Barometr Regionalny
Rok 2023 Tom 19 nr 1

An Ontological Generator of Knowledge Tests from Texts on the Example of the Knowledge of the Geography of Poland

Krzysztof Pancerz, Andrzej Burda
Academy of Zamość, Poland

Piotr Grochowalski, Wiesław Paja
University of Rzeszów, Poland

Abstract

In the paper, we describe the construction of a prototype of an ontological generator of knowledge tests from texts. Closed-ended questions for knowledge tests are generated on the basis of key concepts extracted from texts as well as on the basis of the knowledge structures represented in the form of the OWL ontology. The OWL ontology, presented in the paper as an example, covers selected issues in the field of the geography of Poland. All of the important stages in the presented research are depicted — i.e., the creation of the OWL ontology, among others, on the basis of the publicly available information resources, developing a mechanism for generating the questions, and a prototypical implementation of the generator.

Keywords: OWL ontology, generation of knowledge tests, semantic data analysis, geography of Poland

DOI: 10.56583/br.2165

Introduction

Automation of the generation of knowledge tests is becoming an essential need in the context of an increasing number of online testing tools. The ability to quickly prepare questions for tests covering the indicated range of issues is a great help for people conducting various types of exams. In our research, a prototype of a generator of knowledge tests, based on semantic resources of knowledge structures in specific fields, has been developed. First of all, these sources take the form of ontology. Ontologies are an effective tool of knowledge representation in modern computer tools (Neches et al. 1991). In particular, OWL ontologies are widely used.[1] OWL is an ontology language designed for the Semantic Web. There are many definitions and interpretations of the term ontology in the literature (Guarino and Giaretta 1995). In general, the knowledge included in ontologies is expressed as a set of concepts together with the relationships which have been defined between them, comprising the vocabulary from a given area (Neches et al. 1991).

The approach that was used in our prototype can be seen as a white box approach. The structure of the knowledge on the basis of which the knowledge tests are generated is explicitly presented

1. See: OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004 edited by Mike Dean and Guus Schreiber, available at https://www.w3.org/TR/2004/REC-owl-ref-20040210/.

E-mail addresses and ORCID digital identifiers of the authors
Krzysztof Pancerz • e-mail: krzysztof.pancerz@akademiazamojska.edu.pl • ORCID: 0000-0002-5452-6310
Andrzej Burda • e-mail: andrzej.burda@akademiazamojska.edu.pl • ORCID: 0000-0002-1962-0264
Piotr Grochowalski • e-mail: piotrg@ur.edu.pl • ORCID: 0000-0002-4124-412X
Wiesław Paja • e-mail: wpaja@ur.edu.pl • ORCID: 0000-0002-6446-036X

in the form of ontology which is directly readable for humans. This approach is in opposition to the black box approaches represented recently by very popular tools which use the so-called large language models (LLMs) (Bowman 2023). One of the most popular tools of this kind is GPT-4 (GPT-4 Technical Report 2023). LLMs are models consisting of artificial neural networks with a huge number of parameters (tens of millions to billions), trained on large quantities of unlabeled texts containing up to trillions of tokens, using self-supervised learning or semi-supervised learning (Cheng 2021). Black boxes refer to systems with internal workings that are invisible to the user.

In our research, we distinguished several important stages:

- Developing the OWL ontology of the knowledge structures for an exemplary field (selected issues in the field of the geography of Poland). The knowledge from this field can be easily structured in the form of the OWL ontology. In the presented prototype, we used a part of the ontology of places in Poland (Pancerz, Grochowalski, and Derkacz 2016) which had been created in our earlier research. This stage is described in Section 1: OWL ontology.
- Developing a mechanism for using publicly available resources of the knowledge of selected issues in the field of the geography of Poland to generate a part of the ontology, especially the part including individuals and relationships between them. This stage is described in Section 2: Data acquisition and generation of the OWL ontology.
- Developing a mechanism for generating questions with possible answers (correct and incorrect) for knowledge tests. In our research, we focused on knowledge tests including closed-ended questions (single-choice questions or multiple-choice questions). The question generation mechanism was based on several key operations: tagging documents (texts), matching tagged documents to the OWL ontology of the knowledge structures, and generating questions on the basis of matching results. This stage is described in Section 3: Generator implementation.
- Implementing a prototype of the ontological generator of knowledge tests from texts in the form of REST-API as well as in the form of the Web application. This stage is described in Section 3: Generator implementation.

