IRSTI 20.53.19: 28.23.02

Yu.A. Zagorulko ©

PhD

https://orcid.org/0000-0002-7111-6524

A.P. Ershov Institute of Informatics Systems,

Novosibirsk, Russia

zagor@iis.nsk.su

https://doi.org/10.55956/ZIYQ7931

DEVELOPMENT OF A METHOD FOR AUTOMATIC EXTRACTION OF ONTOLOGY ENTITY NAMES FROM NATURAL LANGUAGE TEXTS

Abstract. Domain ontologies play a crucial role in organizing, sharing, and reusing domain-specific knowledge within software systems. However, developing a software ontology is a time-consuming and complex process. To create such an ontology, a large number of relevant publications must be analyzed. This task of enriching the ontology with data from these sources can be streamlined and accelerated through the use of lexical and syntactic patterns derived from ontological design. This paper presents a method for the automated construction of ontologies within a scientific domain, leveraging a system of heterogeneous ontological design patterns (ODPs). This system includes ODPs tailored for ontology developers and automatically generated lexical and syntactic patterns, which can be used to enrich ontologies with information extracted from natural language texts.

Keywords: scientific subject area, patterns, ontological design, patterns of content, automatic pattern generation, ontology replenishment.

Zagorulko Yu.A. Development of a method for automatic extraction of ontology entity names from natural language texts // Mechanics and Technology / Scientific journal. – 2024. – No.3(85). – P.486-494. https://doi.org/10.55956/ZIYQ7931

Introduction. Ontologies are currently utilized extensively to formalize and organize knowledge and data in scientific subject areas (SSAs). They help describe a scientific discipline or a domain of scientific knowledge in all its dimensions, covering key objects and research subjects, methods used in the field, ongoing projects, and results achieved [1]. The creation of such an ontology involves several stages, with the primary ones being the development of the terminological section, where taxonomies of concepts and relationships are constructed, and the specification of their properties, followed by the population of the ontology through the addition of instances of concepts and relationships. While the first stage establishes the framework of the ontology, the second stage populates it with content.

To create an ontology that accurately represents an SSA in a comprehensive manner, a large volume of scientific papers and information resources from the domain must be processed. To streamline and expedite this process, methods for automatically enriching ontologies based on natural language texts [2-3] and web documents [3-4] have been developed. For automatic text processing, clustering-based methods, utilizing widely known clustering and statistical techniques, are employed, alongside template-based approaches that rely on linguistic templates. However, the former requires substantial text corpora to function effectively, making linguistic template-based methods more prevalent. The linguistic approach originated from an idea proposed in [5], suggesting the automation of semantic relationship construction using diagnostic contexts structured as lexico-syntactic patterns. This technique, known as Hearst patterns, was devised for processing unstructured texts in English. It has been broadly adopted for extracting generic relationships, involving the identification of ordered word pairs from document collections that match pre-constructed patterns. Hearst's method has been applied and refined by numerous other scholars and extended to additional languages.

Several studies have proposed formal frameworks for recording lexical and lexico-syntactic patterns. For instance, [6] introduces an XML-based language schema for formalizing these patterns, aimed at populating ontologies. The Alex system [7] along with its auxiliary tool DigLex [8] provide adaptable methods for defining word forms and combinations using templates, which are later applied for the automatic identification of these elements within texts. These systems extend the functionalities of conventional lexicographic tools by supporting alternatives, pattern references, repeaters, context conditions, distant context, etc. The language allows not only for the recognition of textual objects but also for defining their lexical and semantic properties. Nonetheless, Alex and DigLex lack built-in mechanisms for specifying grammatical attributes of recognized lexical units or for ensuring grammatical agreement between multiple units, which is critical for the precise identification of language constructs (e.g., noun groups). The LSPL language proposed in [9] overcomes this limitation by allowing the specification of grammatical properties for its constituent elements.

Research focused on addressing the challenge of automatic or semi-automatic ontology enrichment utilizes patterns (templates) that map linguistic structures in texts to corresponding ontology components such as concepts, relationships, and instances of both. These patterns are either lexicon-syntactic, which utilize lexical structures and syntactic rules [10-11], or lexicon-semantic, which integrate lexical, syntactic, and semantic aspects in the extraction process [12-13].

