applied sciences

MDPI

Article

Digital Transformation of the Etymological Dictionary of Geographical Names

Tomasz Kubik

Faculty of Electronics, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; tomasz.kubik@pwr.edu.pl

Abstract: This article aims to contribute to the methodology of the structuring of etymological dictionaries of geographical names and the popularization of knowledge regarding the origin of Silesian toponyms. It is based on experiences gathered during the digitization and publication in an electronic form of the SENGŚ ("Etymological Dictionary of Geographical Names of Silesia") and addresses the problems encountered. The article discusses the rules applied in the compilation of the SENGŚ and presents two information models used during the digitalization of this dictionary: a relational model and a graph model. The first one corresponds to standard approaches when designing electronic versions of dictionaries. The second allows the creation of solutions conforming to the idea of Linked Open Data, which are deployable as parts of the Semantic Internet. An important aspect also considered was the linking of historical materials listed in the dictionary entries with the corresponding records maintained in digital repositories. This association was realized using the AZON platform ("Atlas of Open Scientific Resources").

Keywords: etymological dictionary digitization; information model; geographical names; toponomastics

Citation: Tomasz, K. Digital Transformation of the Etymological Dictionary of Geographical Names. *Appl. Sci.* **2021**, *11*, 289. https://doi.org/10.3390/app11010289

Received: 7 December 2020
Accepted: 24 December 2020
Published: 30 December 2020

1. Introduction

One of the major challenges accompanying the development of human civilization is knowledge management. Dictionaries, as collections of words of a language compiled together with informative descriptions, play an important role in this. Dictionaries can be divided into two groups: (i) linguistic dictionaries, focused on lexical units and their linguistic properties; and (ii) encyclopedic dictionaries, oriented towards the extra-linguistic world, which offer terms that have a denotative character [1]. There exists a relationship between a dictionary and vocabulary concepts. Simply put, a dictionary has a physical existence that preserves a selected part of vocabulary—an intangible stock of words in a language known by a person or a community. Vocabularies are sometimes recognized as alphabetized collections of words, formally defined or explained. In that sense, vocabulary is a synonym for dictionary. This is especially true for controlled vocabulary—"an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching" [2]. Controlled vocabularies can be materialized in the form of books or information systems (called e-vocabularies or e-dictionaries). Apart from accumulating essential knowledge, they can also perform additional functions. They serve as reference material, which is perfectly suited for classifying, supplementing, and describing a variety of resources and performing other tasks [3]. Quite often, they play the role of thesauri—collections of words arranged in conceptual groups or alphabetically.

Dictionaries (or controlled vocabularies) may contain a great deal of useful information; however, their perception sometimes becomes difficult. Everything depends on the assumptions made concerning editing, organizing, and presenting dictionary entries. In the case of traditional, printed dictionaries, these matters are ruled by appropriate editorial guidelines. The guidelines specify not only the expected format of the particular

paragraphs, but also mechanisms to condense, refer to, index, and underline their content. They also put some sanctions on the overall dictionary structure.

The most common method of collecting knowledge and compiling dictionaries was cataloguing data records in form of files, notes, alphabetical card index, etc. These records were then converted into dictionary entries through a series of iterations. Once edited, the final material was difficult to change. With advances in information technology, the process of editing dictionaries has evolved. Although the rules regarding the merits were retained, proper information modelling attracted more attention. Mainly, because the right models ensure flexibility of solutions built, and facilitate efficient data processing. Additionally, the separation of data storage (databases) from the place of publication (interfaces of web-based applications) elevates routines to the next level. Nowadays reshaping e-dictionaries is a matter of applying proper styling to the gathered data rather than the re-edition of directly printable data sources. The cross-sectioning of data and visualization of obtained results on customized views are also possible. Thus an e-dictionary can offer both lemma oriented and concept oriented front ends. Additionally, access to information can be opened to a wider group of interested people and also to machines.

Etymological dictionaries of geographical names share characteristics of both aforementioned groups of dictionaries: encyclopedic and linguistic. They offer "proper nouns applied to natural, man-made, or cultural features on Earth" [4] along with an etymological description. Compilation of their entries requires specialist knowledge, often from distinct domains, and is hard to automate. This can be explained by the fact that the discovery of name origins must comprise linguistic and extra-linguistic factors, where the latter often outweighs the former. Deep understanding of geographical, cultural, and historical contexts is also essential.

