ТЕЗАУРУСЫ РУССКОГО ЯЗЫКА В ВИДЕ ОТКРЫТЫХ СВЯЗАННЫХ ДАННЫХ

Усталов Д. А. (dmitry.ustalov@urfu.ru)

Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, Екатеринбург, Россия;
NLPub, Екатеринбург, Россия

Важной тенденцией последних лет являются открытые лингвистические данные, дающие исследователям и разработчикам возможность построения собственных решений на основе готовых и выверенных словарей, корпусов, тезаурусов, и других ресурсов. При этом опубликованные данные хранятся в разных форматах, что затрудняет их эффективное использование, а также привязывает пользователей к поставщику. Данная работа посвящена представлению популярных тезаурусов русского языка в виде открытых связанных данных: описаны существующие форматы данных и подходы к их преобразованию, выполнено отображение трёх популярных открытых русских тезаурусов в схемы Семантической паутины. Полученный набор данных опубликован в формате Turtle и доступен на ресурсе NLPub для использования на условиях лицензии Creative Commons.

Ключевые слова: связанные данные, открытые данные, лексические ресурсы, интеграция данных, семантическая паутина, русский язык

RUSSIAN THESAURI AS LINKED OPEN DATA

Ustalov D. A. (dmitry.ustalov@urfu.ru)

Ural Federal University, Yekaterinburg, Russia;
NLPub, Yekaterinburg, Russia

Open linguistic data is a good recently established trend allowing both researchers and developers in the field of natural language processing to create their own applications using high-quality dictionaries, thesauri, corpora, etc. At the same time, the published open data are stored in different formats making them difficult to be used in an efficient way without falling within vendor lock-in. This paper is devoted to the problem of representing popular lexical resources of the Russian language in the form of Linked Open Data. It summarizes the recent work in the field of thesauri representation formats and approaches to converting such formats to those of Linked Data. It also proposes an approach to converting popular Russian thesauri to the vocabularies that are the essential parts of the Linguistic Linked Open Data Cloud. The proposed approach has been implemented in open source software and the resulted dataset has been made publicly available on NLPub in the Turtle format under the terms of a Creative Commons license.

Key words: Linked Data, Open Data, lexical resources, data integration, Semantic Web, Russian language

1. Introduction

Open linguistic data is a good recently established trend allowing both researchers and developers in the field of natural language processing to create their own applications using high-quality dictionaries, thesauri, corpora, etc. At that, the published open data are stored in different formats making them difficult to be used in an efficient way without falling within vendor lock-in. Hence, both the Semantic Web and natural language processing for Russian fields could benefit from representing the popular Russian thesauri in the form of Linked Data allowing applications to use the Semantic Web technologies including the powerful reasoning tools.

The work, as described in this paper, makes the following contributions: 1) it summarizes the recent work in the fields of thesauri formats, thesauri conversion approaches, and the thesauri for Russian, 2) it proposes and implements an approach to convert popular Russian thesauri to the form Linked Data, and 3) presents the results under a Creative Commons license. The rest of this paper is organized as follows. Section 2 is devoted to the survey on the related work. Section 3 proposes an approach to converting the thesauri to the Linked Data representation. Section 4 describes the implementation, presents and evaluates the resulted dataset. Section 5 concludes with final remarks and directions for the future work.

2. Related Work

The following three directions of the related work are considered: 1) thesauri representation formats, 2) approaches for converting thesauri into Linked Data, and 3) publicly available electronic thesauri for Russian.

2.1. Representation Formats

Princeton WordNet, the most recognized and influential electronic lexical ontology, has been represented in the form of semi-structured text files [3]. This format is widely used and the majority of WordNet's derivatives[1] utilize it to keep compatibility with the original database. In spite of the WordNet's popularity, there are many software libraries for various programming languages allowing one both to read and write the linguistic data in this format. However, dealing with such a format has a significant drawback: embedding it into an application requires either conversion into the form of relational database or utilizing special software to integrate the present data schema with WordNet's. This makes the resulting data model less uniform and requiring additional maintenance. There also exist many linguistic resources operating with their in-house developed custom data formats, hence their formats will be denoted as *custom*.

