Technologies for automation creation of an ontology of urbanonyms in the aspect of historical changes (on the example of Almaty)

Technologies for automation creation of an ontology of urbanonyms in the aspect of historical changes (on the example of Almaty)

Vladimir Barakhnin[1], Madina Mansurova[2], Bekzhan Dossanov[3] and Marzhan Kyrgyzbayeva[3]

[1] Federal Research Center for Information and Computational Technologies, Lavrentiev av. 6, Novosibirsk, 630090, Russian Federation,
[2] Novosibirsk State University, Novosibirsk, 1, Pirogova str., Novosibirsk, 630090, Russia
[3] Al-Farabi Kazakh National University, Republic of Kazakhstan, 050040, Almaty city, 71 Al-Farabi Ave.

e-mail: bar@ict.nsc.ru, olgakozhemyakina@mail.ru, mansurova.madina@gmail.com, dossanbekzhan@gmail.com, marzhan.kyrgyzbaeva@gmail.com

Abstract. When creating geoinformation systems of a city scale, relation this or that information from Internet, there comes the task of creation an ontology of urbanonyms taking into account their historical changes. The account of historical changes is necessary, for example, to process messages about urban event from blogs: since more and more representatives of the middle and older generations are becoming active Internet users, the massages often contain the former names of urbanonyms. Let us note that it is the accounting of historical changes that is required to solve this problem that determines the need to create not a thesaurus, which is sufficient, as shown in [1], to take into account geographical names commonly used (at least in natural science articles) in their actual from, but an ontology. Taking into account the specifics of the task of creating an ontology of Almaty, in should be bilingual: in the Kazakh and Russian languages.

1. 1Introduction

Ontology in this work is considered as a general part of the domain model connecting knowledge about the world with knowledge about language. In an ontology, connections between descriptors are not only explicitly stated, but also classified. Many works [1–3] are devoted to the issues of automated construction of ontologies. At the same time, it is emphasized that this is a laborious and difficult task. One of the approaches is the creation of ontologies based on the subject index of specialized encyclopedias [4, 5]. This technology provides a highly qualified description of the subject area using reliably verified terms.

Ontologies related to the representation of geographic objects, linking a system of concepts and terms to geographic regions are of particular value. They represent the real world in an abstract way and offer a universal and stable framework for discontinuity and data reconciliation. A special place among such ontologies is occupied by the ontologies of urbanonyms. According to [6], this is due to the fact that in the next three decades, more than half of the world's population will live in cities. Although cities cover about 2% of the world's land mass, they generate over 80% of world GDP [7],

which represents a large economic footprint. Cities also contribute over 70% of the world's greenhouse gas emissions [8].

A series of papers is devoted to the development of ontologies for urban development [9], which provides a valuable overview of the main current problems in the field of urban ontologies and many useful and different approaches [10–13].

Ontology-based Semantic Web technologies have extremely high potential and practical impact, providing the basis for new e-services in urban ecosystems [14]. The use of Semantic Web technologies makes it possible to develop a set of solutions that can work in parallel with the Internet of Things, as well as with embedded systems. The resources for the Semantic Web are structured information data extracted from systems such as DBpedia [15], GeonamesDbpedia [16], Wikipedia [17], which provide encyclopedic knowledge about many different domains.

Over the past decade, opportunities have expanded to address complex urban landscape complexity through a systems approach, taking into account environmental, social and economic requirements. Moreover, the ontology-based approach is very successful. The use of IoT technologies together with ontologies allows obtaining information in real time and informing for decision-making [18].

The purpose of this work is to develop an ontology of urbanonyms using the example of Almaty, which will give a complete picture of knowledge about the geographical objects of the city, including concepts, relationships and properties. Due to the fact that our time is characterized by dynamic changes associated with the creation of new ones, a change in status, renaming of existing geographic objects, the ontology is built in the aspect of historical changes. The historical changes required to solve this problem necessitate the creation of not a thesaurus, which is sufficient, as shown in [19], to take into account geographical names commonly used (at least in natural science articles) in their current form, but an ontology. The paper presents a developed algorithm for the automatic creation of an ontology, as well as a web application that implements it.