The main idea of the described prototype is to generate questions from documents with texts indicated by the user (PDF files, web pages). First of all, the user can select the set of documents from which the questions are generated as well as determine the type of these questions. This functionality facilitates the quick creation of questions from different documents — e.g., those previously indicated for reading to people taking knowledge tests, etc.

1 OWL ontology

One of the key decisions to take in the ontology development process is to select the language in which the ontology will be implemented. Our ontology has been created in accordance with the OWL 2 Web Ontology Language (shortly OWL 2).[2] The OWL ontology consists of three components:

- classes (representations of concepts from the domain of interest, interpreted as sets that contain individuals),
- individuals (instances of classes that represent objects in the domain of interest), and
- properties (binary relations on individuals).

There are two main types of properties in OWL 2:

- object properties linking an individual to an individual, and
- data properties linking an individual to a data value.

Classes and their hierarchical structure as well as object properties and data type properties were defined for the OWL ontology of the knowledge structures covering selected issues in the field of the geography of Poland (e.g., cities and the administrative structure of Poland, National Parks, water objects, landforms — mountains and mountain ranges). At this stage, we partially used the ontology of places in Poland (Pancerz, Grochowalski, and Derkacz 2016) which had been created in our earlier research. Selected classes and examples of their individuals in the created OWL ontology are collected in table 1. The hierarchy of classes and object properties is shown in figure 1 (in Polish).

2. See: OWL Web Ontology Language Reference..., op. cit.

Table 1. Selected classes and examples of their individuals

English term	Polish term	Examples of individuals (in Polish)
Classes representing categories of administrative districts		
Country	kraj	Polska
Voivodship	województwo	lubelskie, podkarpackie
County	powiat	powiat hrubieszowski, powiat zamojski
City with County Status	miasto na prawach powiatu (powiat grodzki)	miasto Lublin, miasto Zamość
Commune	gmina	–
Urban Commune	gmina miejska	gmina Lublin, gmina Zamość
Urban-Rural Commune	gmina miejsko-wiejska	gmina Szczebrzeszyn, gmina Zwierzyniec
Rural Commune	gmina wiejska	gmina Nielisz, gmina Radecznica
Classes representing categories of places		
Place	miejscowość	–
City/Town	miasto	Lublin, Zamość
Village	wieś	Nielisz, Radecznica
Classes representing physiographic objects		
Physiographic Object	obiekt fizjograficzny	–
Landform Object	obiekt ukształtowania terenu	–
Mountain	góra	–
Mountain Range	pasmo górskie	Karpaty, Sudety
Peak	szczyt	Rysy, Śnieżka
Water Object	obiekt wodny	–
Flowing Water	woda płynąca	–
Canal	kanał	Kanał Elbląski, Kanał Gryfiński
River	rzeka	Odra, Wisła
Standing Water	woda stojąca	–
Lake	jezioro	Jezioro Białe, Jezioro Wigry
Sea	morze	Morze Bałtyckie
Classes representing categories of forms of nature protection		
Form of Nature Protection	forma ochrony przyrody	–
National Park	park narodowy	Bieszczadzki Park Narodowy, Poleski Park Narodowy
Landscape Park	park krajobrazowy	Krasnobrodzki Park Krajobrazowy, Skierbieszowski Park Krajobrazowy
Classes representing categories of physical-geographical units		
Physical-Geographical Unit Protection	jednostka fizycznogeograficzna	–
Physical-Geographical Macroregion	makroregion fizycznogeograficzny	Europa Zachodnia
Physical-Geographical Megaregion	megaregion fizycznogeograficzny	Pozaalpejska Europa Środkowa
Physical-Geographical Mesoregion	mezoregion fizycznogeograficzny	Niż Środkowoeuropejski, Wyżyny Polskie

