ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter214892020.pdf
https://doi.org/10.30534/ijeter/2020/214892020

Development of the Methodology for Metadata Extraction from Documents in the Course of Information System Integration Using the Ontological Model of the Subject Field

Sadirmekova Zh.B.[1], Tusupov J.A. [2], Sambetbayeva M.A. [3], Nurgulzhanova A.N.[4], Doshtaev K.Zh.
[1] L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan, janna_1988@mail.ru
[2] L.N. Gumilyov Eurasian National University, Nur-Sultan, **5**Kazakhstan, tusupov@mail.ru
[3] L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan, madina_jgtu@mail.ru
[4] Kazakh Academy of Transport and Communications named after M. Tynyshpayev, Almaty, Kazakhstan, nurgulzhanova@mail.ru
[5] Kazakh Academy of Transport and Communications named after M. Tynyshpayev, Almaty, Kazakhstan, kuntu@inbox.ru

ABSTRACT
The article deals with the development of an information system model designed for scientific and educational activities; discusses the state of ontologies as a semantic integration tool; considers the issue of the formation of metadata element sets; and proposes an approach to automatizing the collection of information on scientific activities in the given knowledge domain, which would combine the metasearch and information extraction methods based on ontologies.

Key words: Information System Metadata Extraction, Ontology, OWL, Semantic Integration

1. INTRODUCTION

Currently, two information system types are developed and used (here and after referred to as "IS"): documentary and factographic. Documentary IS are repositories of documents containing metadata, by means of which documents are classified and searched for. Factographic IS accumulate and store data as a set of instances of a single or several types of structural elements (information entities), each of which instances or a certain set of them represent data on a certain fact, standalone event, in isolation from other data and facts.IS that include the features of both information system types mentioned above are becoming the most demanded means of information support of scientific and educational activities. These IS help meet the information needs of skilled users in conformity to the scheme "document – consideration – fact." Note that this pattern complies with the RDF scheme of associated data [1].Interoperability is the key requirement for the IS designed to support scientific and educational activity [2].The interoperability of any IS is understood as the extent of its correlation with other information systems, including with the human. It is impossible to ensure the interoperability of systems without strict abidance by the respective international standards and guidelines.

In this case, the standards should also apply to:

- data access protocols and interfaces;
- search engine languages and interfaces;
- data representation schemes and formats;
- homogenous data visualization interfaces;
- information encoding rules;
- data access control rules.

The authors aim at developing information resources and services that would meet the demands of even the least numerous consumer groups (down to a single user). This aim corresponds with a user friendly tendency for software development, e.g. the successful integration of heterogeneous data with preliminary check of its genuineness [3] or developing algorithms against website phishing to ensure data safety [4]. When developing such software bundles, the authors consider the following properties of a modern IS:

- ensuring the preservation of knowledge accumulated in institutes and scientific organizations;
- provision of both academic researchers and students with the access to scientific knowledge and information;
- ensuring a multi-faceted information support for scientific research based on modern networking technology;
- guarantee of prompt utilization of modern knowledge;
- development ofa toolset for analytic research of the processes of scientific knowledge production, scientific information, and communication of the academic community.

2. THEINFORMATION SYSTEM MODEL

ISNOD is an information system that provides for the systematization and integration of scientific knowledge and information resources, substantive effective access to them (search and navigation) and their intellectual processing tools. One of the key problems of the article is the construction of a model that would accurately represent the software system. The model construction involves ontology [5]. Several approaches to the definition of this concept are known, but no commonly accepted interpretation has been found so far, as it is convenient to interpret this term differently depending on each particular task: from loose definition to the description of ontologies in logical and mathematical terms and constructs. In our turn, we will consider ontology as the formal specification of separable conceptualization, which takes place in a certain context of the subject field. We understand conceptualization as not only collection of concepts, but also all the related information: properties, relations, restrictions, axioms, and statements required to describe and solve problems in the chosen subject field.

Ontologies are the new intellectual means for searching resources in the Internet, special innovative methods to represent and process knowledge and inquiries. They can accurately and effectively describe the semantics of data for a certain subject field and address the issue of concept incompatibility and inconsistency. Ontologies feature their own processing tools (logical conclusion) that correspond to the problems of semantic information processing. Thanks to ontologies, the user using a search engine will be able to obtain resources that are semantically relevant to the query. Therefore, ontologies have become widely popular in solving the problems of knowledge representation and engineering, semantic integration of information resources, information search, etc.