This paper proposes a method to automate the enrichment of SSAs ontologies through the application of lexicon-semantic patterns (LSPs), which enhance traditional lexicon-syntactic ontology design patterns. A key feature of this approach is that the LSPs are generated automatically from other ontology design patterns (ODPs) [14], which are incorporated into a system that facilitates the automated creation of ontologies using diverse ODPs [15-16]. It should be emphasized that developing and implementing such patterns for the Kazakh language will necessitate extensive research, given that these patterns are greatly shaped by the language's unique grammatical features. Therefore, the objective of this article is to enhance and streamline the process of knowledge extraction, ontology construction, and handling large volumes of natural language text.

Materials and methods. The ontology of any SSA includes not only descriptions of its inherent conceptual system, tasks, and methods for processing and analyzing information, but also descriptions of related information resources. In this context, it is practical to represent the ontology of an SSA as a collection of interconnected ontologies: the ontology of the knowledge domain, the ontology of tasks and methods, and the ontology of scientific web resources (Fig. 1).

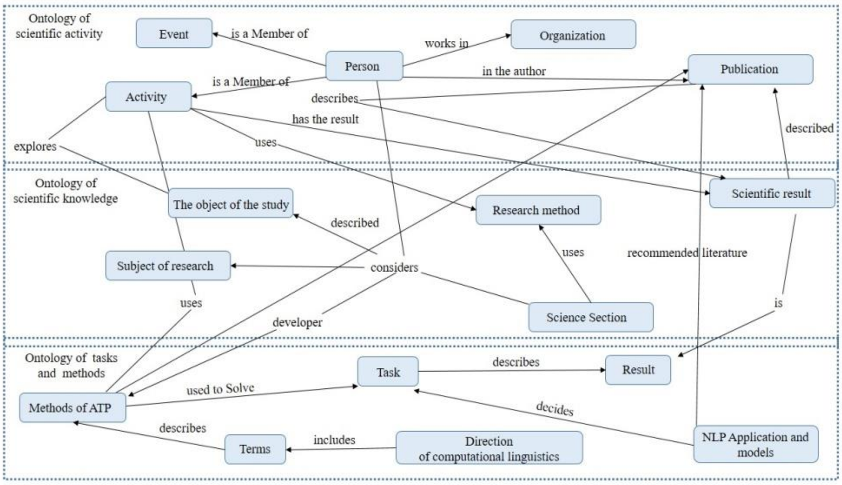


Fig. 1. An illustration of a SSAs ontology

The knowledge domain ontology outlines a system of concepts and relations aimed at a detailed portrayal of the modeled SSA and the research activities conducted within it. The ontology of tasks and methods defines the tasks addressed in a specific SSA and the methods used for their resolution. The ontology of scientific Internet resources captures the online information resources relevant to the given SSA.

Building a specific SSA ontology using foundational ontologies and a system of ODPs involves two key steps:

1) developing the components of the SSA ontology based on foundational ontologies by refining and extending them. At this stage, the structural logical patterns and content patterns provided in the base ontologies are tailored to suit the specific SSA;

2) enriching the SSA ontology by detailing structural logical and content patterns either present in the base ontologies or adapted from them through specialization to the specific SSA.

Specializing patterns may involve renaming elements or specifying particular names and values for properties like attributes and relations. An illustration of such specialization is depicted in Figure 2, which follows the structural logic of the "Binary attributed relation" pattern.

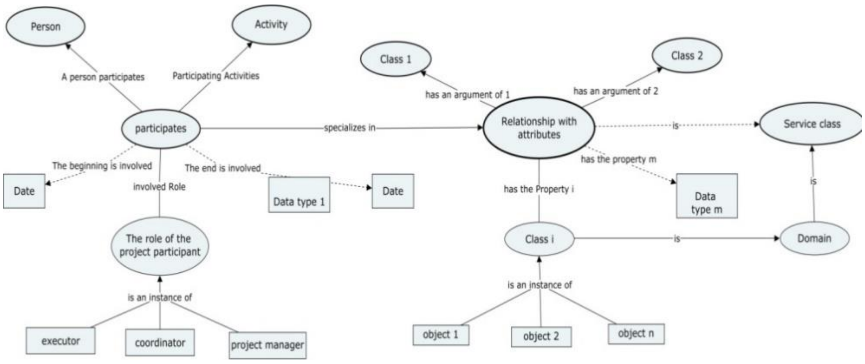


Fig. 2. An illustration of using a pattern

In this pattern, the core element is the Relation class with attributes, to which base classes representing the binary relation arguments are linked through the “is an Argument” and “has an Argument” relations. The pattern dictates that exactly one instance of each argument is required, as shown by the relationship labels. Attributes of a binary attributed relation are represented as properties of the Relation class, using attributes such as “has Attribute” and “has Attribute from Domain”. In general, this kind of relation may not include attributes, which is also demonstrated by the relationship labels describing these properties.