Geographical context makes these dictionaries similar to gazetteers. Gazetteers collect structured data about spatially referenced terms and names [5]. They are implemented as services of spatial data infrastructures (SDIs), providing geographical coordinates based on object names and often offering details on physical properties (such as area, dimension, and geometry), together with statistics (population, income, employment level, etc.) and relations to other physical objects. However, in their case, finding historical data beyond one hundred years prior is rarely possible. This option is available only in domain-specific solutions. Etymological dictionaries of geographical names published on an open license can be considered as candidates for linkage with the Linguistic Linked Open Data (LLOD) cloud (http://linguistic-lod.org/) [6]. The criteria for when a source can be regarded as forming part of this cloud refine the principles coined by Tim Berners-Lee [7] dictating: the use of URIs as data identifiers (which enables the retrieval of machine-readable descriptions using the HTTP protocol), application of web standards, such as HTML and RDF or JSON-LD, for data serialization, and the offering of additional links to related resources. They enforce demands on how e-dictionaries should be modeled and deployed (with a consequent need to implement appropriate network infrastructure).

All the issues mentioned so far have emerged in the course of the digitization and publication of *Słownik Etymologiczny Nazw Geograficznych Śląska* (SENGŚ: "Etymological Dictionary of Geographical Names of Silesia"). This multi-volume dictionary was originally compiled to gather specific knowledge about the origins of Silesian geographical names and their evolution, as confirmed by historical sources, taking into account the coexistence of two or more languages in one area.

The dictionary is addressed to institutions and persons interested in the local names of Silesia—primarily linguists, historians and geographers, ethnographers, regionalists, and other lovers of the region. However, the list of its users is not limited in any case. It can be also attractive, for example, to fans of prose by Andrzej Sapkowski, author of *Wiedźmin* ("The Witcher"), tracing the fate of the characters of his Hussite Trilogy: *Narrenturm* ("The Tower of Fools"), *Boży bojownicy* ("Warriors of God"), and *Lux perpetua* ("Ceaseless Light") set in the Lands of the Bohemian Crown.

The digitization of SENGŚ is intended to increase access and allow further extension of domain-specific knowledge. Scans of the fifth to the seventeenth volume have been deposited and made available under an open license in *Atlas Zasobów Otwartej Nauki* (AZON: "Atlas of Open Scientific Resources")—a platform that stores tens of thousands of scientific resources: books, articles, magazines, teaching materials, presentations, photos, 3D scans, audio and video files, databases, and many more (see https://zasobynauki.pl). A prototype of the SENGŚ as a web application can be accessed at http://sengs.e-science.pl (see Figure 1). In the following section, the scope of the dictionary and its historical context are presented. Next, some remarks are given on how the subsequent volumes were prepared, with attention paid to the method originally used for collecting data. The initially considered area of Silesia is illustrated on the map. After discussing the structure of dictionary entries, the article focuses on proposals of information models for an electronic version of the dictionary. The designed entity-relationship and graph models are presented on the relevant diagrams along with explanations of their details. The article ends with a discussion.