XML, eXtensible Markup Language, is designed to be both human-readable and machine-readable [12]. It is used by many production-grade software in order

[1] http://globalwordnet.org/

to describe data in almost every existing domain including lexical resources. The main advantages of XML are its wide support, representation uniformity, and the ability to be validated against a predefined schema. However, processing of large XML documents containing hundreds megabytes of data is computationally hard as it requires either construction of an expensive document object model (DOM) to be stored entirely in a computer memory, or through the use of stream-oriented SAX parsers, which are much less convenient in development.

Resource Description Framework (RDF) is proposed to be a uniform representation of any subject-predicate-object entity for the Semantic Web [13]. The RDF is just an abstract syntax that should be encapsulated by serialization formats, e. g. RDF/XML, Turtle, N3, etc. It should be noted that the Simple Knowledge Organization System (SKOS) is an RDF extension designed especially for vocabularies and thesauri [11]. The syntax of RDF triplets is simple to be understood by human, but it is difficult for an end user to support all the available representation formats. RDF/XML is the most popular hence it has the same drawbacks as XML.

The recently published ISO 25964 standard is designed to formalize fitting concepts, terms and relationships together to make a thesaurus [7]. It is focused on the knowledge engineering aspect of thesauri and does not propose a representation format leaving such a task to the user. Another standard, ISO 24613:2008 describes LMF, a lexical markup framework, which is aimed at providing an XML-based representation for vocabularies without dealing with word senses and their relations [4].

2.2. Linked Data Conversion

The problem of converting a thesaurus into the Linked Data has been approached for several times since the appropriate data schemas have been issued.

van Assem et al. in 2006 proposed a method to convert a thesaurus to the SKOS format and assessed the applicability of such a representation [10]. The proposed method has three steps: 1) analyzing thesaurus, 2) mapping data items into SKOS, and 3) creating a conversion program. This method has been evaluated on three thesauri (IPSV, GTAA and MeSH) and it has been confirmed that SKOS is suitable for thesauri resembling to the ISO 25964 standard.

McCrae et al. in 2011 presented a model called lemon (Lexical Model for Ontologies) that supports sharing terminological and lexicon resources on the Semantic Web [6]. The lemon model is an RDF-native form making it possible to expose a thesaurus to the Linked Data in the similar way as LMF does, and also represents word senses and their relations.

In 2012, Navigli & Ponzetto released BabelNet, which is a very large multilingual semantic network constructed automatically on the basis of WordNet, Wikipedia and other databases [8]. BabelNet integrates seamlessly into the Semantic Web[2] through the alignment to underlying data sources, and exposes itself in the form of RDF employing such vocabularies as SKOS and lemon.

[2] http://babelnet.org/rdf/

2.3. Thesauri for Russian

There are three notable electronic thesauri for Russian that are publicly available under open licenses: 1) RuThes-lite, 2) the Russian Wiktionary, 3) the Universal Dictionary of Concepts, and 4) Yet Another RussNet (the more detailed survey is presented on [5, 9]).

RuThes-lite[3] is a subset of the RuThes lexical ontology having been developed since 1994 for addressing the information retrieval tasks in various applications for the Russian language [5]. The format of the original RuThes is unknown, nevertheless RuThes-lite is available under the terms of the CC BY-*NC*-SA license in the form of quasi-structured HTML pages on the Internet representing approximately 26,000 concepts and 100,000 relations between them.

The Universal Networking Language[4] is a project led by the United Nations dedicated to the development of a computer language that replicates the functions of natural languages. The Russian version of its semantic network—the Universal Dictionary of Concepts—is contributed by the researchers from IITP RAS [2]. UNLDC is distributed under the CC BY-SA license containing approximately 62,000 of the universal words (UWs) and 90,000 links between them.