Specialization or concretization of a pattern involves substituting specific values for the properties defined in it. Figure 3 provides an example of the concretization of the Method content pattern, which was used to represent information about an automatic text processing method.

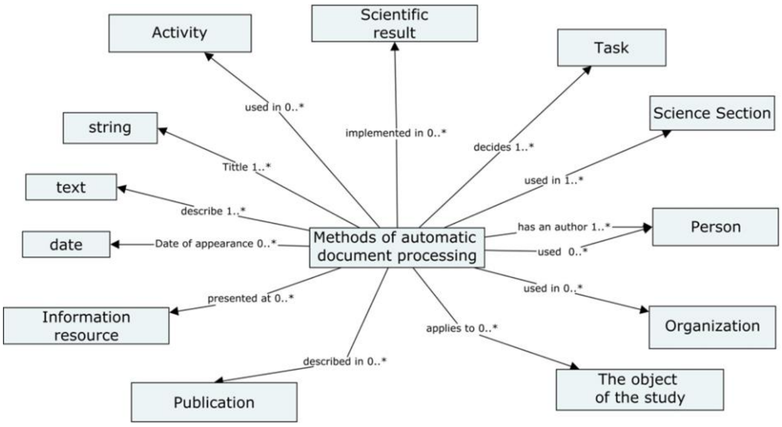
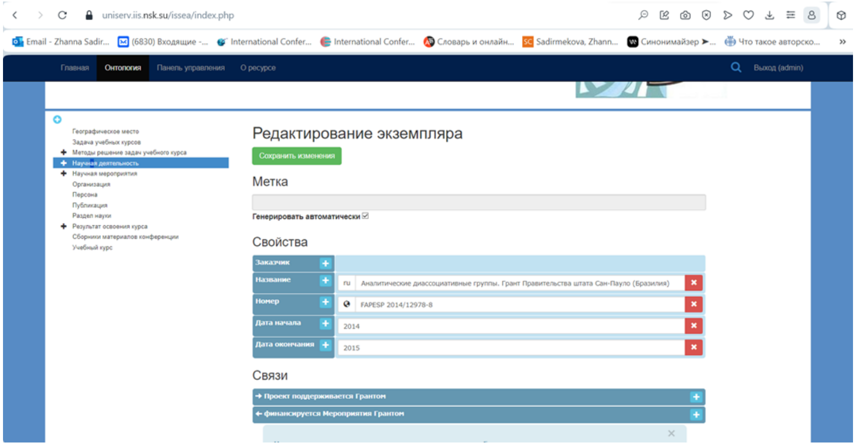


Fig. 3. Content pattern concretization “Automated text processing method” (Methods of ATP)

The process of using content patterns to populate the SSA ontology is facilitated by a specialized data editor (Fig. 4). This tool allows subject area experts to populate the ontology with real data-class objects and their properties. When enriching the ontology, the user selects the desired class from a hierarchy of ontology classes, and the editor retrieves the corresponding pattern by class name. Based on this, it generates a form with fields for the user to fill in the properties of the selected class’s object. The editor can also interpret the attribute relationships defined by the pattern in Figure 4. This allows users to work with the object’s properties, set via these relationships, in the same way as they would with “regular” object properties. The only difference is the need to assign values to the attributes of such a relationship.



Редактирование экземпляра

Метка

Генерировать автоматически ☑

Свойства

Заказчик	
Название	ru Аналитические диассоциативные группы. Грант Правительства штата Сан-Пауло (Бразилия)
Номер	FAPESP 2014/12978-8
Дата начала	2014
Дата окончания	2015

Связи

→ Проект поддерживается Грантом

← финансируется Мероприятия Грантом

Fig. 4. Data Editor

Lexico-syntactic pattern. To enable automatic ontology enrichment, each content pattern is accompanied by a set of lsps (lexico-semantic patterns) that describe various ways of presenting relevant information in scientific texts. These lsps serve as the basis for information extraction.