ISNOD stores the information on employees and their publications, conferences and projects the researchers participated in, as well as the information on organizations, related to particular scientific projects, various scientific publications, and others. Information entities describe the main classes of the entities of the scientific information space, such as Organization, Person, Scientific Activity, Publication, ScientificEvents, Training Course, Subdiscipline, Competence, Geographic Location, Conference Proceedings etc., as well as associations between them.

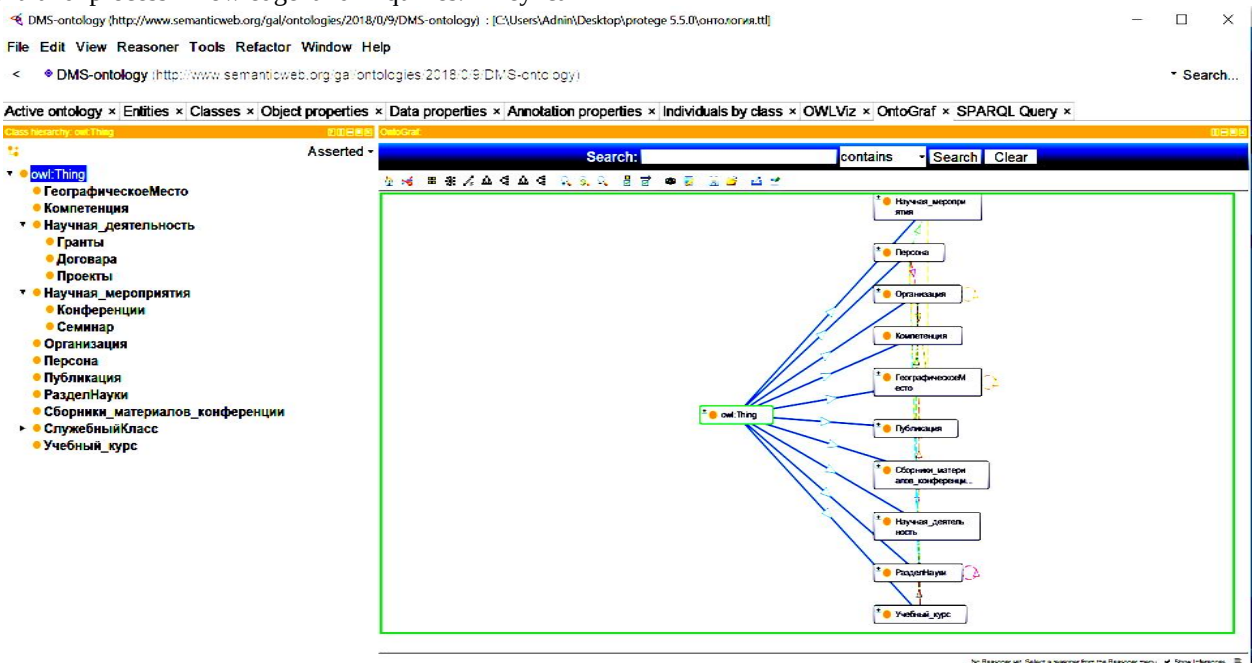

Figure 1: Class visualization

Ontology can be visualized with Protégé (Figure 1). OntoGraf is a plugin used to visualize various entities of the software system and their associations. OntoGraf can be used to activate the search functionality that enables searching for any function, as well as a function and the components associated with it. We can choose any function and expand it to visualize all of its variants and the components that implement it. Figure 2 describes the class of an information space entity – Scientific Activity – as well as its functional properties (title, number, start date, end date), and its correlations with other classes: Subdiscipline, Person, Organization, Publication. The Scientific Activity class reflects the information on the results of a project, grant, agreement, and the like.