2. Methods for creating an ontology of urbanonyms

2.1. The structure of ontology

Let us describe the structure of ontology. As is noted in [20], at the initial stage we mean creating only an ontology framework containing only brief information about entities, and their more detailed description will take place during the functioning of the information system.

The main entities description by the ontology of urbanonyms:

a) names of city districts

b) names of squares (agoronyms), avenues and streets

c) names of separate building (khoronyms)

The description of the entity “name of city districts” according to a minimum scheme consists of the necessary set of triplets of attributes of the form “name” – “valid or not” – “year of assignment of the name” (the latter is optional), while if the district is formed by the merger of two or more districts, this set will contain all the pairs corresponding to these areas, and if, on the contrary the area is formed as part of a divided area, the name of this area will be present in the description of all areas into which it was divided.

In the extended scheme, colloquial names of districts may be added (of course, without attributes “year of assignment of the name” and “valid or not”).

The description of entities “agorononyms” and “godonyms” according to a minimum scheme consists of an identifier of the district (or, if needs, -identifiers of several district) where the square, avenue or street is located, as well as a necessary set of triplets of attributes of the form “name” – “valid or not” – “year of assignment of the name” (the latter is optional), while if the street is formed of several streets, this set will includes pairs of attributes corresponding to all these streets.

In the extended scheme, colloquial names of square avenues and streets may be added (of course, without attributes) “year of assignment of the name” and “valid or not”.

And finally, the description of the entities “horonyms” according to a minimum scheme consists of the identifier of the street where the object is located (in this case we are talking about a legal address), the house number as well as the necessary set of triplets of attributes of the form “name” – “valid or not” – “year of the assignment of the name” (the latter is optional) while there can be several valid names which may have an official character in this or that degree. For example, in the article of the same name on Wikipedia there are the following four names of the university: the Kazakh national university, the Kazakh national university named after al-Farabi, al-FarabiKazNU, KazNU, (and it should be taken into account that “named” can be reduced to “n”, but it is better to take into account such language peculiarities not in the ontology but in word processing software).

In the extended scheme, colloquial names of organizations and building they occupy (of course, without attributes “year of the assignment of the name” and “valid or not”) as well as additional addresses (for organizations occupying several building in fact) may be added.

2.2. The ontology creation algorithm

The process of ontology creation consists of the following stages:

1. creation of a list of urbanonyms;
2. compiling for each urbanonym a list of its historical changes (if possible, including the date of change);
3. Reference of urbanonyms from the created list to the city map.

Let us describe the technology of automation of stages 1 and 2 based on the use of information contained in Wikipedia in relation to the city of Almaty (for other large cities the situation with Wiki-articles is similar).

a) The historical aspect of the names of Almaty districts is reflected in Wikipedia.
https://ru.wikipedia.org/wiki/Административно-территориальное_деление_Алма-Аты

The number of urbanonymes of this type is not great, therefore the information can be entered into the ontology database (in the Kazakh and Russian languages) manually without using automatic parsing of the webpage.

b) The historical aspect of the names of squares (agoronyms), avenues and streets (godonyms) of Almaty is reflected in Wikipedia, too:
https://ru.wikipedia.org/wiki/Список_улиц_Алма-Аты

There are more than 2200 names in the list (taking into account the fact that for some godonyms each of 2–3 consecutive historical names has a separate line), therefore, it is advisable to enter if into the database by means of automatic parsing. However, the matter is complicated by the fact that the historical aspect of the names is written without any common pattern, therefore, it is necessary to fix the main types of regular expression that determine the historical changes in the godonym as well as to check to exclude the possibility of entering into the database duplicate information from different lines of the wiki-article. Unfortunately, in the Kazakh language there is no wiki-article with a list of godonyms of Almaty in their historical aspect, therefore, to automate the translation of the Russian language ontology to the Kazakh language, if is expedient to use the list of wiki-articles about the streets of Almaty:
https://kk.wikipedia.org/wiki/Санат:Алматы_көшелері

c) creation of a list of names of separate buildings (khoronyms) is the least algorithmizable subproblem.