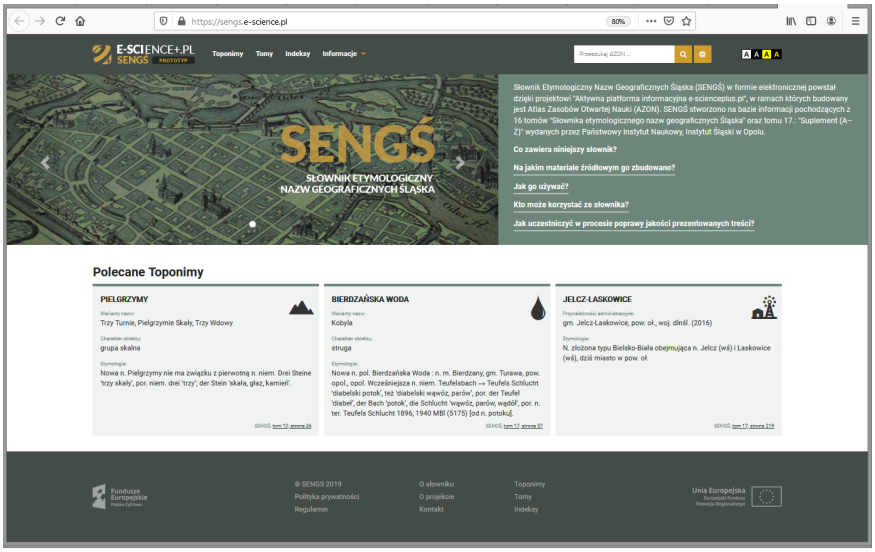


Figure 1. View of the SENGŚ portal prototype.

2. The Etymological Dictionary of the Geographical Names of Silesia

2.1. Scope of the Dictionary

The scope of SENGŚ goes beyond that set for traditional etymological dictionaries of geographic names. As a multi-volume publication compiled over nearly 70 years, SENGŚ integrates results of the collective effort of several generations of linguists. All identified and reconstructed geographical names of Silesia were assigned with etymology, characteristics, locations, and references to the historical sources documenting their continuity. The list of historical and contemporary names includes, but is not limited to, local (inhabited settlements), field (fields, meadows, forests, roads, pastures), water (rivers, streams, lakes, ponds), and mountain (peaks, hills, slopes, rocks, gorges, caves) names. Due to the richness of medieval, modern, and new sources, as well as long-lasting bohemization, germanization, and polonization processes, compilation of the dictionary was a big scientific challenge. The following short historical introduction will highlight the matter [8].

Over the centuries, the political and state configurations of Silesia, as well as its ethnic and denominational conditions, have changed several times. During the turbulent period of the first West Slavic countries, the Silesian territory belonged to the area of Great Moravia, and then of its heir, the Duchy of Bohemia. At the end of the 10th century, Silesia was taken over by the first historical ruler of Poland, Mieszko I. In the 14th century, the situation reverted and Silesia became a part of the Bohemian Kingdom. In the third decade of 16th century, it fell under the rule of the Habsburg monarchy of Austria. The first half of the

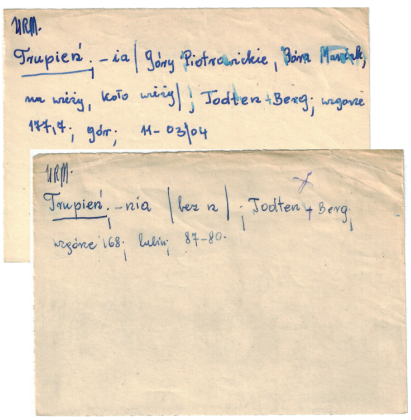
18th century yielded further changes. The area of Silesia was captured by the Kingdom of Prussia. Next, in the second half of the 19th century, it became a part of the German Empire. The end of the First World War resulted in the division of Silesia between Germany, Poland, and Czechoslovakia. After the Second World War, the German part of Silesia was merged with Poland.

Silesia, as a region with a rich historical past, is a very attractive research subject. Its multicultural and multi-ethnic roots generate many research topics in different fields of formal, natural, social, and even applied sciences. SENGŚ gives the readers a strong basis for independent studies on the nature of names, the causes of changes, and other naming issues, and serves as reference material for linguistic research and conducting scientific work in related disciplines. The East Slavic proper names attracted the attention of Slavists from various countries, making room for sharing experiences. The development of a common onomastic method in Poland and Germany confirmed this. "The unifying factor, ready for cooperation and agreement, was (and is) the magazine 'Onomastica Slavogermanica' that was established in 1964 in Wrocław, together with a university center in Leipzig" [9].

During the compilation of SENGŚ, more than 1400 relevant sources were used and thousands of toponyms were identified. Some of the sources have already been digitized, and many others are ready to be explored [8]. This can be illustrated in the example of the *Historische Ortsverzeichnis von Sachsen* dictionary, which became a data source for a web service implementation ("Digital Historical Gazetteer of Saxony", https://hov.isgv.de/) or *Słownik geograficzny Królestwa Polskiego i innych krajów słowiańskich* ("The Geographical Dictionary of the Polish Kingdom and Other Slavic Countries"), whose searchable scans can be accessed on the Internet (http://dir.icm.edu.pl/Slownik_geograficzny/). A similar historical flavor characterizes the Meyers Gazetteer (https://www.meyersgaz.org) and *Elektroniczny słownik hydronimów Polski* (ESHP: "Electronic Dictionary of Hydronyms in Poland") (https://eshp.ijp.pan.pl/).