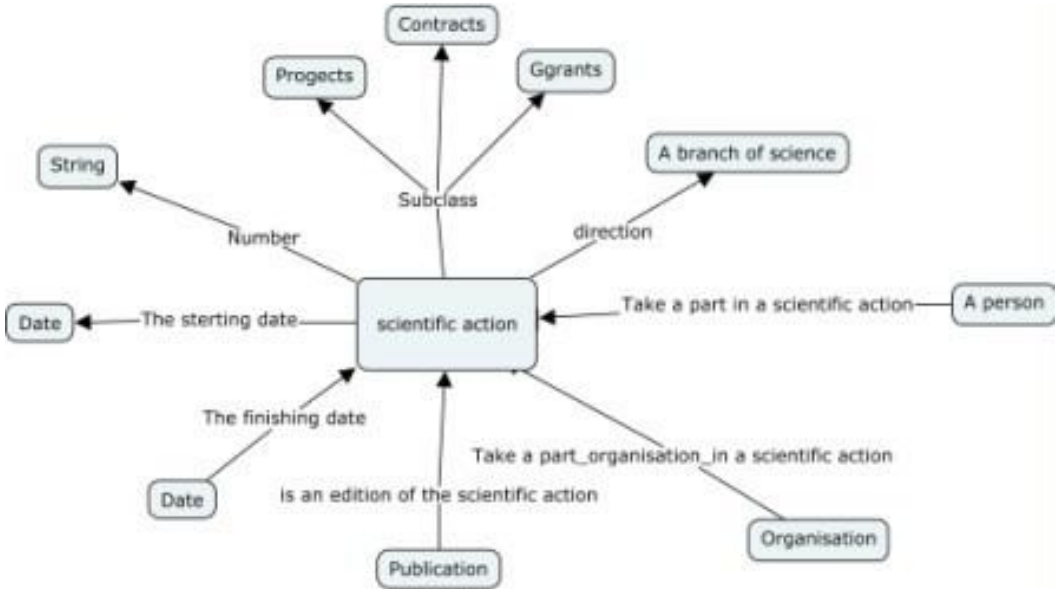

Figure 2: Description of the *Scientific Activity* class

Below is a fragment of the OWL implementation of the content used to describe *Scientific Activity* in the Turtle format.
The set of properties of the *Scientific Activity* class (*Object Property* and *Data type Property* according to the OWL language notation) is as follows:

:**hasDirection**_Subdiscipline*rdf:typeowl:ObjectProperty*;
rdfs:domain : Scientific Activity;
rdfs:range :Subdiscipline;
:participates_Organization_inScientific_Activity*rdf:typeowl:ObjectProperty*;
rdfs:domain : ScientificActivity;
rdfs:range :Organization;
:**is**Publication_ofScientific_Activity*rdf:typeowl:ObjectProperty*;
rdfs:domain : ScientificActivity;
rdfs:range :Publication;
:**participatesIn**_Person_inScientific_Activity*rdf:typeowl:ObjectProperty*;
rdfs:domain :ScientificActivity;
rdfs:range :Person;
:**participatesIn**_Person_inScientific_Activity*rdf:typeowl:ObjectProperty*;
rdfs:domain : ScientificActivity;
rdfs:range : Organization;
:**Scientific Activity_Title***rdf:typeowl:DatatypeProperty*;
rdfs:domain :ScientificActivity;
rdfs:rangerdfs:Literal;
:**Scientific Activity_Number**rdf:typeowl:DatatypeProperty;
rdfs:domain :ScientificActivity;
rdfs:rangerdfs:string;

Further, the *Functional Property* entity is used to provide each instance (representative) of the *Scientific Activity* class with the requirement to specify only one start date and one end date:

:**ScientificActivity_EndDate***rdf:typeowl:DatatypeProperty, owl:FunctionalProperty*;
rdfs:domain :ScientificActivity ;
rdfs:rangerdfs:Date;
:**ScientificActivity_StartDate***rdf:typeowl:DatatypeProperty, owl:FunctionalProperty*;
rdfs:domain :ScientificActivity ;
rdfs:rangerdfs:Date;

Figure 3 below schematically presents the*Person* class. Persons are the actors/individuals, who have the following properties:*surname, name, patronymic, initials, the list of WOS publications, the list of Scopus publications, the list of RSCI publications, the list of Mathnet.ru publications.* Besides, this class is associated with other classes, namely: *Study Course, Scientific Events, Scientific Activity, Organization, Publication*. For example: *teaches Person Study Course, participates Person in Scientific Activity, works Person at Organization, participates Person in Scientific Event, is Author of Publication.*