The visualization of the hierarchy of classes was prepared using the Simple Ontology Visualization API (SOVA). SOVA is a Protege plug-in for full ontology visualization.[3] The yellow-rounded rectangles represent classes. Each arrowhead points to the superclass. There are two equivalent classes, *Mountain* and *Peak* (*Góra* and *Szczyt* in Polish), which are marked with a two-way arrow. The blue-rounded rectangles represent object properties which link individuals together. In figure 1, domains and ranges of object properties are marked. Object properties are a key element of our prototypical generator because it is mainly on their basis that the structure of the questions is built. For example, we distinguished the following object properties:

- *flows into* (*wpływa do* in Polish), *flows through* (*przepływa przez* in Polish), *lies on* (*leży nad* in Polish)
- *is home to* (*jest siedzibą dla* in Polish), *is located in* (*ma siedzibę w* in Polish)

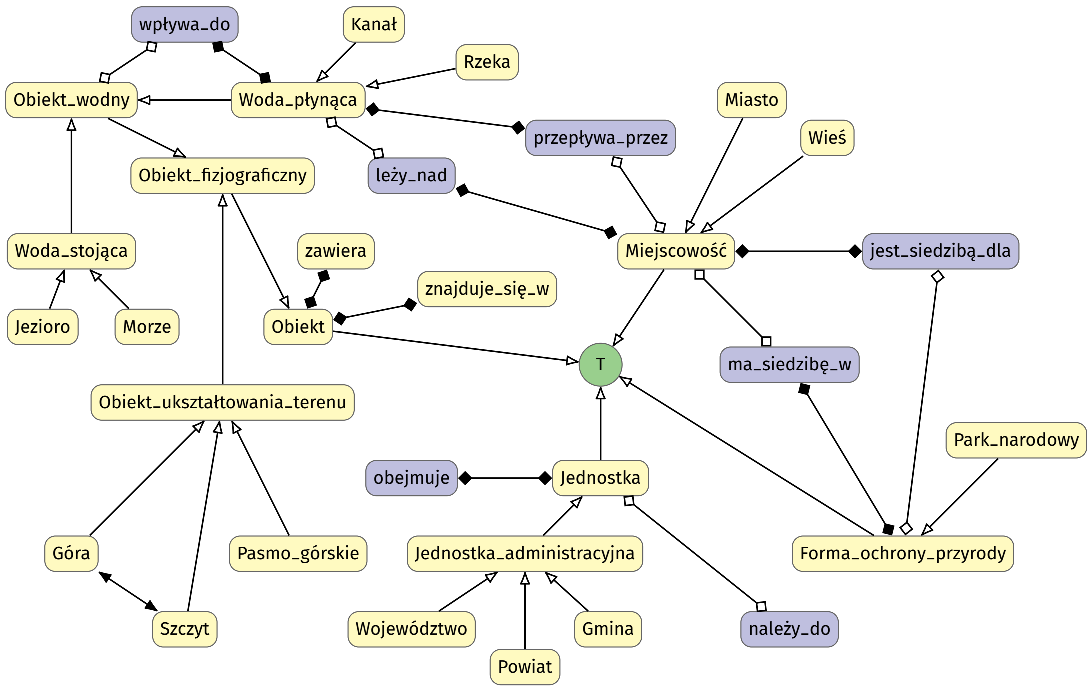

Figure 1. The OWL ontology of the knowledge structures covering selected issues in the field of the geography of Poland (a fragment)

2 Data Acquisition and Generation of the OWL Ontology

Data acquisition and knowledge exploration mechanisms were developed to generate the OWL ontology. The developed mechanisms obtain data from available resources in the form of web documents and spreadsheet files. The developed mechanisms use, among others, text data processing methods and web scraping methods. The ontology generator usage environment is depicted in figure 2.