The Russian Wiktionary[5] is the eighth largest Wiktionary composed of more than 520,000 articles—one article represents a lexical entry—written by more than 120,000 users (only 164 users are active participants) since 2004. The native format of the Wiktionary pages is a quasi-structured wiki syntax, which is quite hard to parse. However, there exists the Wikokit[6] project that parses the Russian and English Wiktionaries and renders them in the machine-readable form of a relational database available under the terms of the CC BY-SA license.

Yet Another RussNet[7] is an open project established in 2013 and aimed at creation of a large electronic thesaurus for Russian through the use of crowdsourcing [1]. At the moment, this resource contains 111,895 words and approximately 18,000 synsets. Initially, the project deliverables had been available in the XML format with the correspondent XSD although recently the synsets have been made available in the CSV and RDF formats, which are more convenient to parse and to use. All the content is published under the CC BY-SA license. Yet Another RussNet includes the lexicon and synsets of the Russian Wiktionary among several others resources licensed under the same license.

It should be noted that all these resources utilize their own *custom* data representation formats embarrassing their evaluation and forcing end users into vendor lock-in. Since that the Yet Another RussNet project includes the lexicon and synsets of the Russian Wiktionary, only three resources will be considered in this study: RuThes-lite, UNLDC, Yet Another RussNet.

[3] http://www.labinform.ru/pub/ruthes/index.htm

[4] http://www.undl.org/

[5] https://ru.wiktionary.org/

[6] https://code.google.com/p/wikokit/

[7] http://russianword.net/

3. Representing the Thesauri

In order to represent the above-mentioned resources in the form of Linked Data, it is necessary to make the following assumptions. Firstly, the primary applications of the present work are natural language processing and information retrieval, thus the resulted resource may not cover the complete set of natural language entities and relations. Secondly, the resulted dataset should not reinvent the Linked Data vocabularies, but should use the popular ones as soon as possible. Finally, both humans and machines should easily understand the resulted data format.

Each thesaurus has been analyzed to find out how the data items can be mapped to the Linked Data vocabularies, and each thesaurus will be presented in a separate ontology. Since the RuThes-lite thesaurus is widely applied in various practical tasks, the types of its concept relations have been considered as the only concept relation types with one exception: the antonymy relation, which is widely used in UNLDC.

The choice of the Linked Data vocabularies is mostly inspired by that of BabelNet, hence the following vocabularies have been used: Simple Knowledge Organization System (SKOS) to represent concepts, Lexicon Model for Ontologies (lemon[8]) to represent lexical senses, lexical entries, definitions and usage examples, LexInfo[9] to represent the morpho-syntactic labels. RDFS, OWL and Dublin Core have expressed the ontology description. The Turtle format has been chosen to store the processing output because of its readability and popularity.

Table 1 demonstrates the result of thesaurus entities' mapping. The most challenging part of the mapping process was the selection of the appropriate concept relation representation. For instance, SKOS provides the special terms for expressing hypernymy and hyponymy, but does not provide such terms for holonymy and meronymy—although LexInfo does.

Table 1. Entities, relations, vocabularies

Entity/Relation	**Vocabulary Term**
Concept	`skos:Concept`
Lexical Sense	`lemon:LexicalSense`
Definition (Gloss)	`lemon:SenseDefinition`
Usage Example	`lemon:UsageExample`
Lexical Entry	`lemon:LexicalEntry`
Lemma	`lemon:Form`
Class-Subclass (is-a)	`skos:{broader,narrower}`
Part-Whole (part-of)	`lexinfo:{holonymTerm,meronymTerm}`
Asymmetric Association	`lemon:subsense`
Symmetric Association	`skos:related`
Antonomy	`lexinfo:antonym`
Sense-Concept Mapping	`lemon:{sense,isSenseOf}`