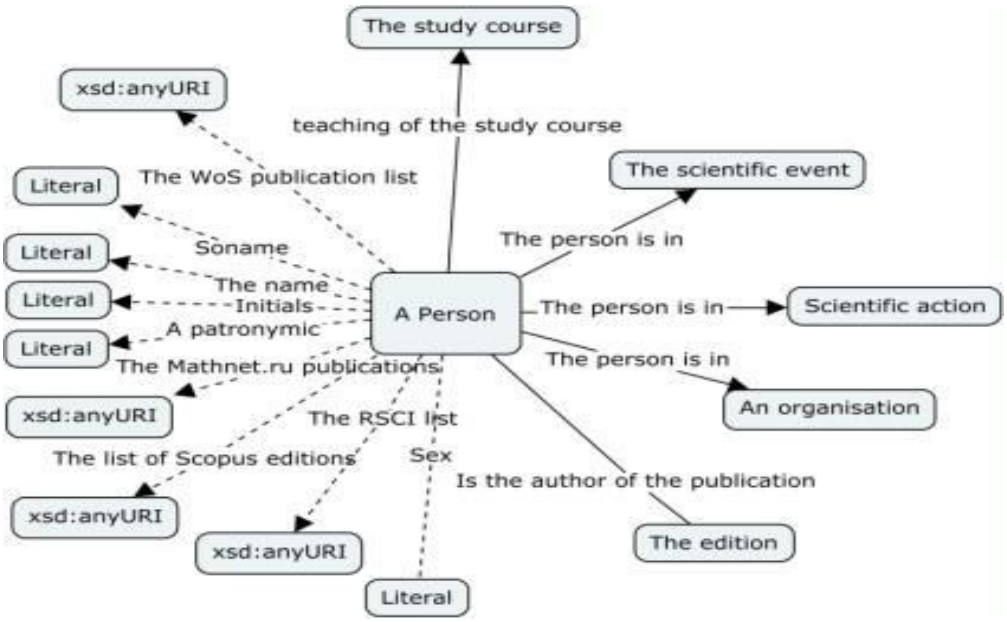

Figure 3: Description of the *Person* class

The set of properties of the *Person* class (*Object Property* and *Datatype Property* according to the OWL language notation) is as follows:

:**teaches**_Person_StudyCourse:*typeowl:ObjectProperty*;
rdfs:domain :Person;
rdfs:range :StudyCourse;
:participates_Person_inScientific_Event:*rdf:typeowl:ObjectProperty*;
rdfs:domain :Person;
rdfs:range :Scientific Event;
:**participates_**Person_inScientific_Activity: *rdf:typeowl:ObjectProperty*;
rdfs:domain :Person;
rdfs:range: Scientific_Activity;
:is_Author_ofPublication: *rdf:typeowl:ObjectProperty*;
rdfs:domain :Person;
rdfs:range : Publication;
:**teaches_**Person_StudyCourse:*rdf:typeowl:ObjectProperty*;
rdfs:domain :Person;
rdfs:range :Study Course;
:**Person_Surname***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs:Literal;
:**Person_Name** rdf:typeowl:DatatypeProperty;
rdfs:domain :Person;
rdfs:rangerdfs:Literal;
:**Person_Patronymic***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs:Literal;
:**Person_List of Publications _Mathnet.ru** *rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs:xsd:anyURI;
:**Person_List of Publications _inScopus***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs:xsd:anyURI;
:**Person_List of Publications __inWOS** *rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs:xsd:anyURI;
:**Person_List of Publications __inRSCI***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs:xsd:anyURI;
:**Person_Gender***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Person;
rdfs:rangerdfs: Literal.

The *Publication* class correlates with the following classes: *Person, Scientific Activity, Scientific Events,*and *Subdiscipline*. Besides, it has the following properties: *Document Language, Publication Date, Title, Edition, Keywords, URI, Abstract, Bibliographic Reference,* etc.
The set of properties of the *Publication* class (*ObjectProperty*and *Datatype Property* according to the OWL language notation) is as follows:

is_Publication_ofScientific_Activity*rdf:typeowl:ObjectProperty*;
rdfs:domain :Publication;
rdfs:range :Scientific Activity;
:**has**_Author_Person*rdf:typeowl:ObjectProperty*;
rdfs:domain :Publication;
rdfs:range :Person;
:**is**_Publication_ofScientific_Event*rdf:typeowl:ObjectProperty*;
rdfs:domain :Publication;
rdfs:range:Scientific Event;
:**describes_**Publication_Subdiscipline*rdf:typeowl:ObjectProperty*;
rdfs:domain :Publication;
rdfs:range :Subdiscipline;
:**Publication_Title***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs:Literal;
:**Publication_Edition**rdf:typeowl:DatatypeProperty;
rdfs:domain :Publication;

rdfs:rangerdfs:Literal;
:**Publication_Publication Date***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs:xsd:int;
:**Publication_ Keywords***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs:Literal;
:**Publication_ URI***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs: xsd:anyURI;
:**Publication_Abstract***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs:Text;
:**Publication_Reference***rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs: *Literal*;
:**Publication_Document_Language** *rdf:typeowl:DatatypeProperty*;
rdfs:domain :Publication;
rdfs:rangerdfs: *Literal*;

The ISNOD information model should be tiered and consist of at least two elements [6]: the data storage subsystem and the information resource management service subsystem. In this case, the data storage function must be separated and be independent from other functions and services of the system. The data storage subsystem – digital repository – is one of the critical elements of the system and serves to provide for the "function" of long-term storage of information resources. Services can change, but the data must be stored in a secure place (Figure 4).

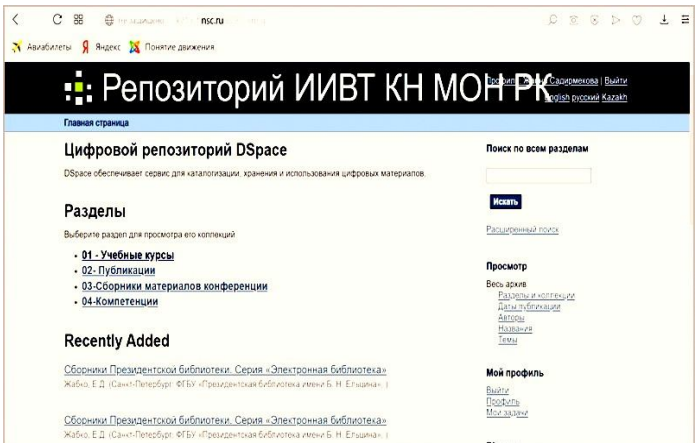

Figure 4: Repository view window

Based on this model, the concept of "institutional repository" was created as a system of long-term storage and accumulation of information, provision of secure access to digital entities, being the result of the intellectual activity of a research or educational institution [7].

Publication has a basic set of attributes, which is based on the Dublin Core data scheme, extended in compliance with the MECOF requirements [8].

The type of Publication determines the set of mandatory descriptive metadata and rules of their display. Publication is the only class of documents that can contain information content. The latter is, as a rule, external with respect to the IS and is stored in a digital repository [9]; in the system metadata, it is represented with a link to the resource (Figures 5, 6).