One of the possible algorithms of partial automation of its solution is based on the proposed by us in [1, 21] algorithm of extraction of metadata from homogenous weakly structured documents. The algorithm based on man-machine interactions typical for intelligent information systems comes down to performing sequential operations:

- creation of a pattern for the site being processed;
- creation of a list of addresses where the documents are located;
- keeping information up to date.

A document template is needed to automatically highlight its main metadata (in the task considered – as a minimum, the name of khoronym and its address). In each concrete case, the template is created comparatively easily using the language of regular expressions in this or that format (RegEx, Posix, etc.). However, the problem is that different sites, even those of one type (for example, wiki-pages referring to different categories) have a different structure of description and presentation of documents. Therefore it is desirable to create an algorithm realized in the form of a web-application which allows a user, even if he does not possess the languages of processing regular expressions, to generate templates for different sites of course, if a user knows the foundations of using regular expressions, he can indicate expressions describing the data format, as such indication is able in some situations to increase the efficiency of the algorithm performance.

In a general case, it is enough for the user-cataloguer to enter in the form fields the tags surrounding the values of each of the metadata elements of one document of the processed site in the HTML-code of the article, and also specify the data separator in case of a plurality of some metadata element, after which the template of the web-page of the site document is created and saved.

The most general information is contained in the section:

https://ru.wikipedia.org/wiki/Категория:Алма-Ата

In order to get to a specific web page with data about a city object, it is necessary to go through a number of subcategories. For example, for this or that university such subcategories subsequently will be:

https://ru.wikipedia.org/wiki/Категория:Образование_в_Алма-Ате
https://ru.wikipedia.org/wiki/Категория:Учебные_заведения_Алма-Аты
https://ru.wikipedia.org/wiki/Категория:Высшие_учебные_заведения_Алма-Аты
https://ru.wikipedia.org/wiki/Категория:Университеты_Алма-Аты

The problem of processing lower-level pages is that these pages are written very heterogeneously: for example, on the wiki-pages of most research institutes of Almaty there is their legal address but sometimes there is not. Such cases should be detected automatically for the subsequent implementation of the missing information manually.

After processing all categories we will get a sufficiently complete list of khoronyms of Almaty with the addresses of the corresponding objects.

Attention should be paid to the fact that the structure of description of objects is not a tree in a sense of the graph theory: one object may have several parent subcategories. For example,

https://ru.wikipedia.org/wiki/Категория:Музеи_Алма-Аты
you can come as a chain
https://ru.wikipedia.org/wiki/Категория:Алма-Ата
https://ru.wikipedia.org/wiki/Категория:Культура_Алма-Аты
so along the chain
https://ru.wikipedia.org/wiki/Категория:Алма-Ата
https://ru.wikipedia.org/wiki/Категория:Архитектура_Алма-Аты
https://ru.wikipedia.org/wiki/Категория:Здания_и_сооружения_Алма-Аты

Therefore, it is necessary to include for possible duplication in the procedure for adding a new object to the ontology.

Let us note that in the process of parsing the wiki-pages, only the main official name of the organization, the legal address (if available in the organization card on the wiki-page) as well as the previous names of the organization (if available in the organization card on the wiki-page) will be entered. Variants of the official name of organization and, what is more, its non-official names will be entered in the process of further development of ontology including the use of crowdsourcing.

3. Results

The stage of designing and building an ontology indicating classes, data relations, object relations was carried out in the OWL environment [22]. Below are screenshots of some classes of the constructed ontology (Figures 1–5).