2.2. Dictionary Evolution

The first volume of SENGŚ, including names starting with A and B, was published in paper form in 1970. The last, seventeenth volume was issued in 2016. It was a supplementary volume, containing names starting with A to Ż. However, the initial work on the dictionary dates back to around 1957. Preliminary studies were done between 1957 and 1970 on the identification of historical and contemporary materials and compilation of the alphabetical card index (see Figure 2).



TRUPIEŃ, -pnia (Sargberg), też *Góra Śmierci*, g. 481 m, Pog. Złotoryjskie, Sudety Zach. (pomiędzy Wilkowem w pow. jaw. a Kondratowem w pow. złotor., dlnśl.): *Trupień* (zamiast niem. *Sarg.-B.*) 1949 M.P. 44; *Trupień* (niem. *Sargberg*) R s.v.; *Trupień*. też *Góra Śmierci*, niem. *Sarg Berg // Sargberg* SGTS 7, 592; **2.** ~ (Todten Berg), wzgórze 178 m w d. pow. gór.: URM 11-03/04; **3.** ~ (Todten Berg), wzgórze 168 m w d. pow. lubin.: URM 87-80.

Ad_1 Pierw. n. niem. *Sargberg* 'trumienna góra', por. *der Sarg* 'trumna', *der Berg* 'góra'. Pol. n. *Trupień* : *trupień* 'trup' nawiązuje do n. niem. Ad_{2-3} Niem. n. *Todten Berg* 'góra zmarłych', por. *tot* 'zmarły, nieżywy'.

Figure 2. Cards about *Trupień* from the manually edited alphabetical card index (**left**) and the corresponding dictionary entry, as published in the sixteenth volume of SENGŚ (**right**).

The research undertaken was based on the assumption that the entire dictionary will cover the area of Silesia within the historical borders determined by Semkowicz, W. (1933) [10] and Arnold, S. (1927) [11], approximately 60,000 km^2 (see Figure 3). Thus, in the

first volume, the full geography of the region was considered, i.e., place names, hydronyms, oronyms, microtoponyms, and even ethnonyms, which were also originally choronyms. The edited entries were intended to hold not only the Polish names, but also German, Czech, Lusatian (Upper and Lower), and even prehistoric (Celtic, Illyrian) names, along with references to historical documentation and etymological descriptions. It was decided that each entry should begin with the Polish name (with its variants in German) or with the German name (with its variants in Polish), followed by a comprehensive description. Adjustable editorial rules have been adopted to ensure the compactness of the resulting text. Although not all assumptions worked well, this was a good trial experiment, and the experience gained helped in the compilation of the following volumes.

Figure 3. The area of Silesia considered in the first volume of SENGŚ (in EPSG:2180).

Editorial rules were verified in the second volume, and the territorial range was reduced to the areas within Polish borders, mainly due to the publication of the *Místní jména na Moravě a ve Slezsku* ("Local names in Moravia and Silesia") dictionary [12]. The general principles were clarified and broader presentation of historical documentation and etymological descriptions were suggested. Further improvements to the dictionary continued over time. A growing range of source materials and methodological enhancements were also noticed. The first four volumes did not deal with the etymology of names that were official administrative "baptisms" (proposed and assigned to places by an administrative body); neither were German names explained without the existence of their Polish counterparts. This changed in the fifth volume, where the descriptions of all modern, reconstructed, colloquial, and administrative "baptisms" were provided. The dictionary's scope has been extended beyond the strict geographical domains, and names assigned to the objects resulting from human activities, particularly for "mines", were considered.

The sixth volume focused on explaining all Polish names, as well as Slavic and Pre-Slavic ones (assigned mainly to rivers), and all Silesian German names. Furthermore, special care has been taken to list historical materials in chronological order, as this simplifies the analysis of phonetic and morphological changes. In the next volume, the seventh, the list of historical sources was supplemented. Basically, in each newly published volume, the list of sources has been expanded according to the growing scope of research conducted. However, the core choice was left unchanged: diplomatic codes, regests, dictionaries, maps, archival files, etc.

