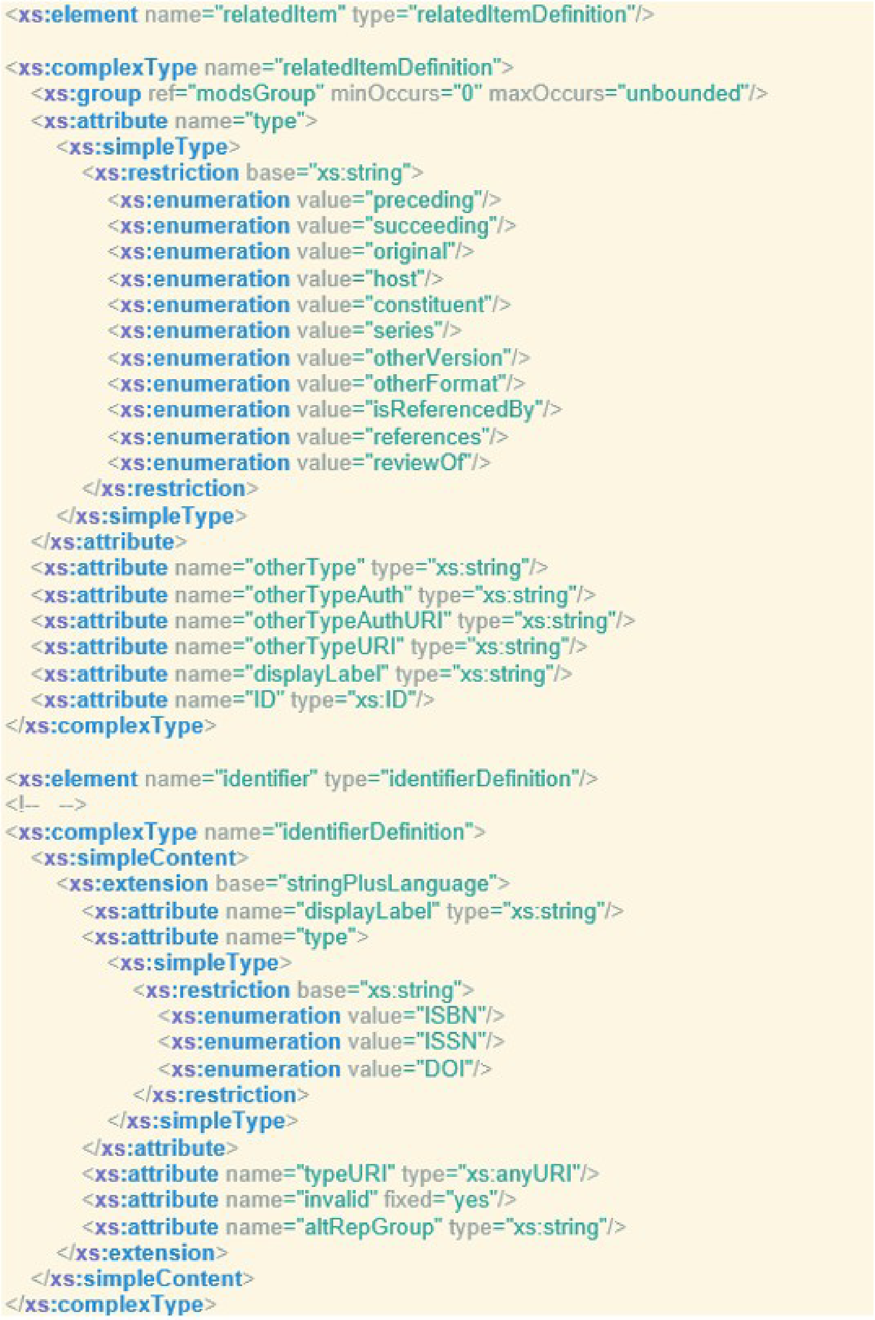
Phase 6 - Defining dependencies between elements

The final stage before publishing the application profile is to define the dependencies between the individual elements. Defining dependencies serves primarily to explain the logic dependence of some elements and to reduce duplicate entries.

Phase 7 - Publishing the application profile

The final stage is to create an application profile, represented as an XML schema, an RDF format, or a text list containing the element names, their properties and constraints.