Figure 5: View of Publication class information

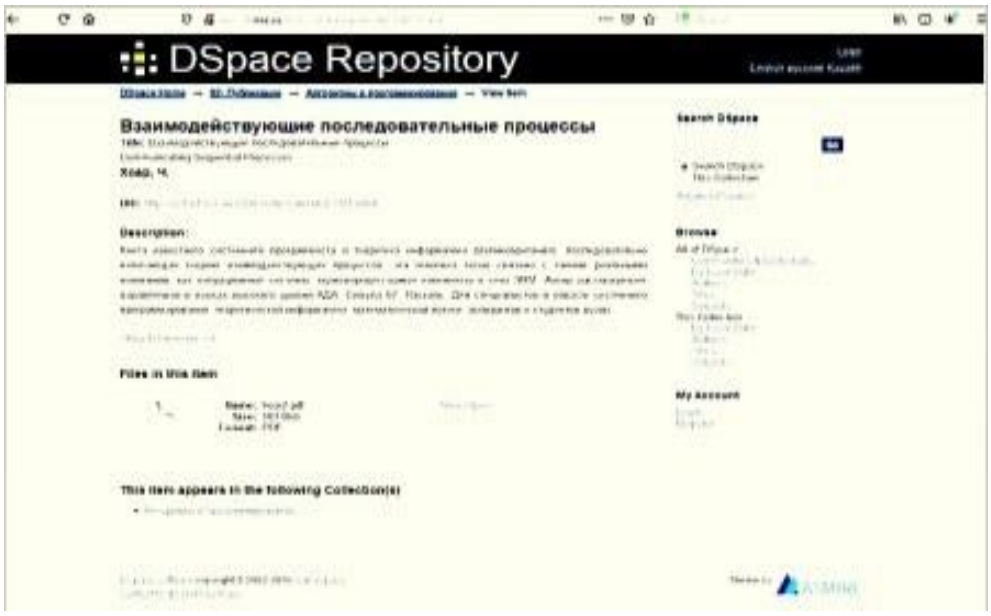

Figure 6: Link to repository

3. ISNOD METADATA SCHEME BASICS

The main requirements for the development of the ISNOD data model were:

- the possibility of adequate information representation;
- interoperability with other systems containing similar information.

With account of these requirements, we included the following types of classes in the model: Organization, Person, Scientific Activity, Publication, Scientific Events, Study Course, Subdiscipline, Competence, Geographic Location, Conference Proceedings. The development of the class structure requires defining the characteristic properties of each information entity that sufficiently describes it. The requirement for interoperability complicates this problem. Ensuring interoperability is one of the most important tasks, as significant information resources are already digitally represented, and the main problem arising when using them is the lack of links between them. In order to provide for the interaction with a wide range of information systems, it is necessary to ensure the model's compatibility with the common standards of data representation. We familiarized ourselves with the proposals and standards of meta description of information classes included in ISNOD, such as Dublin Core [10] and CERIF [11].Dublin Core was chosen as the basis for the implementation of the *Publication* resource from the variety of available options. This choice is substantiated by the following advantages of the standard:

- the set of core semantic elements is compact and at the same time allows setting virtually all the required attributes;
- the semantics of every element can be refined using qualifiers, both standard, known and comprehensible for everyone, and those specifically designed for accurate specification of the semantic meaning of a certain attribute when exchanging data within a small community;
- the standard makes it possible to use various semantic schemes, vocabularies, etc.;
- it has a defined mechanism for extracting information from description using unconventional extensions of the name space;
- the standard is gaining increasing popularity among the world community.

The ISNOD publication data model enables specifying any base element of *Dublin Core* (*Title, Author, Abstract*, etc.). It provides for the use of qualifiers refining the semantics of the base elements (such as *Parallel Title, Member–Editor*, etc.). This facilitates exchanging literature information based on this standard. However, a serious obstacle for the interoperability of such subscheme of the ISNOD model with other systems is the fact that the majority of systems consider the standalone properties of Publications, such as *Author, Publisher, Source*, as regular text attributes, while they are connections with other entities (*Persons, Organizations, Publications*). Such models do not contradict the Dublin Core, but result in a certain incompatibility with the ISNOD model, in particular when integrating data into the system [8].

We chose the CERIF standard to represent information of other classes: *Organization, Person, Scientific Activity, Scientific Events, Study Course, Subdiscipline, Competence, Geographic Location*. It is based on the data model that includes entities such as *Project, Organization and Person*, links between them, as well as the attributes of these entities. The standard defines three tiers of detailing in class description:

1) complete class description, which contains an advanced set of attributes and allows describing various resource schemes;
2) a set of attributes for exchanging data between various systems;
3) a reduced set of attributes, which is the metadescription of classes.

Worth noting is the fact that each standard offers its own data model, and often its own information recordingsyntax. The IS approach consists in using a single model of RDF-defined data and syntax for exchanging metadata. Only the semantics of the attributes of certain resources is taken from the relevant standard. If it is impossible to find a suitable element in the standard name spaces, one can create his own space by defining it with URI and add the elements required by the semantics. This approach facilitates the interaction between ISNOD and other systems significantly [12].

4. METADATA EXTRACTION FROM DOCUMENTS IN THE COURSE OF INFORMATION SYSTEM INTEGRATION

In order to populate ISNOD with content, information is collected from sources, such as websites of organizations, associations, projects, and conferences, knowledge portals, scientific social media, etc. As mentioned above, these sources are used to obtain information *on Projects, Organizations, Persons, Conferences*, all the basic classes of the activity ontology, except the information on *Publications.* The information about *Publications* is extracted from repositories (Dspace).For each of these classes, a separate method for information extraction is developed, which includes a set of templates generated based on ontology. Documents can be published in the Internet in various formats. But since HTML is today the main format of information representation in the Internet, the proposed methods for extracting the information *on Projects, Organizations, Persons, Conferences* are tuned to work with HTML pages.