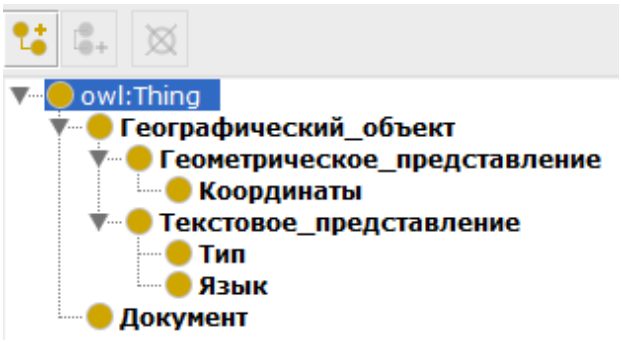

Figure 1. Ontology class hierarchy

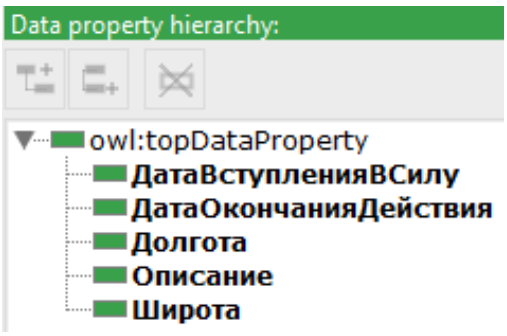

Figure 2. Ontology data relationships

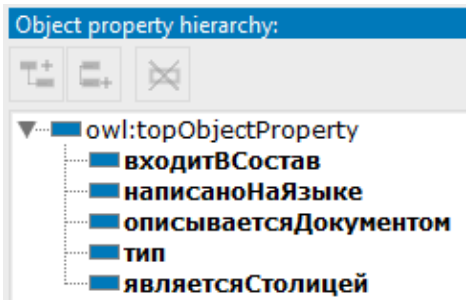

Figure 3. Ontology object relationships

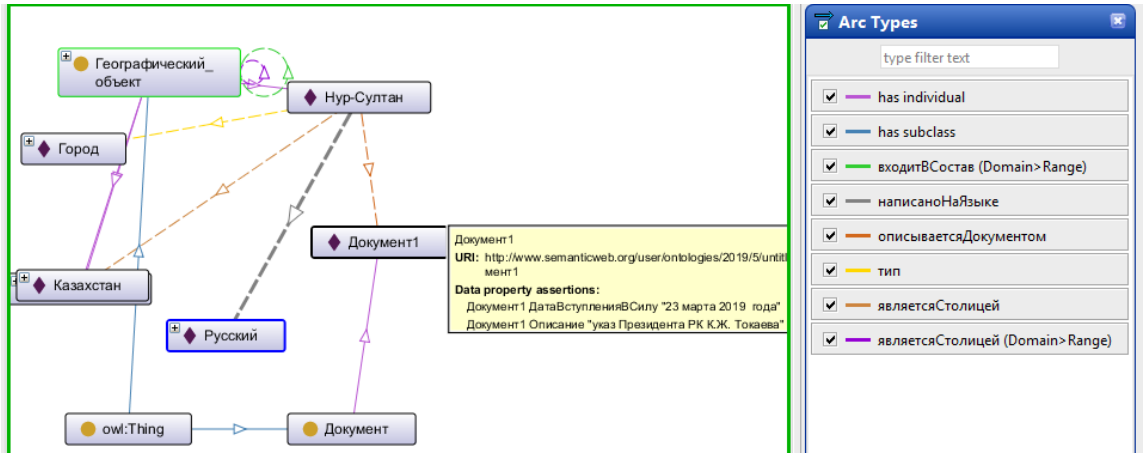

Figure 4. Example of relations of the "Nur-Sultan" object

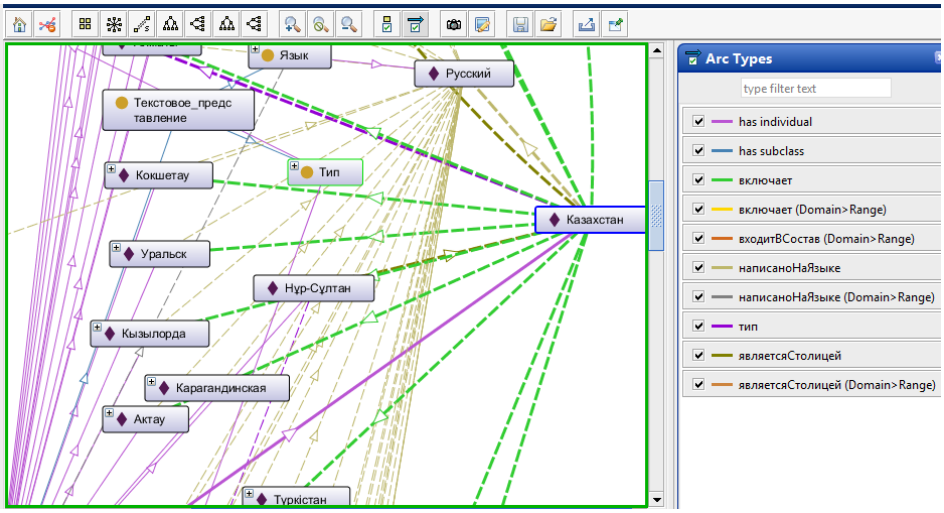

Figure 5. Example of relations of the object "Kazakhstan"

Figure 6 shows the architecture of a web application developed based on the ontology of urbanonyms. The application is implemented in the Java programming language using the popular SpringBoot framework using the DataRest module [23]. The data is stored in the MySql database. An interface for interacting with data has been created, eliminating the need to directly interact with the database. Practice has shown that this approach simplifies interaction with the storage, and the modularity of the system makes it flexible.

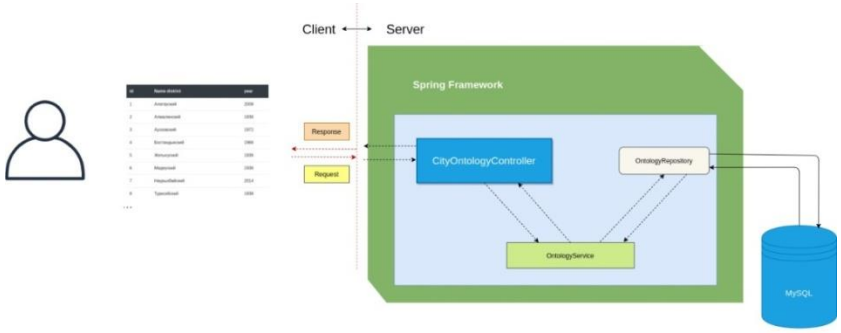

Figure 6. The architecture of the developed web application