The influence of foreign structures (German, Czech, and Lusatian) on the lingual picture of the region became more and more evident. In the ninth volume, one could even see traces of the Praslavian lexical system in the reported river names. A characteristic feature of the tenth volume was its orientation toward historical and contemporary local names, the vast majority of which appeared in historical records until the 15th century. In the eleventh volume, attention was drawn to the old Slavic and German names. The phenomenon of the displacement of Polish names by the later German was pointed out in the twelfth volume. Interesting evidence of name changes during the Middle Ages was given in the thirteenth volume. By the fourteenth volume, some descriptions of water, mountain, and local names were elaborated thanks to an expanded range of sources. Particularly valuable river names belonging to the pre-Slavonic and Indo-European lexical systems were embraced in the fifteenth volume. The sixteenth volume closes the alphabetical range of the vocabulary. Its content was improved thanks to benevolent pieces of advice given by the reviewers. However, the need to complete the list of entries and correct the existing ones was recognized. This was met by the supplementary, seventeenth volume.

2.3. Details of Entries

The descriptions embedded inside dictionary entries are very compact. Their authors applied some rules to make the running text as short as possible. As a side effect, some parts of the descriptions are difficult to read, especially those from the first volumes. This is mainly due to the use of abbreviations and transitive dependencies. Thus, before diving into vocabulary volumes, the reader should be aware of these rules.

The dictionary consists of entries arranged alphabetically according to their lexical labels. These labels can be treated as headwords or lemmas accompanied by several pieces of information. This construction complies with the general lexicographic principles given in [13] and the ideas of presentation/representation of entries described in [14]. In fact, there are two kinds of entries in SENGŚ (see Figure 4):

- Descriptive entries, with detailed descriptions of real objects;
- Referring entries, with labels of other entries listed for comparison.

KLINICZKI, -ek (Fahrhauser). os., gm. Bojadła, ziel.: *Fährhäuser* 1845 K 126; 1908 SOV 61; *Klenica,* niem. *Fährhäuser b. Kleinitz* Pasterniak 35; PRL 475; Wyk II, 84.

N. powojenna *Kliniczki: klin* zamiast niem. *Fährhäuser* = 'domy przy promie', por. *die Fähre* 'prom'.

AUGUSTÓW, -towa (Augustenhof), os., gm. Wądroże Wielkie, pow. jaw., dlnśl.: *Vorwerk Augustenhof* 1825 SGTS 19, 51; *Augustenhof* 1845 K 721; *Augustenhof* 1887 Glex 264; *Augustenhof* 1941 SOV 16; *Augustów* 1948 M.P. A-78; R s.v.; *Augustów* PRL; *Augustów* Wyk I 29 i 2013 Dz.U. poz. 200; **2.** ~ (Augu stthal), przys. wsi Giebułtów, gm. Mirsk, pow. lwów., dlnśl.: *Augustthoal* 1738 SGTS 2/1, 87, *Augustenthal* 1761 SGTS 2/1, 87; *Augustthal* 1887 Glex 298; *Augustthal* 1908 SOV 8, *Augustthal* 1941 SOV 16; *Augustów* 1948 M.P. A-78; R s.v.; *Augustów* PRL 25; *Augustów* Wyk I 29 i 2013 Dz.U. poz. 200.

Ad$_1$ Pierw. n. niem. *Augustenhof* 'dwór Augusta' (por. *der Hof* 'dwór, folwark', im. niem. *August*) przetłumaczona urzędowo na j. pol. jako *Augustów.* Ad$_2$ Nowa n. pol. *Augustów* tłumaczy n. niem. *Augustthal* 'dolina Augusta', por. *das Tal* 'dolina', im. niem. *August.*

BERGHOF cf. GÓRECZKI.

BIAŁKA cf. BIAŁA; BIAŁA GŁUCHOŁASKA; BIAŁA WISEŁKA; BIAŁA WODA; BIELA.