[8] http://www.lemon-model.net/

[9] http://lexinfo.net/

Entity/Relation	**Vocabulary Term**
Sense-Lexeme Mapping	*lemon*:{reference,isReferenceOf}
Sense-Definition Mapping	*lemon*:definition
Sense-Example Mapping	*lemon*:example
Lemma Indication	*lemon*:canonicalForm
Part-of-Speech	*lexinfo*:partOfSpeech

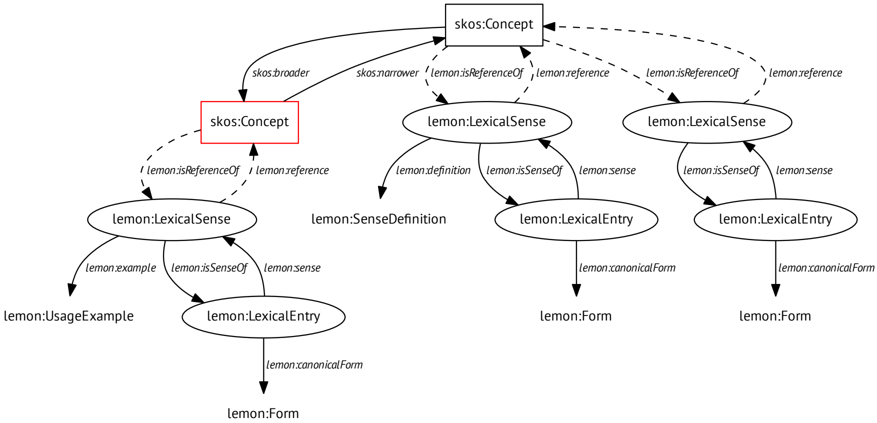

Fig. 1. Elements of the uniform ontology: concepts, lexical entries and their senses, definitions, usage examples, and lemmas

An example of the resulted ontology is depicted at Fig. 1. The present example shows two defined concepts: one has two lexical senses and is denoted as a hypernym to the other (red colored) concept having only one sense. Each sense is provided with the corresponding lexical entries. Each lexical entry hopefully has a canonical form (lemma). The picture also illustrates that the sense definitions and the usage examples are connected to the lexical senses instead of the concepts.

3.1. RuThes-lite

RuThes-lite comes in the form of four schema-less XML files representing lexical entries, concepts, their relations, and mappings between the concepts and lexemes. Despite RuThes-lite containing some valuable morpho-syntactic information, it is written in barely parsable form, and such information has been—unfortunately—omitted, i. e. the fields like `synt_type` and `pos_string`. Table 2 summarizes the mapping process.

Table 2. Mapping RuThes-lite to the Linked Data vocabularies

Data Item	Feature/Function	Property/Class
//entry[@id]	Lexical Entry Lemma Lemma Indication	*lemon*:LexicalEntry *lemon*:Form *lemon*:canonicalForm
//concept	Concept	*skos*:Concept

//entry_rel	Lexical Sense
	Sense-Concept Mapping
	Sense-Lexeme Mapping
		lemon:LexicalSense
		lemon:{sense,isSenseOf}
		lemon:{reference,isReferenceOf}

//rel[@name="ВЫШЕ"]	Class-Subclass (is-a)	*skos*:broader
//rel[@name="НИЖЕ"]		*skos*:narrower
//rel[@name="ЧАСТЬ"]	Part-Whole (part-of)	*lexinfo*:holonymTerm
//rel[@name="ЦЕЛОЕ"]		*lexinfo*:meronymTerm
//rel[@name="АСЦ1"]	Asymmetric Association	*lemon*:subsense
//rel[@name="АСЦ2"]		*lemon*:subsense
//rel[@name="АСЦ"]	Symmetric Association	*skos*:related

3.2. The Universal Dictionary of Concepts

UNLDC is published[10] in the form of CSV files representing universal words (UWs) and links between them. Since the UNLDC universal words are unambiguous by design, they have been mapped into lexical senses as described in Table 3. The main problem of the UNLDC mapping is the necessity to parse the domain-specific relations stored within such UWs as *tongue(icl>concrete_thing,pof>body)*, therefore such descriptors were omitted and the resulted dataset has no relations. The synsets[11] derived from the UWs have been mapped to concepts.

