The development of a system for automated ontology building based on heterogeneous ontology design patterns

Y A Zagorulko, O I Borovikova and G B Zagorulko
The A.P. Ershov Institute of Informatics System SB RAS, 6, pr. Acad. Lavrentieva, 630090, Novosibirsk, Russia
E-mail: zagor@iis.nsk.su

Abstract. Currently, ontologies have become the main means of formalizing and systematizing knowledge and data in scientific subject domains (SSDs). In this regard, there is an urgent need for the methods and software tools allowing to involve specialists from different scientific subject domains in the process of ontology building. A possible way to solve this problem is to use ontology design patterns (ODPs). The paper presents an approach to the development of a system supporting the automated construction of the SSD ontologies based on ontology design patterns, which are the formal descriptions of field-proven solutions to typical ontological modeling problems. The main components of this system are repositories of heterogeneous ODPs and basic ontologies, including concepts common to most SSDs, and editors of patterns and ontologies used to build and populate ontologies.

1. Introduction
Currently, the development of ontologies of scientific subject domains (SSDs) retains its relevance. When solving practical problems, researchers working with large volumes of heterogeneous data need a high-quality systematization of these data, description of their semantics, acquisition of new knowledge based on their analysis, as well as formation of common terminology and support for the required data search in its terms.

Nowadays, ontologies intended for describing the various aspects of SSDs, in particular, scientific activities and scientific results, have been created. For example, paper [1] provides an overview of ontologies for annotating scientific publications and documents. The Scholarly Ontology described in [2] presents scientific activity as a special type of “business process.” The central concept of this ontology is Activity, which is described in three perspectives (aspects). The first perspective concerns the participants and the goals of this activity, the second perspective concerns the intellectual framework and work organization, and the third concerns the resources used, consumed or produced in the course of the activity.

To simplify and facilitate the process of ontology constructing and populating, various automated approaches and methods have been proposed, many of which are focused mainly on extracting the concepts and relations between them from the texts and building ontological structures on this basis. To apply these methods successfully, additional linguistic resources are required. They are necessary both for solving the related sub-tasks, such as extracting and classifying named entities, extracting relations between entities, disambiguating, etc., and for using them as a source of ontological information. For example, methods for building and populating the ontology using thesauri [3], encyclopedias [4], explanatory dictionaries [5], dictionaries of synonyms [6], and Wiktionaries [7] have been developed.
In [8], the authors propose a combined approach to solving the ontology population problem. Apart from linguistic resources (Verbnet, FrameNet, WordNet), this approach uses additional ontological resources represented by top-level ontologies from the field of biomedicine such as BioTop (https://bioportal.bioontology.org/ontologies/BT) and OBO (http://obofoundry.org).

Recently, an approach using ontology design patterns (OPDs) [9], which are documented descriptions of proven solutions to the typical problems of ontological modeling, has been intensively developed. In particular, the OPDs can serve to expand the expressive capabilities of the ontology description language (structural logical patterns), to represent typical fragments of ontologies (content patterns), to increase the readability and usability of ontologies (presentation patterns), to expand the capabilities of logical inference machines and ontological processing information (reasoning patterns), to solve the problems of reengineering and alignment of ontologies (correspondence patterns), to automate the construction and population of ontologies based on natural language texts (lexico-syntactic patterns).

The use of ontology design patterns greatly facilitates the process of ontology constructing, allowing not only to save human resources, but also to involve the experts in modeled subject domains in this process, which is especially valuable when building the SSDs ontologies. For this, the experts need to be equipped with the tools supporting the ontology building process and providing them with necessary resources. When applying the OPDs, these tools should include both systematized sets of patterns focused on the presentation of scientific knowledge and ontology development tools supporting the construction and use of the OPDs.

The paper presents an approach to the development of a system that supports the automated construction of the SSD ontologies on the basis of ontology design patterns.

2. Features of the scientific subject domain ontology

A scientific subject domain ontology is understood as a subject area encompassing a certain scientific discipline or field of scientific knowledge in all its aspects, including objects, subjects and methods of research specific to it, classes of problems to be solved, scientific activities, and the scientific results obtained.

In order for the ontology to represent an SSD in its entirety, it should describe not only the aspects above, but also publications and information resources representing scientific activities in this SSD and scientific results obtained. In this regard, a SSD ontology is structured as three interconnected ontologies: ontology of the knowledge domain, ontology of tasks and methods, and ontology of scientific Internet resources. The first ontology defines a system of concepts and relations intended for a detailed description of an SSD and research activities carried out within its framework. The second ontology is intended to describe the tasks and methods specific to this SSD, and the third helps to describe scientific information resources found on the Internet.