Figure 4. Exemplary entries with detailed descriptions (KLINICZKI—one object, AUGUSTÓW—two objects) and labels listed for comparison (BERGHOF—one label, BIAŁKA—five labels).

The content of the descriptive entries can be split into the following parts:

- Lexical label—holds the name of the object or objects being described. By default, the lexical label opens the description of the first object. If there are more objects,

their descriptions are given in enumerated sections (the second and the consecutive section are marked with numbers, and the use of a repetition character "~" at the beginning is allowable). These sections can be mapped to the *senseGroups*, and the section number to the *senseNumber* [14]. The label can be marked with a character denoting a certainty qualifier: "∗" if the reconstructed name differs from the common form, "†" if the name was considered to be lost, and "?" if the reconstruction of the name is doubtful or unclear. All the names are given in the primary, reconstructed, and current forms (at the time of dictionary publication).

- Genitive ending (or endings)—if given, it represents the genitive of the Polish name (the second case in Polish grammar).
- Name variant (or variants)—one or more words representing a name variant. By assumption, the words placed in brackets are official names, and others are treated as unofficial names. In most cases, if a name was given in Polish, an official variant was in the German language.
- Physical characteristics of the object—these define the nature of the object (e.g., city, river, village).
- Location of the object—informs about the administrative affiliation of the object and/or about names of the associated river basin, river inflow, or mountain peak, depending on the object's nature.
- Historical material—a chronologically ordered aggregation of: a transliterated form of the name (as mentioned in the historical source), year, and bibliographic reference to the historical source (usually an abbreviation of the name, followed by the volume, numbering, or position).
- Etymology—explains the origins of the name. This part appears at the end of the entry. Its paragraph without any annotations applies to all objects described, and annotation with numbers (or a range of numbers) applies to objects described in the relevant numbered sections.

The referring entries are compiled from two parts, which are separated by the abbreviation "cf." (a short form for the Latin confer ("compare")):

- Source lexical label—contains the name used as a comparison source, sometimes assisted with a complement (additional text);
- Destination lexical labels—represent names (at least one) serving as comparison destinations, pointing at other descriptive or referring entries, sometimes assisted with complements (additional texts).

3. Model Design

In the software engineering domain, an information model design is a process in which a conceptual description of modeled objects, relationships, restrictions, rules, etc. is created at the level of abstraction that ensures implementation neutrality. Elements of the information model may be declared in a formal way, e.g., by using class diagrams of Unified Modeling Language (UML), Entity-Relationship (ER) models, or Resource Description Framework (RDF) graph models. The use of natural language is also acceptable. The conceptual description can be implemented in different ways; thus, different data models can be created based on the same information model. However, sometimes, it is difficult to draw a clear line between the information model and the data model [15]. This happens in cases of complex problems when the transition from the idea to the implementation goes through many phases.

The design of the information model for a dictionary is a complex task, and depends on the dictionary's scope, structure, and offered features. For that reason, linguistic and encyclopedic dictionaries cannot be shaped in the same way. Moreover, the models might vary when it comes to ensuring proper presentation or to facilitating data processing (some issues related to the typography and information structures of different views were discussed in [16]).

Usually, controlled vocabularies in the form of taxonomies, glossaries, thesauri, or subject headings are modeled by utilizing inheritance and associations [2]. Their information models have no expressivity to declare other linguistic properties. This can be spotted in the most commonly used data models derived from the ISO 25964-1 standard [17] or the SKOS W3C Recommendation [18]. The information models of lexical databases, like WordNet [19] or Słowosieć [20], are different. This allows grouping of words into synsets, hypernyms, hyponyms, etc., and provides, among other things, attributes for preserving short definitions and usage examples. Some parts of the collected data can be automatically inferred from the language corpora, and some need human attention. In general, this model incorporates paradigmatic relationships among lexemes (so-called lexical relations), and its implementation may follow custom approaches or standardized ones.

Currently, proposals of models oriented at machine-readability and data linking exist. In general, ontologies (understood in the IT sense) give greater possibilities in this respect. Usually, they are designed with the use of RDF, RDF Schema (RDFS), and/or Web Ontology Language (OWL) constructs, are serialized in several forms (RDF/XML, OWL/XML, Turtle, N3), and are managed by a number of software tools (editors, repositories, inference engines). For example, WordNet 3.1 (accessible at http://wordnet-rdf.princeton.edu/) has an RDF export implemented based on, amongst others, *Lemon—The Lexicon Model for Ontologies* [21].