To facilitate the analysis, the HTML page is represented as a DOM tree according to the DOM (Document Object Model) standard, which regulates the methods for document content representation (in particular, HTML page) by means of an object set [13]. Based on a relevant template, the DOM tree of each page is analyzed, and the information described with this template is extracted.To extract the metadata of the publication from repositories in batches, the data are exported in the Dublin Core format (Figure 7).

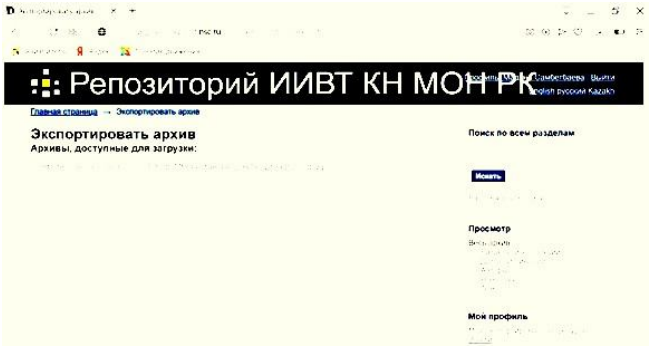

Figure 7: Data extraction from the repository

The template is an XML document, which specifies the markers signaling about the location of this object, relation, or attribute for the objects, relations and attributes of ontology.Figure 8 shows a fragment of the template used to extract publication metadata from the repository. This template enables the extraction of the attributes of the *Publication* class, such as *Title, Author, Abstract, Document Type, Document Language, Publication URI, Bibliographic Data, Keywords, Publication Date,* etc.

Figure 8: Information extraction from the data repository using a template

We developed a PHP code in order to automatically extract all publications from the repository along with Dublin Core metadata in the ontology into the Protégé software (Figure 9). It facilitates populating the ontology, one of the tasks of which is the automation of the system population with data.

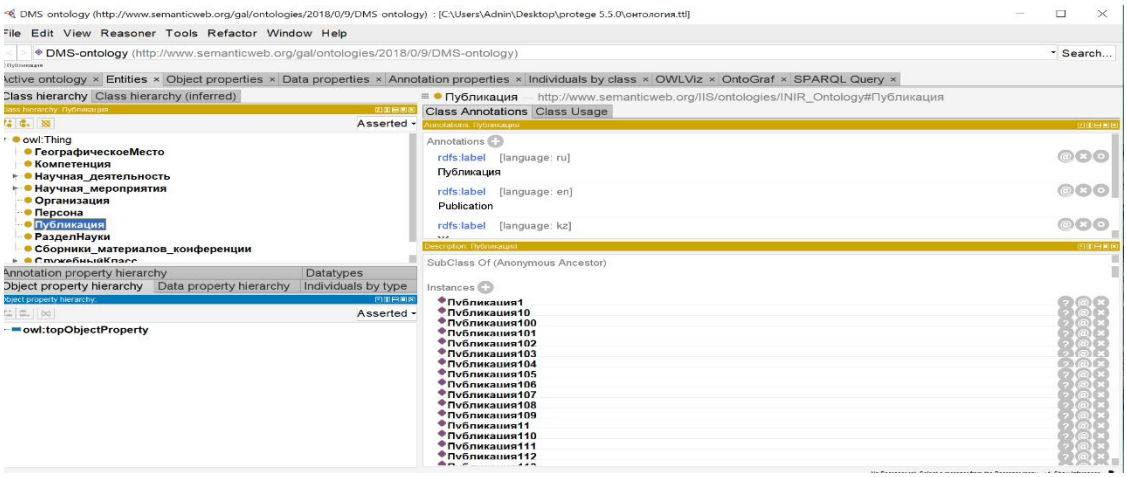

Figure 9: Publication metadata extraction from the repository

In this way, the following functions of the information systems for the academic activity are revealed:

1) providing access to the information on various aspects and participants of the academic activity, researchers, as well as groups, communities, and organizations involved in the research process;
2) integration of the related resources in the Internet (relational databases, XML and HTML resources, newsfeeds, etc.) in a uniform information space;
3) providing users with a means to find interesting information in the entire information space of the portal;
4) providing information resource users with support (e.g. announcement of various events and occasions);
5) maintaining a flexible interface, which makes it possible to take into account user's preferences when working with the resource and services rendered.