To facilitate and simplify the process of constructing an ontology of SSD, we propose a method on the basis of base (nuclear) ontologies. These include only the most general entities (concepts and relations) independent of a specific SSD. In base ontologies, such entities are represented using the ontology design patterns that are specialized and concretized (instantiated) when the ontology of a specific SSD is constructed.

At the subsequent stages, the completion and population of the SSD ontology is also performed using the ODPs but of different type, namely, the lexico-syntactic patterns. Such patterns automatically complete and populate the ontology based on a collection of texts. This collection is selected so as to include texts covering as many descriptions of the entities of the SSD modeled as possible.

To support this method of ontology building, a system for automated ontology construction is being developed. It is described in the next section.
3. The system for automated construction of the SSD ontologies

The proposed system for the automated construction of the SSD ontologies (see figure 1) includes the repositories of base ontologies and ontology design patterns, formats and languages for describing the ODPs, dictionaries of general scientific and subject lexis (terms), editors of ontologies and ODPs, as well as an information-analytical Internet resource (IAIR) supporting the use of ontology design patterns.

Figure 1. The system for automated construction of SSD ontologies.

3.1. Repository of ontology design patterns

The repository of ontology design patterns currently includes four types of the ODPs: structural logical patterns, content patterns, presentation patterns, and lexico-syntactic patterns. These heterogeneous patterns form a system of interconnected patterns, which are shared when an ontology of a specific SSD is built.

The basis of such a system are content patterns defining the ways of representing the typical fragments of ontologies describing the SSD complex entities. To represent the content patterns and complex ontological structures, structural logical patterns expanding the expressive capabilities of the ontology description language are included in the repository.

To represent scientific knowledge, the repository has been supplemented by the content patterns developed in our project. These patterns serve to present concepts such as the Object of research, Subject of research, Research method, Division of science, Scientific result, Activity (Scientific activity), Project, Person, Organization, Publication, Information resource, etc.

To represent complex entities and constructions relevant in the construction of the SSD ontologies, the list of structural logical patterns includes ontology design patterns for representing the ranges of permissible values, n-ary and attributed relations (binary relations with attributes) [10].

Content patterns are tuned to a modeled SSD using import, specialization, generalization, composition, and expansion operations [11].

The ontology is populated with specific objects and facts using the instantiation operation of patterns performed on structural logical and content patterns.
To make the ontology more readable, convenient to use, and tailored to various types of users, the repository includes presentation patterns. These patterns define conventions for naming, annotating, and visualizing the ontology elements. In particular, presentation patterns allow us to adjust the view of elements and structures of the ontology to the various needs of its users.

The names of ontological elements (in one or several natural languages) are defined in the presentation patterns using labels denoting the names of entities in various texts and terminological systems. These names can later be used to recognize the corresponding entities by the lexico-syntactic patterns included in the repository of the ODPs.

As mentioned above, the lexico-syntactic patterns [12] are designed to automate the construction and population of ontologies based on natural language texts (NL). Using these patterns, language structures are mapped to ontology elements. The structural elements of lexico-syntactic patterns can be groups of words and phrases of a specific NL, which are found in general scientific and specialized dictionaries and correspond to the descriptions of ontological constructions defined both in the ontology description language and in structural logical patterns, content and presentation patterns.

Lexico-syntactic patterns are constructed for each structural logical and content pattern. This can be done manually or automatically. At the same time, several lexico-syntactic patterns can be built for one complex pattern in order to extract entities from the text corresponding to the various parts of the complex pattern and different ways of representing them in the text.

3.2. Formats and languages for describing the ontology design patterns

The ODPs included in the repository are described in a format built on the basis of that proposed on the ODPA association portal (http://ontologydesignpatterns.org) [13]. According to this format, the description of a pattern includes information about its author and the field of application, its graphical representation, text description, links to other patterns, a set of scripts and examples of use. The content pattern can additionally be provided with links to the patterns specializing it, as well as a set of Competency questions [11, 13]. We can use competency questions both to develop the content pattern itself and to find the desired pattern in the repository when developing a specific ontology [14].

The pattern description format used in the system is expanded by relations reflecting the interaction of the patterns of different types with one another. For example, the relation “uses” defines the connection between a content pattern and a structural logical pattern, the relation “is used for presentation” links a presentation pattern with a content pattern, the relation “corresponds” specifies the connection between a lexico-syntactic pattern and a content pattern, etc.

To be used in the construction of ontologies, patterns should be described (implemented) in a specific ontology description language [10]. Thus, content patterns and structural logical patterns are defined in the OWL language [15], recommended by the W3C consortium as a standard, which allows solving the problem of ontology reuse. The graphical representation of patterns uses the popular modeling language UML [16].

The presentation patterns specifying the rules for naming the ontology elements are described in a natural language, and the presentation patterns responsible for the annotation and visualization of ontology elements are described using the annotation properties provided by the OWL language.