The methodology was verified by compiling the application profile and transforming dozens of scientific records into the desired format. Figure 4 shows a part of the XML schema created by EDSM and demonstrates a description of a digital object (references, citation).



```
<xs:element name="relatedItem" type="relatedItemDefinition"/>

<xs:complexType name="relatedItemDefinition">
  <xs:group ref="modsGroup" minOccurs="0" maxOccurs="unbounded"/>
  <xs:attribute name="type">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="preceding"/>
        <xs:enumeration value="succeeding"/>
        <xs:enumeration value="original"/>
        <xs:enumeration value="host"/>
        <xs:enumeration value="constituent"/>
        <xs:enumeration value="series"/>
        <xs:enumeration value="otherVersion"/>
        <xs:enumeration value="otherFormat"/>
        <xs:enumeration value="isReferencedBy"/>
        <xs:enumeration value="references"/>
        <xs:enumeration value="reviewOf"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="otherType" type="xs:string"/>
  <xs:attribute name="otherTypeAuth" type="xs:string"/>
  <xs:attribute name="otherTypeAuthURI" type="xs:string"/>
  <xs:attribute name="otherTypeURI" type="xs:string"/>
  <xs:attribute name="displayLabel" type="xs:string"/>
  <xs:attribute name="ID" type="xs:ID"/>
</xs:complexType>

<xs:element name="identifier" type="identifierDefinition"/>
<!--  -->
<xs:complexType name="identifierDefinition">
  <xs:simpleContent>
    <xs:extension base="stringPlusLanguage">
      <xs:attribute name="displayLabel" type="xs:string"/>
      <xs:attribute name="type">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:enumeration value="ISBN"/>
            <xs:enumeration value="ISSN"/>
            <xs:enumeration value="DOI"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:attribute>
      <xs:attribute name="typeURI" type="xs:anyURI"/>
      <xs:attribute name="invalid" fixed="yes"/>
      <xs:attribute name="altRepGroup" type="xs:string"/>
    </xs:extension>
  </xs:simpleContent>
</xs:complexType>
```

Source: Author

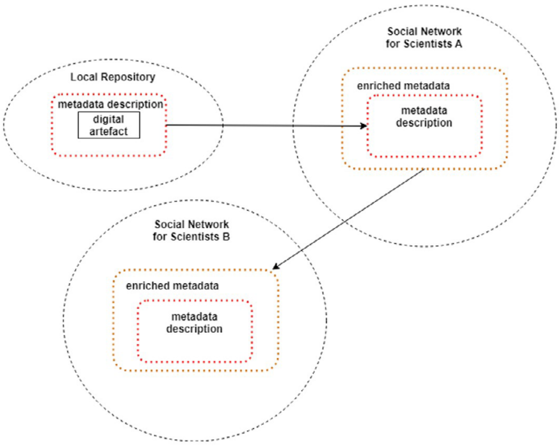
Figure 4: XML schema description of a digital object (references, citation).

Conclusion

The proposed method is designed for social network services for scientists who also publish their results from an agriculture area, forestry, aquaculture or rural development. It is used to create the application profile for the metadata of the digital artefact (object) that occurs in the given application. The application profile will allow social networking application users to import or export the metadata describing a digital artefact (object) from/to the application.

The main benefit of the EDSM methodology is to simplify the transfer of metadata descriptions between social networking applications for scientists. Figure 5 describes the principle of transferring metadata from a local repository to a social networking application for scientists and the subsequent transfer of enriched metadata to another application. Transferring between a local repository and an application may occur by automatic retrieval (e.g., OAI-PMH) or by writing the data by users. Metadata from application A is exported in XML format and then transferred to application B. Transmission can be done manually by the user or using the OAI-PMH protocol.

Using this methodology, an application profile is created to create specific metadata records of specific digital artefacts. These artefacts can be easily transformed and used to transfer data among other applications to support scientific work. The resulting record is expected to be both human-readable and machine-readable.



Source: Author

Figure 5: Transfer of metadata between applications.

Acknowledgements

The knowledge and data presented in the present paper were obtained as a result of the Grant No. 20181014 of the Internal Grant Agency titled "Permanent and globally unique identification of author entities".

Corresponding authors:
Ing. Michal Stočes, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Prague, Czech Republic
Phone: +420 224 382 277, E-mail: stoces@pef.czu.cz

References

[1] Al-Aufi, A. S. and Fulton, C. (2013) "Use of Social Networking Tools for Informal Scholarly Communication in Humanities and Social Sciences Disciplines", *3rd International Conference on Integrated Information (IC-ININFO)*, Sep 5-9, 2013, Prague, Czech Republic, Vol. 147, pp. 436-445. DOI 10.1016/j.sbspro.2014.07.135.

[2] Asmi, N. A.and Margam, M. (2018) "Academic social networking sites for researchers in Central Universities of Delhi: A study of ResearchGate and Academia", *Global Knowledge, Memory and Communication*, Vol. 67, No 1/2, pp. 91-108. DOI 10.1108/GKMC-01-2017-0004.

[3] Arms, W. Y. (2000) "*Digital Libraries*", Cambridge: MIT Press. ISBN 0-262-01880-8.

[4] Carey, M. J., Onose, N. and Petropoulos, M. (2012) "Data Services", *Communications of the ACM*, Vol, 55, No. 6, pp. 86-97. DOI 10.1145/2184319.2184340.

[5] Deming, E. W. (2018) "PDSA Cycle", The W. Edwards Deming Institute, [Online]. Available: https://deming.org/explore/p-d-s-a [Accessed: 1 Feb. 2018].