LSPs are structured in a hierarchical system comprising terminological patterns (T-LSPs) and information patterns (I-LSPs). T-LSPs are utilized to describe fundamental linguistic constructs and to extract new terms, whereas I-LSPs are designed to extract factual data from the text and create corresponding ontology components. These facts are structured as triples: the Object, which is an SSA entity identified in the text, the Property, representing an attribute of the entity, and the Value, which corresponds to the attribute’s value. When aligned with an ontology, the Object becomes an instance of an ontology class, the Property represents the name of a relation (such as a type, attribute, or object property), and the Value is linked to the respective ontological value.

Each LSP implements a model that includes three elements: Arguments (a set of semantic arguments, where SSA terms or objects are matched), Constraints (semantic, syntactic, or positional conditions on these arguments), and Results (which can be a new term or a generated ontology fragment, depending on the LSP type).

To automate the ontology population using LSPs, it is necessary to extract specific SSA terms from the text. This process is supported by a subject dictionary and terminological patterns that extract new terms (like object names or specific predicates). The subject dictionary is an extension of a general scientific vocabulary, organizing terms based on their semantics and storing all the necessary information for retrieving and analyzing terms from the text.

The subject dictionary is structured as a system in the form D = <W, P, M, G, S>, where W represents a set of lexemes, each associated with details about its forms; P is a collection of multi-word terms, represented by a pair <N-gram, structure type>, where the N-gram denotes a sequence of lexemes and the structure type defines a vertex along with matching rules; M signifies the morphological model, encompassing morphological classes and attributes; G specifies rules for extracting multi-word terms; and S defines lexical-semantic features specific to the subject area.

The semantic part of the dictionary includes two separate lexical-semantic class hierarchies: a universal hierarchy inherited from a general scientific dictionary, and a subject-specific hierarchy based on the SSA ontology. A method has been developed for the automatic generation of lexical-semantic characteristics, which uses the names of ontology elements to automatically construct the dictionary.

Each term in the dictionary is annotated with attributes from both the subject-specific and universal hierarchies. For example, the verb “to use” is assigned four syncretic features, each combining the universal class “Application” with ontology-based attributes such as Method.applies_to, Method.uses_to, and Method.realizes.

Lexical-semantic features from the dictionary are employed in LSP descriptions (both in arguments and results) to reference SSA terms with specific semantics, where prior detailed knowledge about these terms is not required.

Automatic generation of LSP. Lsps are generated automatically based on ontology design patterns, dictionaries of general scientific and subject-specific vocabulary, and the current version of the ssa ontology. Figure 5 illustrates the relationships between the system components involved in generating lsps.

The LSP generation process begins with building and populating the subject vocabulary. Terms, including lexemes and term-like N-grams, are extracted from the ontology and content pattern descriptions. These terms are then grouped into lexical-semantic classes, with each term tagged with its corresponding semantic features.

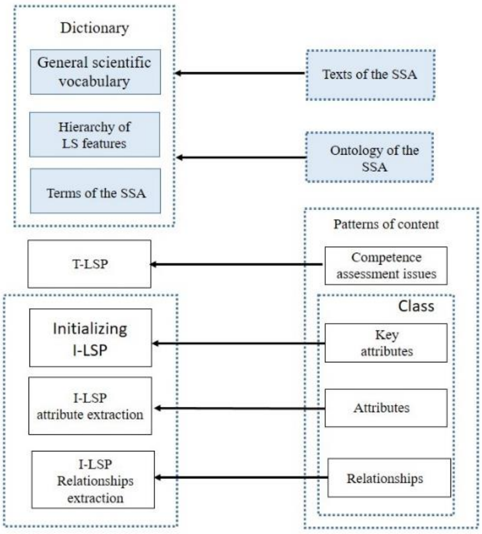


Fig. 5. Scheme of interconnections of system components involved in LSP generation

Based on an analysis of the structure of the content patterns, variables in the metapatterns are labeled, leading to the formation of T-LSPs (Terminological Lexico-Semantic Patterns) and I-LSPs (Information Lexico-Semantic Patterns).

Initially, the generated T-LSPs are applied to analyze competency assessment questions and scientific corpus texts. Competency assessment questions, expressed in natural language, help extract predicate terms and set initial syntactic constraints on the extracted facts, which are later refined based on the scientific text corpus.