Figure 7 shows an application web page with short information about the city and a list of districts. Figure 8 displays information about the street: the current as well as the previous street names, a list of objects located on this street with addresses.

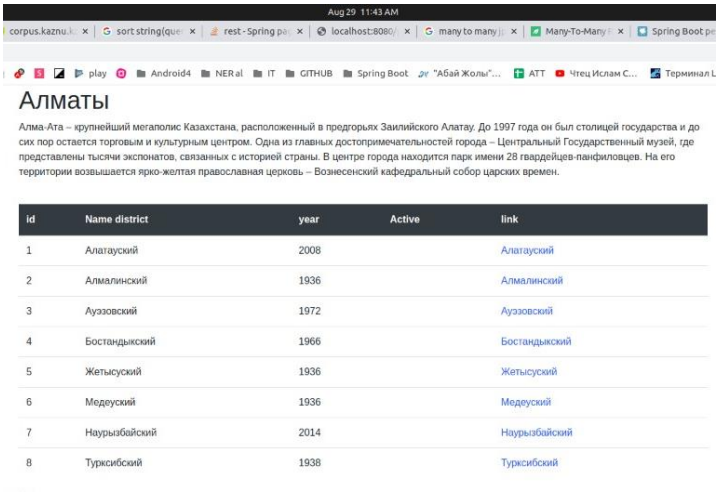

Figure 7. The architecture of the developed web application

Figure 8. Display street information with a historical perspective

4. Conclusion

Over the past few years, various ontologies have been developed related to the development of the city's information infrastructure. As stated in the introduction, the aim of this work was to develop an ontology of urbanonyms and to identify its potential role in urban development. This paper describes the technology of automating the creation of an ontology of urbanonyms in the aspect of historical changes (on the example of Almaty). The paper presents a developed algorithm for the automatic creation of an ontology, as well as a web application that implements it.