To describe lexico-syntactic patterns, we use a modification of the language proposed in [17] for extracting information from scientific and technical texts.

3.3. Repository of Base Ontologies

As mentioned above, the SSD ontology consists of three interconnected ontologies that describe the SSD components, such as the knowledge domain, tasks and methods, and scientific Internet resources. Building such a complex ontology from scratch is a very difficult and time-consuming process. To simplify and facilitate the process, we use a repository of base ontologies developed as part of the technology for building thematic intellectual scientific Internet resources [18]. At the moment, the repository includes four base ontologies described in the OWL language: the ontology of scientific knowledge, ontology of scientific activity, base ontology of tasks and methods, and base ontology of
Internet resources. Each of these ontologies is a starting point in constructing the corresponding part of the SSD ontology.

3.4. Repository of dictionaries of general scientific and subject lexis
The core of this repository is a dictionary containing general scientific lexis used in the texts of most scientific subject domains. However, this lexis is not enough to represent the constituent elements of lexico-syntactic patterns and presentation patterns. Therefore, this repository will be supplemented by subject dictionaries, including lexis specific to the SSDs modeled. In addition, it will include the dictionaries of synonyms, necessary for recognizing various spellings of entity names and their properties (classes, class instances, properties and their values) in specific texts and terminological systems.

3.5. Editors of Ontologies and Patterns
To build the SSD ontology and to specialize the ODPs, the popular ontology editor Protégé (https://protege.stanford.edu) is used. At the moment, the use of ODPs is supported by a special editor (data editor) providing knowledge engineers and domain experts with convenient graphical tools for populating the ontology with actual data (class instances and their properties).

To support the work on the creation and edition of structural logical patterns and content patterns, we develop an editor that will provide visual design and edition of patterns and their subsequent mapping in the OWL format.

4. Support for the use of patterns
To support the use of patterns in the construction of an SSD ontology, an information-analytical Internet resource (IAIR) “Ontology Design Patterns” has been developed. The IAIR systematizes knowledge and information resources about the ODPs, methods and tools for ontology construction using them and also provides a user with content-based access to this information.

The conceptual basis of the IAIR is the ODP ontology containing formal specifications of various types of patterns, methods and tools for their development and application, publications and information resources related to the problems of “Ontology Design Patterns.” Based on these specifications, structures are built to present information on specific patterns, integrated information resources, and methods for developing and using patterns in the IAIR content. Also, on the basis of the ontology, convenient navigation through the IAIR content and content-based search for necessary information are organized.

The ODP ontology is built on the basis of the scientific activity ontology previously developed by our team [18] and implemented using the OWL language.

The core of this ontology is the class Ontology Design Pattern, which defines the basic properties of patterns and its subclasses, which serve to represent various types of patterns. These classes are Structural Pattern, Content Pattern, Presentation Pattern, etc.

In addition to the classes listed above, the ODP ontology includes classes representing concepts related to scientific activity and describing the context in which the patterns are used: Scope of application, Project, Task, Publication, Event, Person, Organization, Information resource, Geographical location. Using these concepts, each pattern can be associated with subject areas, organizations and projects where it is used and with publications and information resources where it is described.

Also, relations between the patterns are defined. They are “complements,” “imports,” “generalizes,” “consists of,” and “specializes.” These relations correspond to the operations by which a specific pattern could be tuned to SSD modeled [11].

The IAIR content built on the basis of this ontology includes descriptions of the patterns we have developed and included in the repository and descriptions of the patterns presented on the ODPA Association portal (http://ontologypedesignpatterns.org) and suitable for developing the SSD ontologies.
Figure 2 shows the IAIR page. The ODP ontology is on the left side of the page, and the description of the content pattern designed to describe the research methods used in scientific activity is on the right side of the page. The pattern description includes its name and purpose, link to its graphical representation, and reference to the project where it was developed. The graphical representation of the pattern in figure 2 shows that when defining a specific research method in ontology, you need not only to indicate its name, creation date, author, and give its text description, but also to link it with instances of the ontology classes Activity, Division of Science, Scientific result and Task by the relations "is used in," "is realized in," and "decides."

Figure 2. A page of the IAIR “Ontology Design Patterns.” A pattern for describing a research method.

5. Conclusion
The paper proposes an approach to the development of a system for the automated construction of ontologies based on the ontology design patterns of various types. A combination of heterogeneous patterns into a single system allows their effective application to the construction of the ontologies of specific SSDs. An additional convenience for ontology developers is the information-analytical Internet resource “Ontology Design Patterns,” which systematizes information on available ODPs and provides content-based access to them.

We are planning further development of the system of ontologies automated construction and of the information-analytical Internet resource discussed here.