A proper script was developed for the R environment[4] that implements the developed mechanisms of data acquisition and knowledge extraction and generates the OWL ontology (classes) and their hierarchical structure, individuals, object properties and data type properties of the knowledge covering selected issues in the field of the geography of Poland. The ontology was recorded in the OWL/XML format. The fragment of the generated OWL ontology (in Polish) created in the Protege environment[5] is shown in figure 3.

We also developed the mechanisms for tagging text data from PDF files or web documents and generating test questions on the basis of the information extracted from these documents and the knowledge contained in the generated OWL ontology of the knowledge structures covering selected

3. See: SOVA by Piotr Kunowski and Tomasz Boiński, available at http://protegewiki.stanford.edu/wiki/SOVA.
4. See: RStudio Desktop software, available at https://posit.co/download/rstudio-desktop/.
5. See: Protégé Software, available at https://protege.stanford.edu/.

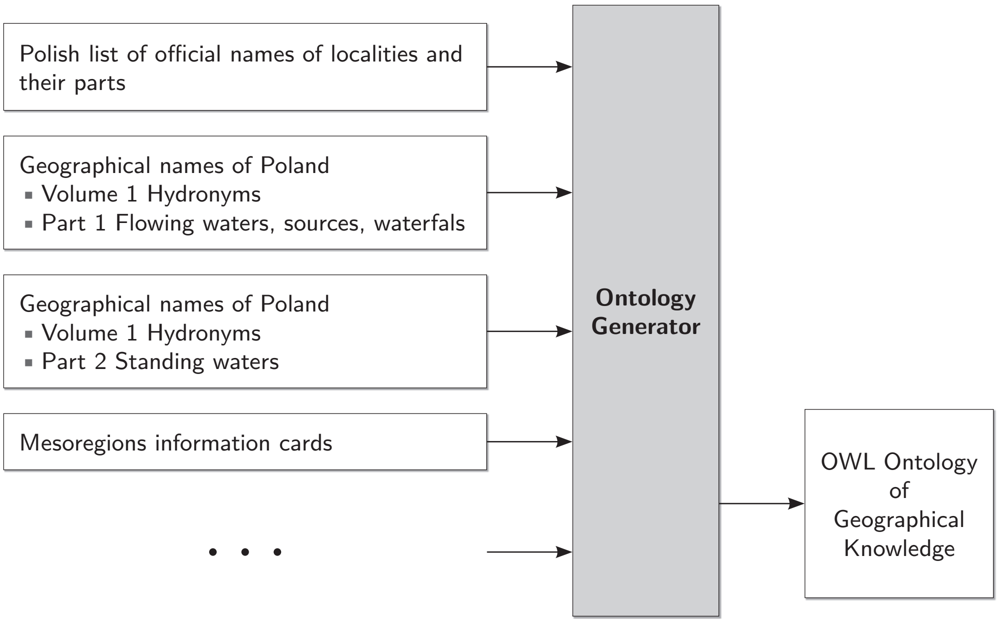

Figure 2. The ontology generator usage environment

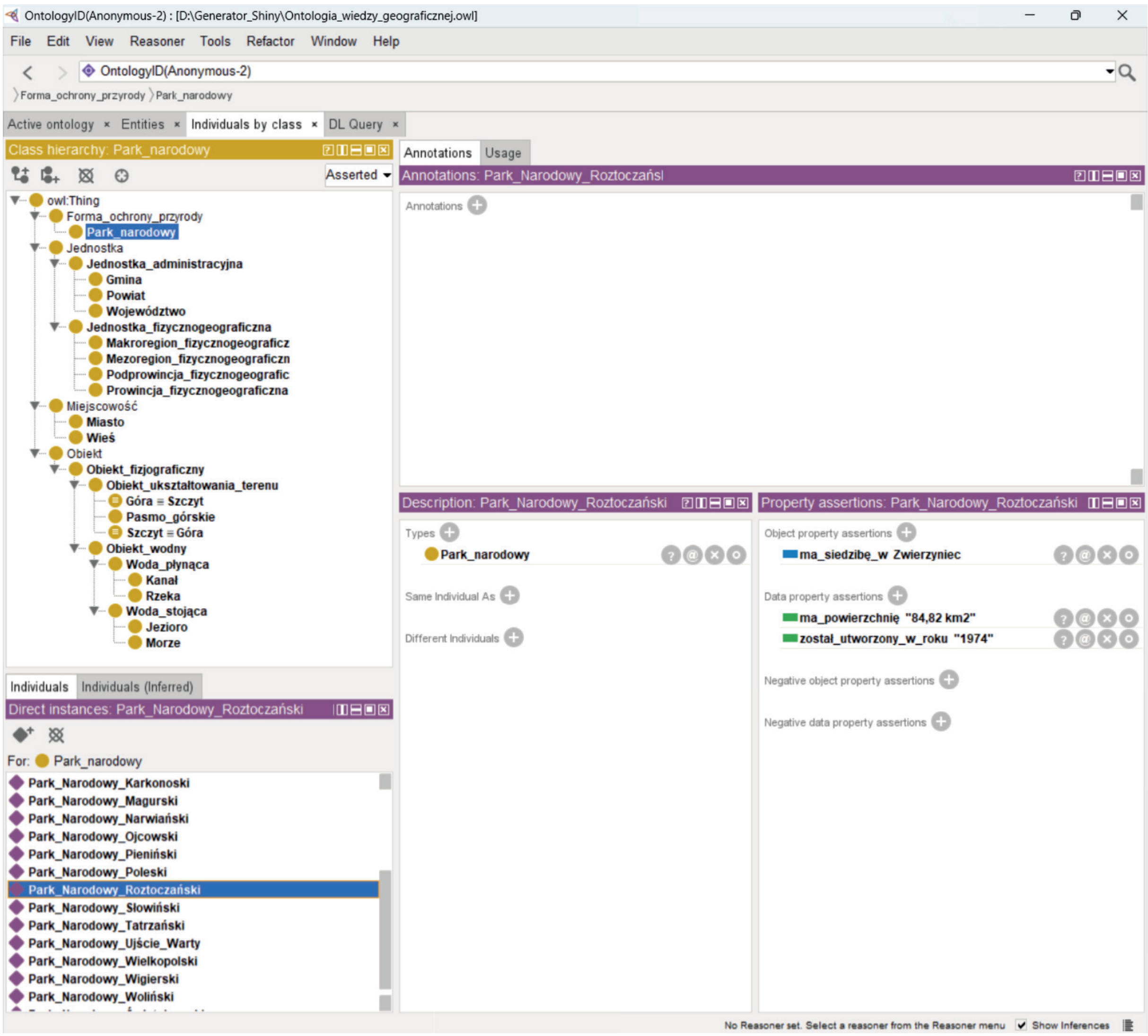

Figure 3. The fragment of the generated OWL ontology (in Polish) in the Protege environment

issues in the field of the geography of Poland. These mechanisms use, in particular, natural language processing methods — for example:
- punctuation removal
- numerical values removal
- removal of words with little meaning (words from the so-called stop list)
- lemmatization of words
- tagging
- determining the frequency of the occurrence of words

The mechanisms developed for tagging text data and generating test questions were implemented in the R environment. In particular, two external services provided in the REST API by the clarin-pl project[6] were used:
- determination of geolocation data, and
- lemmatization of words.

The main idea of the generation of questions and possible answers (correct and incorrect) is as follows. Key concepts are extracted on the basis of the frequency of their occurrence. In particular, proper names are identified. Proper names are searched in the set of individuals in the OWL ontology. Next, according to the OWL ontology, object and data properties linking individuals together or linking individuals with data values are identified respectively. Object and data properties determine the structures of the questions generated by the tool. The information about individuals linked together or individuals linked with data values enables us to determine incorrect answers (if the linking is not present).