Table 3. Mapping UNLDC to the Linked Data vocabularies

Data Item	Feature/Function	Property/Class
Lemma	Lexical Entry Lemma Lemma Indication	*lemon*:LexicalEntry *lemon*:Form *lemon*:canonicalForm
Part-of-Speech	Part-of-Speech	*lexinfo*:partOfSpeech

Universal Word	Concept
	Lexical Sense
	Sense-Concept Mapping
	Sense-Lexeme Mapping
		skos:Concept
		lemon:LexicalSense
		lemon:{sense,isSenseOf}
		lemon:{reference,isReferenceOf}

Relation `icl`	Class-Subclass (is-a)	*skos*:{broader,narrower}
Relation `iof`		
Relation `pof`	Part-Whole (part-of)	*lexinfo*:{holonymTerm,meronymTerm}
Relation `equ`	Symmetric Association	*skos*:related
Relation `ant`	Antonomy	*lexinfo*:antonym

[10] https://github.com/dikonov/Universal-Dictionary-of-Concepts/tree/master/data/csv

[11] https://github.com/dikonov/Universal-Dictionary-of-Concepts/tree/master/data/misc

3.3. Yet Another RussNet

The Yet Another RussNet software is implemented in the Ruby on Rails framework with active use of the ActiveRecord object-relational mapping [9]. Table 4 shows that its data models[12] have been mapped to those of Linked Data.

Table 4. Mapping Yet Another RussNet to the Linked Data vocabularies

Data Item	Feature/Function	Property/Class
Word	Lexical Entry Lemma Lemma Indication Part-of-Speech	*lemon*:LexicalEntry *lemon*:Form *lemon*:canonicalForm *lexinfo*:partOfSpeech
Synset	Concept	*skos*:Concept
SynsetWord	Lexical Sense Sense-Concept Mapping Sense-Lexeme Mapping Sense-Definition Mapping Sense-Example Mapping	*lemon*:LexicalSense *lemon*:{sense,isSenseOf} *lemon*:{reference,isReferenceOf} *lemon*:definition *lemon*:example
Definition	Definition (Gloss)	*lemon*:SenseDefinition
Example	Usage Example	*lemon*:UsageExample

4. Results

The conversion and the supplementary programs have been implemented in the Ruby programming language. The resulted software is available on GitHub under the MIT license: https://github.com/nlpub/rtlod. During the implementation, it has become necessary to port the lemon and LexInfo vocabularies to the syntax of the used RDF.rb library, which resulted in releasing of the *rdf-lemon*[13] library for Ruby.

The resulted dataset consisting of the converted RuThes-lite, UNLDC and Yet Another RussNet thesauri in the Turtle format is available on NLPub: http://nlpub.ru/RTLOD. Thorough evaluation of such a resource is a very interesting topic that is complicated enough to conduct a specialized study. Nevertheless, in order to compare the resulted ontologies quantitatively, brief statistics of them have been calculated and demonstrated in the Table 5. The lexical intersection between the converted thesauri has also been assessed (Table 6).

It seems that Yet Another RussNet that is created through crowdsourcing has the widest lexical coverage although the number of its concepts is relatively low. It is also the only resource provided with the word usage examples due to its crowdsourcing schema requiring users to consider such examples [1]. High number of lexical senses is caused by the presence of many duplicated synsets generated by users. Despite this resource having no established concept relations, it still may be still useful

[12] http://nlpub.ru/YARN/API

[13] https://github.com/nlpub/ruby-rdf-lemon

as a synonyms' dictionary in some applications. Both RuThes-lite and UNLDC are mature resources with developed concept relations [5], but UNLDC is a dictionary of a controlled language [2], hence its number of concepts is significantly smaller, although these concepts are tightly connected to each other.