In particular, the ODP repository will be supplemented with new content patterns. As part of this activity, we are going to conduct experiments on the automatic recognition of the closed typical fragments of the SSD ontologies, which may become the prototypes of content patterns.
In addition, in the near future we are planning to carry out experiments on the automatic generation of lexico-syntactic patterns based on structural logical and content patterns included in the repository, using the dictionaries of general scientific and specialized lexis.

Acknowledgments
The research has been supported by Russian Foundation for Basic Research (project no. 19-07-00762) and by grant funding for scientific and (or) scientific and technical research for 2018–2020 MES RK (No. AP 05133546).

References
[1] Ruiz-Iniesta A, Corcho O 2014 A review of ontologies for describing scholarly and scientific documents Proc. 4th Workshop on Semantic Publishing co-located with the 11th Extended Semantic Web Conf. (ESWC 2014) vol 1155 (CEUR Workshop Proceedings)
[2] Pertsas V, Constantopoulos P 2017 Scholarly Ontology: modelling scholarly practices Int. J. on Digital Libraries 18 173–190
[3] Navigli R, Velardi P 2008 From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions Ontology learning and population: bridging the gap between text and knowledge vol 167 ed Buitelaar P and Cimiano P Series: Frontiers in artificial intelligence and applications (Amsterdam; Washington, DC: IOS Press) pp 71–87
[4] Bocharov V, Pivovarova L, Rubashkin V, Chuprin B 2010 Ontological Parsing of Encyclopedia Information Lecture Notes in Computer Science vol 6008 pp 564–579
[5] Kiselev Y, Porshnev S, Mukhin M 2015 Method of Extracting Hyponym-Hypernym Relationships for Nouns from Definitions of Explanatory Dictionaries J. Software Engineering 10 38–48
[6] Orobinska E 2012 Automatic Method of Domain Ontology Construction Based on Characteristics of Corpora POS-Analysis Proc. 15th Conf. Internet and Modern Society (IMS-2012) (MPSS: Saint Petersburg) pp 209–212
[7] Krizhanovsky A, Smirnov A 2013 An approach to automated construction of a general-purpose lexical ontology based on wiktionary J. of Computer and Systems Sciences International 52 (2) 215–225
[8] Ruiz-Martínez J, Valencia-Garcia R, Martinez-Bécker R, Hoffmann A 2012 BioOntoVerb: A top level ontology based framework to populate biomedical ontologies from texts Knowledge-Based Systems 36 68–80
[9] Gangemi A, Presutti V 2009 Ontology Design Patterns Handbook on Ontologies ed Staab S and Studer R (Berlin: Springer Verlag) pp 221–243
[10] Zagorulko Y, Borovikova O, Zagorulko G 2018 Implementation of Content Patterns in the Methodology of the Development of Ontologies for Scientific Subject Domains Communications in Computer and Information Science Artificial Intelligence. RCAI 2018 vol 934 ed Kuznetsov S et al (Springer, Cham) pp 260–272
[11] Blomqvist E, Hammar K, Presutti V 2016 Engineering Ontologies with Patterns: The eXtreme Design Methodology Ontology Engineering with Ontology Design Patterns. Studies on the Semantic Web vol 25 ed Hitzler P et al (IOS Press) pp 23–50
[12] Maynard D, Funk A, Peters W 2009 Using lexico-syntactic ontology design patterns for ontology creation and population Proc. Workshop on Ontology Patterns (WOP2009) vol 516 (CEUR Workshop Proceedings) pp 39–52
[13] Karima N, Hammar K, Hitzler P 2017 How to Document Ontology Design Patterns Advances in Ontology Design and Patterns. Studies on the Semantic Web (IOS Press: Kobe, Japan) vol 32 pp 15–27
[14] Wiśniewski D, Potoniec J, Ławryniewicz A, Keet C M 2019 Analysis of ontology competency questions and their formalizations in SPARQL-OWL J. of Web Semantics 59 100534
[15] Antoniou G, Harmelen F 2009 Web Ontology Language: OWL. *Handbook on Ontologies* ed Staab S and Studer R (Berlin: Springer Verlag) pp 91–110

[16] Booch G, Rumbaugh J E, Jacobson I 2005 *The Unified Modeling Language user guide second edition* (Addison-Wesley Professional) p 496

[17] Bolshakova E, Efremova N, Noskov A 2010 LSPL-patterns as a tool for information extraction from natural language texts *New trends in classification and data mining* ed Markov K et al (ITHEA: Sofia) pp 110–118

[18] Zagorulko Y, Zagorulko G 2015 Ontology-Based Technology for Development of Intelligent Scientific Internet Resources *Intelligent Software Methodologies, Tools and Techniques Proc. 14th Int. Conf. (SoMet 2015)* vol 532 ed Fujita H and Guizzi G (Springer International Publishing) pp 227–241