The information model of SENGŚ has been designed to reflect the full content of the printed dictionary, with some extensions. The aim was to create a foundation for the implementation of basic e-dictionary functions, such as: browsing (using indexes), searching (using a full-text search engine), and delivering dereferenceable identifiers (for published materials). The work started with an analysis of the editorial rules, which were interpreted as a description of the information model expressed in a natural language. However, because of the many exceptions detected inside the printed dictionary entries, some parts of the proposed model became very specific. Further modifications would be needed before applying this model to the more general case of an e-dictionary.

To get used to the dictionary, it is good to keep in mind that labels of entries of both mentioned types are not bound to any real objects. They are just headwords representing sets of homographs. In fact, they hold an abstract geographical name. The association with a real object must be inferred from the provided description (if such a description exists within the dictionary entry). It is also worth remembering that the etymological descriptions of the entries evolved. In the first volume, they followed a structural classification, distinguishing the appellative and toponymic derivatives [22]. In subsequent volumes, this structure was replaced by unstructured text, written in natural language, with some abbreviations. This was done to improve the readability of entries for common readers. Scientific considerations, such as those on compounding in Polish described in [23], were not in focus. However, remarks on the formal methodology used were given inside introductory chapters. The following excerpt of etymological descriptions of *Adamowice* from the first and the last volumes demonstrate the effects of this change (instead of numbers "II 3" reflecting a structural classification, a descriptive explanation was introduced: "Dawna n. patr. ..." (pol.)—"an old patronymic name ..."):

> *Adamowice* (vol. 1): II 3—*Adam* n. os.; cf. pol. n. patr. *Adamowice*, czes. 'ts'.
>
> *Adamowice* (vol. 17): Dawna n. patr. *Adamowice* od im. Adam z przyr. - *(ow)ice*.

Since the compilation of the full dictionary took several decades, during which significant changes occurred in the administrative division of Poland, some of the published data became outdated. Therefore, the possibility of updating them would be a valuable extension. Moreover, the final form of the model should be shaped with a view of improving search mechanisms and integration with other systems by linking data. These observations led to the following assumptions:

- Parts of the entries that do not determine real objects can be treated as abstract toponyms. Because of the existence of two kinds of entries, the abstract toponyms can

be divided into: explicated toponyms (which are abstract toponyms associated directly with detailed descriptions of real objects) and compared toponyms (which are abstract toponyms associated with the other abstract toponyms through the comparison list. The entry parts that are related to real objects (real places) can be treated as concrete toponyms. They are characterized by attributes such as: name, genitive, location, etymology, etc.

- There exists an association between concrete toponyms and explicated toponyms belonging to the same entry: the attributes shared among concrete toponyms, like part of the etymological description, are assigned to the corresponding explicated toponym.
- Each concrete toponym is affiliated with the proper administrative unit valid at the time of publication of the particular volume. However, adding the current administrative affiliation should also be possible.
- To facilitate advanced searching and linking, the following model extensions are required: classifying attributes for reasonable cross-sectioning, georeferenced location for positioning in the real world, external links to related Web resources, and references to historical materials managed in external systems.
- Ideally, all references appearing in the descriptive parts of the entries should be converted into hyperlinks. Therefore, the data types used to implement the relevant parts of the model should offer such a possibility.
- There should be some elements available to declare data certainty.
- The form of etymological description should be preserved, but its structuring should be considered. Based on these assumptions, two models were designed: an entity-relationship model and a graph model.

3.1. An Entity-Relationship Model

"The relational model is solidly based on two parts of mathematics: first-order predicate logic and the theory of relations" [24]. Relation is defined as a set of tuples (also known as records), all of the same type. The tuples of the same type have the same set of fields (also known as attributes), and each field can hold a value of a certain type. The operations performed over a relational model are handled by relational algebra. This algebra serves as a procedural query language, which takes instances of relations as input and yields instances of relations as outputs. In the context of relational databases, the concepts of the relational model's relation, attribute, and tuple are transposed into, respectively: tables, columns, and rows [25].