Table 5. Resulted datasets

# of	RuThes-lite	UNLDC	Yet Another RussNet
Lexical Entries	96,700	56,313	111,895
Part-of-Speech Tags	n/a	56,313	111,821
Concepts	26,354	8,896	17,492
Relations	98,976	n/a	0
Lexical Senses	115,106	20,366	69,981
Definitions	10,701	8,896	7,641
Usage Examples	0	0	2,991

Table 6. Lexical intersection

		# of common lexical entries
RuThes-lite	UNLDC	18,596
UNLDC	Yet Another RussNet	26,088
Yet Another RussNet	RuThes-lite	37,920

5. Conclusion

The author believes that the present work—especially the published dataset and software—could facilitate the development of the modern linguistic resources for Russian among their integration into the Linguistic Linked Open Data Cloud[14]. Given the openly published resources, a user can choose between them in order to pick the best option for the particular application. Moreover, it contributes a lot into simplifying conducting thorough studies of these resources by such benchmarks as word sense disambiguation competitions. The present mapping approach is general and could be freely used for adopting more thesauri of the Russian language.

There are several reasons for future work. Firstly, it may be useful to assess the lexical coverage of the given resources with these representations. Secondly, since these datasets are Linked Data, it may be interesting to estimate alignments between concepts of them. Finally, end users may consume the deliverables of this work and link their own data to these.

[14] http://linghub.lider-project.eu/llod-cloud

Acknowledgements

This work is supported by the Russian Foundation for the Humanities, project № 13-04-12020 “New Open Electronic Thesaurus for Russian”. The author would like to thank Natalia V. Loukachevitch from the Lomonosov Moscow State University for the provided RuThes-lite dataset in the XML format, and Vladimir V. Ivanov from the Kazan Federal University and Vyacheslav Dikonov from the Institute for Information Transmission Problems of the RAS for fruitful discussions and valuable suggestions. The author is also grateful to the anonymous referees who offered very useful comments on the present paper.

References

1. *Braslavski P., Ustalov D., Mukhin M.* (2014), A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, pp. 101–104.
2. *Dikonov V. G.* (2013), Development of lexical basis for the Universal Dictionary of UNL Concepts, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Bekasovo, Russia, pp. 212–221.
3. *Fellbaum C.* (1998), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, USA.
4. *Francopoulo G.* (2013), LMF: Lexical Markup Framework, Wiley-ISTE, London, UK.
5. *Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I.* (2014), RuThes-Lite, a Publicly Available Version of Thesaurus of Russian Language RuThes, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Bekasovo, Russia, pp. 340–349.
6. *McCrae J., Spohr D., Cimiano P.* (2011), Linking Lexical Resources and Ontologies on the Semantic Web with Lemon, Springer: Lecture Notes in Computer Science, Vol. 6643, pp. 245–259.
7. *National Information Standards Organization.* (2011), ISO 25964 – the international standard for thesauri and interoperability with other vocabularies, available at: http://www.niso.org/schemas/iso25964/
8. *Navigli R., Ponzetto S. P.* (2012), BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence, Vol. 193, pp. 217–250.
9. *Ustalov D.* (2014), Enhancing Russian Wordnets Using the Force of the Crowd, Springer: Communications in Computer and Information Science, Vol. 436, pp. 257–264.
10. *van Assem M., Malaisé V., Miles A., Schreiber G.* (2006), A Method to Convert Thesauri to SKOS, Springer: Lecture Notes in Computer Science, Vol. 4011, pp. 95–109.
11. *World Wide Web Consortium W3C.* (2004), SKOS Simple Knowledge Organization System, available at: http://www.w3.org/2004/02/skos/
12. *World Wide Web Consortium W3C.* (2008), Extensible Markup Language (XML) 1.0 (Fifth Edition), available at: http://www.w3.org/TR/2008/REC-xml-20081126/
13. *World Wide Web Consortium W3C.* (2014), RDF 1.1 Concepts and Abstract Syntax, available at: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/