When creating I-LSPs, information about the key attributes of ontology classes is crucial, particularly for pattern initialization. For example, the Name attribute is often used, but other key attributes, such as those in the Persona class, may require different generic patterns. I-LSPs also require syntactic and positional constraints based on how they appear within the text corpus.

From the ontological knowledge components, a distinction can be made between the knowledge necessary for extracting information from texts to update the ontology and the various ways in which ontological entities can be described linguistically. Formalizing this knowledge within an LSP system allows existing automated text processing technologies to be used for enriching the SSA ontology.

Research results and discussion. Experimental testing of the automatic ontology replenishment subsystem was conducted on the SSA ontology titled “Modern Methods of Automatic Text Processing.” The experiment focused on the content pattern “Methods of ATP.” A subject-specific vocabulary was generated automatically based on the current state of this NGO ontology and the content pattern description. This vocabulary included 214 terms annotated with 21 lexico-semantic features, along with 34 T-LSPs designed to extract new terms (such as class instance names and predicate words). Labels (rdfs:label) of attributes and relationships from the Methods of ATP class, as well as attribute values for instances of this and related classes, were used to generate the lexico-semantic feature hierarchy. A corpus of scientific publications in Kazakh, totaling 53.9 thousand tokens, was employed to evaluate the quality of the generated LSPs.

By applying T-LSPs aimed at extracting individual names from the Methods of ATP class, 111 occurrences were identified, with 73 unique terms recognized, all classified under the lexico-semantic class Method.Name. Of these, 70 were new terms not previously present in the ontology-based dictionary. Similar T-LSPs were used to extract individual names from other classes related to the Method class. For example, in the Task class, 42 pattern occurrences were found, resulting in the extraction of 33 terms, 30 of which were new, such as “cognitive modeling task,” “ordinal classification task,” and “integer programming task.” Some errors arose, including terms representing sources, created by the same syntactic rule, such as “task of a thesis”.

To extract new predicate terms, T-LSPs were applied to model ontological relationships between two objects. For instance, a T-LSP designed to extract predicate terms for the class Method.solves.Problem found 22 occurrences in the corpus, identifying terms like “put,” “allow to construct,” “allow to allocate”, “search for solutions” and “allows to solve”. Overall, 38 new predicate terms were identified.

Three experts assessed the T-LSP's overall accuracy, which averaged 88% and had an estimated 90% agreement among experts.

Conclusion. The research outlines a method for automating the development and population of SSAs ontologies utilizing heterogeneous ontology design patterns. This approach has been incorporated into a system called SAOC, which automates the construction of ontologies. A distinctive feature of this method is that it allows knowledge engineers and SSA experts to initially develop and populate the SSAs ontology by applying structural and content patterns. Once the ontology is established, the process of further populating it is automated through the use of lexicon-semantic patterns (LSPs), which are generated based on content patterns and the latest version of the SSAs ontology available in the SAOC repository. What sets this method apart from other approaches that leverage LSPs is its capability to automatically generate these patterns, ensuring a more efficient and systematic population of the SSAs ontology.