For example, if the Wisła river is identified in the text as a key concept, then the ontology delivers the information that the river is connected with other individuals by means of the following object properties: *flows into* (*wpływa do* in Polish), *flows through* (*przepływa przez* in Polish). Therefore, in the knowledge test, we can ask where the Wisła flows, through which cities the Wisła flows, etc. The general scheme of the environment for the knowledge test generator is shown in figure 4.

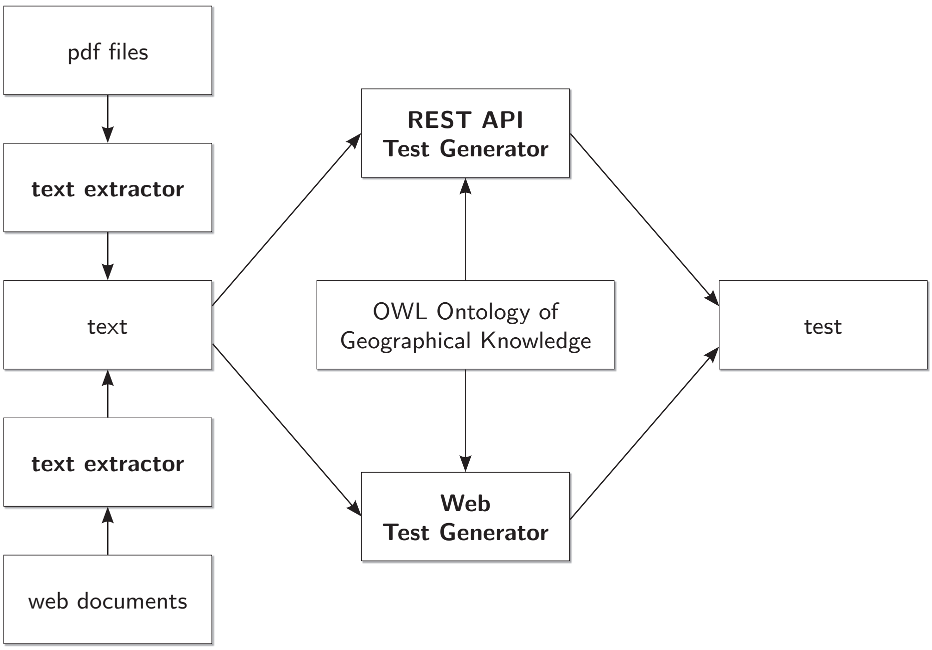

Figure 4. The knowledge test generator usage environment

6. See: https://ws.clarin-pl.eu/.

3 Generator Implementation

The prototype of the ontological generator of knowledge tests from texts was implemented in two versions:

- As a web application using the Shiny package[7] for the R environment. Questions are generated on the basis of two web documents for which the user provides URLs. The user can also set the number of correct answers and the number of incorrect answers. The GUI (in Polish) is shown in figure 5.

Figure 5. The GUI of the web ontological generator of tests from texts (fragment)

- As a REST API server using the plumber package[8] for the R environment. The parameters for the generator of tests (Web document URLs, the number of correct answers and the number of incorrect answers) are sent in the HTTP POST request body in the JSON format, as shown in the example below:

7. See: https://shiny.posit.co/.
8. See: https://www.rplumber.io/.

```
{"urls": ["https://pl.wikipedia.org/wiki/San",
"https://pl.wikipedia.org/wiki/Przemyśl"],
"correct": 2, "incorrect": 4}
```