5. CONCLUSION

ISNOD enables researchers to significantly reduce the time required for providing access to the information of their interest. In this case, the efficiency of using each particular ISNOD directly depends on the extent of correctness of the information represented in it. Such an extent can be achieved by automatizing the information collection process. For this purpose, the subsystem of information sourcing from the Internet is developed.As of today, all the main components of this subsystem have been implemented and methods for extracting information about *Projects*, *Persons*, *Organizations* and *Events* have been developed, including the corresponding templates and processors that implement the information on publications.

ACKNOWLEDGEMENT

The work is supported by the grant of funding of scientific and (or) scientific and technical research for 2018-2020. MES RK (No. AP 05133546).

REFERENCES

[1] Arsky Yu.M., Gilyarevsky R.S., TurovI.S. andChyorny A.I.,**Infosphere: information structures, systems, and processes in science and society**, Moscow, VINITI Press, 1996. – 489 pp.

[2] Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Fedotova O.A., Sagnayeva S.K., Bapanov A.A. and Tazhibaeva S.Z.,**Classification model and morphological analysis in multilingual scientific and educational information systems**,*Journal of Theoretical and Applied Information Technology*, 2016, Vol.86, No. 1. pp. 96-111.

[3] Nadu T., Ramakrishnan I.D.B.,**Novel Heterogeneous Data Integration Using KNN Algorithm**,*International Journal of Emerging Technologies in Engineering Research*, 2019,Vol. 7, Iss. 4. Retrieved from: https://www.ijeter.everscience.org/Manuscripts/Volume-7/Issue-4/Vol-7-issue-4-M-03.pdf

[4] Aarthi S., Narsepalli Vamsi Kishan, Surya Teja V., Harsha Vardhan Gupta N.V., **Classification of Phishing Website Based on URL Features**, *International Journal of Emerging Technologies in Engineering Research*, 2019,Vol. 7, Iss. 5. Retrieved from: https://www.ijeter.everscience.org/Manuscripts/Volume-7/Issue-5/Vol-7-issue-5-M-04.pdf

[5] Zagorulko Y., Borovikova O. and Zagorulko G.,**Pattern-based methodology for building the ontologies of scientific subject domains**,*Proceedings of the 17th International Conference SoMeT_18, Series: Frontiers in Artificial Intelligence and Applications*, Vol. 303, 2018,pp. 529-542.

[6] **DSpace: an open source solution for accessing, managing and preserving scholarly works**. Retrieved from:http://www.dspace.org

[7] Sadirmekova Zh.B. and Tusupov D.A, **Institutional open access repositories**,*Proceedings of the International Scientific and Practical Conference*

"Science and Education Development Prospects in Globalization Conditions", 2019, pp. 483-486.

[8] **Functional requirements for bibliographic records, final report.** Retrieved from:http://archive.ifla.org/VII/s13/frbr/frbr.htm.

[9] Zhizhimov O.L., FedotovA.M. andFedotova O.A.,**Construction of a generic model of an information system for processing scientific heritage documents**,*Herald of the Novosibirsk State University.Series: Information Technology*, 2012, Vol. 10, No. 2.,pp. 5-14.

[10] **ANSI/NISO Z39.88-2004 (R2010) – The open URL framework for context-sensitive services**,National Information Standards Organization, 2010, pp. 122.

[11] **CERIF 2008 – 1.2 Full Data Model (FDM). Introduction and specification**. Retrieved from: http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_FDM.pdf

[12] Braginskaya L., Kovalevsky V., Grigoryuk A. and Zagorulko G., **Ontological approach to information support of investigations in active seismology**, *Proceedings of the 2nd Russian-Pacific Conference on Computer Technology and Applications (RPC)*, 2017, pp. 27-29.DOI:https://doi:10.1109/RPC.2017.8168060.

[13] **Library linked data incubator group final report**. Retrieved from:http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025