References

1. Fernández M. et al. Building a chemical ontology using METHONTOLOGY and the ontology design environment // IEEE intelligent Systems, 1999. Vol. 14, No. 1. P. 37-46.
2. Sure Y., Staab S., Studer R. Ontology engineering methodology // Handbook on ontologies. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. – P. 135-152.
3. Pinto H. S., Staab S., Tempich C. DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies // ECAI, 2004. Vol. 16. P. 393-397.
4. De Nicola A., Missikoff M., Navigli R. A proposal for a unified process for ontology building: UPON // International conference on database and expert systems applications. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. – P. 655-664.
5. Gangemi A., Presutti V. Ontology design patterns //Handbook on ontologies. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2009. – P. 221-243.
6. Blomqvist E., Hammar K., Presutti V. Engineering ontologies with patterns–the eXtreme design methodology // Ontology Engineering with Ontology Design Patterns. – IOS Press, 2016. – P. 23-50.
7. Karima N., Hammar K., Hitzler P. How to document ontology design patterns // Advances in Ontology Design and Patterns. – IOS Press, 2017. – P. 15-27.
8. Zagorulko Y., Zagorulko G. Ontology-based technology for development of intelligent scientific internet resources // Intelligent Software Methodologies, Tools and Techniques: 14th International Conference, SoMet 2015, Naples, Italy, September 15-17, 2015. Proceedings 14. – Springer International Publishing, 2015. – P. 227-241.
9. Zagorulko Y., Borovikova O. Technology of ontology building for knowledge portals on humanities // International Conference on Knowledge Processing in Practice. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2007. – P. 203-216.
10. Borovikova O. et al. Methodology for knowledge portals development: background, foundations, experience of application, problems and prospects // Joint NCC&IIS Bulletin, Series Computer Science, 2012. Vol. 34. P. 73-92.
11. Gamma E. et al. Elements of reusable object-oriented software // Design Patterns, 1995.
12. NeOn Project [Electronic resource] – Access mode: http://www.neon-project.org.
13. Association for Ontology Design & Patterns [Electronic resource] – Access mode: http://ontologydesignpatterns.org.
14. Ontology Design Patterns (ODPs) Public Catalog, [Electronic resource] – Access mode: http://odps.sourceforge.net.
15. Dodds, L., Davis, I.: Linked Data Patterns (2012) [Electronic resource] – http://patterns.dataincubator.org/book.
16. Krisnadhi A., Hitzler P. A core pattern for events // Advances in Ontology Design and Patterns. – IOS Press, 2017. – P. 29-37.

Material received on 27.09.24.

Ю.А. Загорулько

А.П. Ершов атындағы информатика жүйелері институты, Новосибирск қ., Ресей

ТАБИҒИ ТІЛДЕГІ МӘТІНДЕРДЕН ОНТОЛОГИЯЛЫҚ НЫСАН АТАУЛАРЫН АВТОМАТТЫ ТҮРДЕ АЛУ ӘДІСІН ӘЗІРЛЕУ

Аңдатпа. Домен онтологиясы бағдарламалық жүйелерде доменге қатысты білімді ұйымдастыруда, бөлісуде және қайта пайдалануда шешуші рөл атқарады. Алайда, бағдарламалық жасақтаманың онтологиясын жасау көп уақытты қажет ететін және күрделі процесс. Мұндай онтологияны құру үшін көптеген тиісті жарияланымдарды талдау қажет. Онтологияны осы көздерден алынған мәліметтермен байытудың бұл міндетін онтологиялық дизайннан алынған лексикалық және синтаксистік үлгілерді қолдану арқылы жеңілдетуге және жеделдетуге болады. Бұл мақалада гетерогенді онтологиялық дизайн үлгілері (ODP) жүйесін қолданатын ғылыми саладағы онтологияны автоматтандырылған құру әдісі келтірілген. Бұл жүйеге онтологияны жасаушыларға бейімделген ODP және табиғи тілдегі мәтіндерден алынған ақпаратпен онтологияны байыту үшін пайдалануға болатын автоматты түрде жасалатын лексикалық және синтаксистік үлгілер кіреді.

Тірек сөздер: ғылыми пәндік сала, шаблондар, онтологиялық дизайн, мазмұн шаблондары, шаблондарды автоматты түрде құру, онтологияны толықтыру.

Ю.А. Загорулько

Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск, Россия

РАЗРАБОТКА МЕТОДА АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ НАЗВАНИЙ СУЩНОСТЕЙ ОНТОЛОГИИ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Аннотация. Онтологии предметной области играют решающую роль в организации, совместном использовании и повторном использовании знаний, относящихся к предметной области, в программных системах. Однако разработка онтологии программного обеспечения является трудоемким и сложным процессом. Для создания такой онтологии необходимо проанализировать большое количество соответствующих публикаций. Эта задача по обогащению онтологии данными из этих источников может быть упрощена и ускорена за счет использования лексических и синтаксических шаблонов, полученных на основе онтологического проектирования. В данной статье представлен метод автоматизированного построения онтологий в рамках научной области, использующий систему гетерогенных онтологических шаблонов проектирования (ODP). Эта система включает в себя ODP, адаптированные для разработчиков онтологий, и автоматически генерируемые лексические и синтаксические шаблоны, которые могут быть использованы для обогащения онтологий информацией, извлеченной из текстов на естественном языке.

Ключевые слова: научная предметная область, шаблоны, онтологический дизайн, шаблоны контента, автоматическая генерация шаблонов, пополнение онтологии.