The answer to the REST API is a set of questions in the XML format, as shown in the example below (questions and answers are in Polish):


```
<?xml version="1.0" encoding="UTF-8"?>
<test>
 <question id="1" txt="Dopływem rzeki San jest:">
 <answer>Jaskrzanka</answer>
 <answer>Płonia</answer>
 <answer>Węgierka</answer>
 <answer>Wołczyński Strumień</answer>
 <answer correct="true">Sanoczek</answer>
 <answer correct="true">Łukawica</answer>
 </question>
 <question id="2" txt="Rzeka San przepływa przez:">
 <answer>Kamienna Góra</answer>
 <answer>Krasnobród</answer>
 <answer correct="true">Dynów</answer>
 <answer>Skwierzyna</answer>
 <answer correct="true">Przemyśl</answer>
 <answer>Gryfino</answer>
 </question>
 <question id="3" txt="Miasto Przemyśl leży nad rzeką:">
 <answer>Bug</answer>
 <answer>Łyna</answer>
 <answer correct="true">San</answer>
 <answer>Wisła</answer>
 <answer>Odra</answer>
 </question>
 <question id="4" txt="Miasto Przemyśl leży w:">
 <answer>Województwo lubelskie</answer>
 <answer>Województwo śląskie</answer>
 <answer correct="true">Województwo podkarpackie</answer>
 <answer>Województwo lubuskie</answer>
 <answer>Województwo opolskie</answer>
 </question>
</test>
```


Conclusions

In the paper, we have described the created OWL ontology of the knowledge structures covering selected issues in the field of the geography of Poland. Thanks to the use of the universal OWL/XML standard, the presented ontology can be used in different applications. In addition, this ontology can be easily developed and modified.

Moreover, we have presented:

- the mechanisms of data acquisition and ontology generation that have been developed, and
- the mechanisms of generating tests from texts, tagging text data from web documents and generating test questions based on the information extracted from these documents and the knowledge contained in the generated OWL ontology.

These mechanisms were implemented in the prototypical ontological generator of tests from texts. The developed mechanisms can be easily expanded, thanks to which it is possible to create advanced test generators from texts. The implemented applications of the prototypical ontological generator of tests from texts (in the form of a web application and in the form of a REST API service server) can be easily extended in the future for other knowledge fields.

References

Bowman, S.R. 2023. Eight Things to Know about Large Language Models. *arXiv*. doi:10.48550/arXiv.2304.00612.

Cheng, T. 2021. "Supervised, Semi-Supervised, Unsupervised, and Self-Supervised Learning." Towards Data Science. https://towardsdatascience.com/supervised-semi-supervised-unsupervised-and-self-supervised-learning-7fa79aa9247c.

Gómez-Pérez, A., M. Fernández-López, and O. Corcho. 2004. *Ontological Engineering. With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. London: Springer-Verlag.

GPT-4 Technical Report. 2023. doi:10.48550/arXiv.2303.08774.

Guarino, N., and P. Giaretta. 1995. "Ontologies and Knowledge Bases. Towards a Terminological Clarification." In *Towards Very Large Knowledge Bases*, edited by N.J.I. Mars, 25–32. Amsterdam: IOS Press.

Horridge, M. 2011. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition 1.3. Manchester: The University of Manchester.

Neches, R., R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W.R. Swartout. 1991. "Enabling Technology for Knowledge Sharing." *AI Magazine* 12 (3):36–56.

Pancerz, K., P. Grochowalski, and A. Derkacz. 2016. "W stronę ontologii miejscowości w Polsce: przykład województwa mazowieckiego." *Barometr Regionalny. Analizy i Prognozy* 14 (3):137–144. doi: 10.56583/br.500.

Solon, J., J. Borzyszkowski, M. Bidłasik, A. Richling, K. Badora, J. Balon, T. Brzezińska-Wójcik, Ł. Chabudziński, R. Dobrowolski, I. Grzegorczyk, M. Jodłowski, M. Kistowski, R. Kot, P. Krąż, J. Lechnio, A. Macias, A. Majchrowska, E. Malinowska, P. Migoń, U. Myga-Piątek, J. Nita, E. Papińska, J. Rodzik, M. Strzyż, S. Terpiłowski, and W. Ziaja. 2018. "Physico-Geographical Mesoregions of Poland: Verification and Adjustment of Boundaries on the Basis of Contemporary Spatial Data." *Geographia Polonica* 91 (2):143–170. doi: 10.7163/GPol.0115.

Wykaz identyfikatorów i nazw jednostek podziału terytorialnego kraju. Podział terytorialny z dnia 1.01.2015 r. 2015. Zeszyty Metodyczne i Klasyfikacje. Warszawa: Główny Urząd Statystyczny.