[6] Epple, U., Mertens, M., Palm, F. and Azarmipour, M. (2017) “Using Properties as a Semantic Base for Interoperability”, IEEE Transactions on Industrial Informatics, Vol. 13, No 6, pp. 3411-3419. DOI 10.1109/TII.2017.2741339.

[7] Gemma, N. and Ángel, B. (2013) "Use of social networks for academic purposes: a case study", *The Electronic Library*, Vol. 31, No. 6, pp. 781-791. DOI 10.1108/EL-03-2012-0031.

[8] Hodge, G., Templeton, C. and Allen, R. (2008) “A metadata element set for project documentation“, *Science and Technology Libraries*, Vol. 25, No. 4, pp. 5-23. E-ISSN 1541-1109, ISSN 0194-262X. DOI 10.1300/J122v25n04_02.

[9] Kahn, R. and Wilensky, R. A. (2006) "A framework for distributed digital object services", *International Journal on Digital Libraries*, Vol. 2, No. 6, pp. 115-123. E-ISSN 1432-130, ISSN 1432-5012. DOI 10.1007/s00799-005-0128-x.

[10] Masner, J., Vanek, J. and Jarolimek, J. (2016) "Prototype of a Content Creation and Updating Application Module for Agrarian Sector and Regional Development", *Scientific conference Agrarian Perspectives - Global and European Challenges for Food Production, Agribusiness and the Rural Economy*, Sep 14-16, 2016, Czech University of Life Sciences Prague, pp. 206-213. ISSN 2464-4781.

[11] Müller, T. H. and Adelhard, K. (2002) "A web-based central diagnostic data repository", *Studies in Health Technology and Informatics*, Vol. 90, pp. 246-250. ISSN 0926-9630. DOI 10.3233/978-1-60750-934-9-246.

[12] Ortega, L. S. (2015) "Disciplinary differences in the use of academic social networking sites", *Online Information Review*, Vol. 39 No. 4, pp. 520-536. ISSN 1468-4527. DOI 10.1108/OIR-03-2015-0093.

[13] Planková, J. (2008) “Technická řešení pro otevřený přístup“ (in Czech), *Ikaros*, Vol. 12, No. 2. [Online]. Available: http://ikaros.cz/node/12740 [Accessed: 5 Feb. 2018]. ISSN 1212-5075.

[14] Rao, A., Carr, L. P., Dambolena, I., Kopp, R. J., Martin, J., Rafii, F. and Schlesinger, P. F. (1996) “Total Quality Management: A Cross Functional Perspective“, John Wiley & Sons, pp. 656. ISBN 978-0-471-10804-7.

[15] Stočes, M., Masner, J. and Jarolímek, J. (2015) “Mitigation of Social Exclusion in Regions and Rural Areas – E-learning with Focus on Content Creation and Evaluation”, *AGRIS on-line Papers in Economics and Informatics*, Vol. 7, No. 4, pp. 143 - 150, ISSN 1804-1930.

[16] Stočes, M., Šimek, P. and Pavlík, J. (2017)“ Metadata Formats for Data Sharing in Science Support Systems“ , *AGRIS on-line Papers in Economics and Informatics*, Vol. 9, No. 3, pp. 61-69. ISSN 1804-1930. DOI 10.7160/aol.2017.090306.

[17] Stočes, M., Šilerová, E., Vaněk, J., Jarolímek, J. and Šimek, P. (2018) “Possibilities of Using Open Data in Sugar & Sugar Beet Sector”, *Listy cukrovarnické a řepařské*, Vol. 134, No. 3, pp. 117-121. ISSN 1210-3306.

[18] Šimek, P., Vaněk, J., Jarolímek, J., Stočes, M. and Vogeltanzová, T. (2013) “Using metadata formats and AGROVOC thesaurus for data description in the agrarian sector“, *Plant, Soil and Environment*, Vol. 59, No. 8, pp. 378-384. ISSN 1214-1178. DOI 10.17221/261/2013-PSE.

[19] Taheri, S. M. and Hariri, N. A. (2012) "A comparative study on the indexing and ranking of the content objects including the MARCXML and Dublin Core's metadata elements by general search engines", *The Electronic Library*, Vol. 30, No. 4, pp. 480-491. DOI 10.1108/02640471211252193.

[20] Thanos, C., Klan, F., Kritikos, K. and Candela, L. (2017) ”White Paper on Research Data Service Discoverability”, *Publications*, Vol. 5, No. 1. ISSN 2304-6775. DOI 10.3390/publications5010001.

[21] Van de Sompel, H. and Hochstenbach, P. (1999) ”Reference Linking in a Hybrid Library Environment Part 3: Generalizing the SFX solution in the 'SFX@Ghent & SFX@LANL' experiment”, *D-Lib Magazine*, Vol. 5, No. 10. ISSN 1082-9873.