References

[1] Shokin Yu I, Fedotov A M and Barakhnin V B 2010 *Information search problems* (Novosibirsk: Science)
[2] Staab S and Stuckenschmidt H (Eds.) 2006 *Semantic web peer-to-peer, decentralized management and exchange of knowledge and information* (Springer)
[3] Knogstie J, Halpin T and Siau K 2005 *Information modeling methods and methodologies* (Idea group publishing)

[4] Barakhnin V B and Nekhaeva V A 2007 The technology of creating a subject area thesaurus based on the subject index of an encyclopedia *Computational technologies* **12** (2) pp 3–9
[5] Barakhnin V B 2003 Development of the thesaurus of the "Mathematics" subject area *Computational technologies* **8** (3) pp 111–115
[6] Yiqun Chen, SoheilSabri, Abbas Rajabifard, Muyiwa Elijah Agunbiade 2018 An ontology-based spatial data harmonisation for urban analytics *Computers, Environment and Urban Systems* DOI: 10.1016/j.compenvurbsys.2018.06.009.
[7] *Urban world*: Mapping the economic power of cities. 2011. https://www.mckinsey.com
[8] *Global Report on Human Settlements* 2011: Cities and Climate Change. https://unhabitat.org/global-report-on-human-settlements-2011-cities-and-climate-change
[9] Teller J, Lee J R and Roussey C (eds) 2007 Ontologies for Urban Development *Studies in Computational Intelligence* **61** (Springer, Berlin, Heidelberg)
[10] Caglioni M and Rabino G A 2007 Theoretical approach to urban ontology: a contribution from urban system analysis *Studies in Computational Intelligence* **61** (Springer, Berlin, Heidelberg)
[11] Borgo S 2007 How Formal Ontology can help Civil Engineers *Studies in Computational Intelligence* **61**
[12] Kaza N and Hopkins L D 2007 Ontology for Land Development Decisions and Plans. *Studies in Computational Intelligence* **61**
[13] Métral C, Falquet G and Vonlanthen M 2007 An Ontology-based Model for Urban Planning Communication *Studies in Computational Intelligence* **61**
[14] Tarek Abid, Mohamed ridda Laouar, Hafed Zarzour and Mohamed TarekKhadir 2016 Towards a smart city ontology *IEEE/ACS 13th International Conference of Computer Systems and Applications* (AICCSA)
[15] DBpedia https://wiki.dbpedia.org/
[16] Geonames Dbpedia http://dbpedia.org/page/GeoNames
[17] Wikipedia https://www.wikipedia.org/
[18] Corentin Kuster, Jean-Laurent Hippolyte and Yacine Rezgui 2020 The UDSA Ontology: An Ontology to Support Real Time Urban Sustainability Assessment *Advances in Engineering Software* **140** DOI: 10.1016/j.advengsoft.2019.102731
[19] Barakhnin V B, Zhizhimov O L, Kupershtokh A A, Skachkov D M and Fedotov A M 2012 Algorithm for extracting geographic names from text documents that reflect the content *Vestnik NSU. Series: Information Technology* **10** (1) pp 109–120
[20] Barakhnin V B and Fedotov A M 2013 Building a model of factual search *Vestnik NSU. Series: Information Technology* **11** (4) pp 16–27
[21] Barakhnin V B and Vedernikov V V 2006 Automated cataloging of electronic journal publications *Proceedings of the international conference "Computational and information technologies in science, technology and education"* (Kazakhstan, Pavlodar, September 20–22) **I** pp 209–214.
[22] OWL Web Ontology Language Reference https://www.w3.org/TR/owl-ref/
[23] SpringBoot and DataRest setup instructions URL: https://spring.io/guides/gs/accessing-mongodb-data-rest/