The entity-relationship model of SENGŚ reflects the complex data structure recognized within dictionary entries (abstracted from data types) and offers some required extensions (see Figure 5). The proposed tables and their columns have self-explanatory names; nevertheless, some of them require an explanation.

The central part of the model is the `concrete_toponyms` table with the columns:

- `concrete_toponym_id`, `abstract_toponym_id`—primary and foreign keys;
- `object_characteristics_id`—a foreign key to the `object_characteristics` table that stores qualifiers as: city, town, village, etc. (over 370 such terms were identified);
- `toponym_class_id`—a foreign key that refers to the primary key in the table with terms representing toponym classes, like oikonym, hydronym, urbanonym, etc. (an extension introduced to facilitate qualification);
- `attributes`—a place for the specific information about objects being described (e.g., river length, mountain height, lake surface area);
- `concrete_toponym_number`—a number assigned to a concrete toponym;
- `descriptive_location`—a location expressed in natural language, applicable especially to rivers (descriptions of river courses, basins, and inflows) and mountains (descriptions of mountain peaks and ranges);
- `etymology`—a descriptive explanation of the name's origin (potentially encoded with the use of markup language);

- `genitive`—a genitive ending (or endings);
- `georeferenced_location`—a location in the form of geometry represented as Well-Known Text (WKT) strings [26] (an extension introduced together with `georeferenced_location_source` to facilitate georeferencing);
- `georeferenced_location_source`—source of information stored in `georeferenced_location`;
- `language`—a language in which the name is given;
- `name`—a real object name;
- `original_description`—an original, textual description extracted from the dictionary (an extension introduced for data verification);
- `volume_number`, `page_start`, `page_end`—attributes used to represent the physical position of the description in the original dictionary (an extension introduced to preserve information about origins of data);
- `status`—a certainty qualifier.

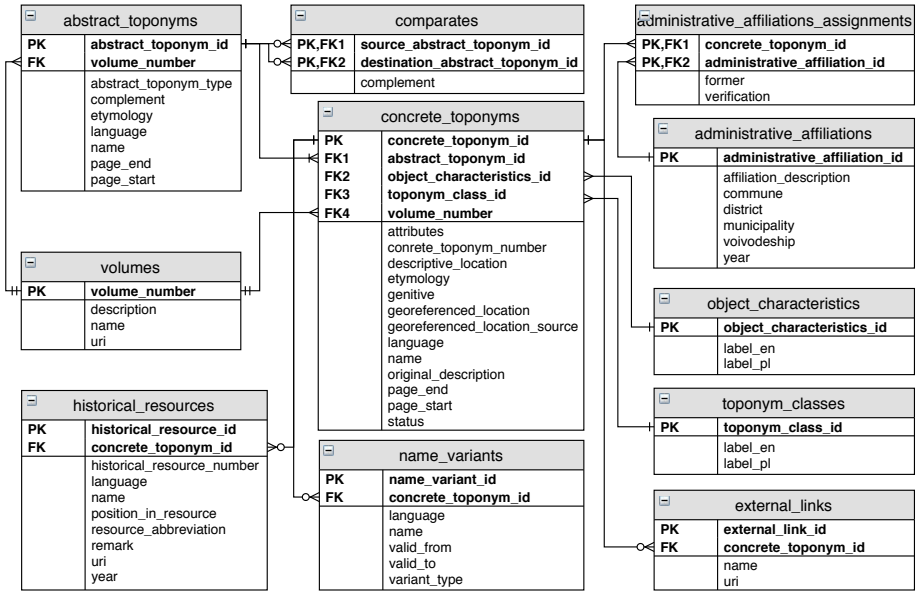


Figure 5. Entity-relationship model of SENGŚ.

Each concrete toponym may have several name variants; therefore, these were moved to the separate `name_variants` table. The `variant_type` attribute informs about whether the variant is official or unofficial. `valid_from` and `valid_to` represent the name validity period (if it is known). To enable automation and simplify machine processing, the textual data can be structured in various ways: from annotated text up to complex data structures. In SENGŚ, the `etymology` attribute is susceptible to structuring (as part of the planned extension), but only by annotation. This restriction gives the possibility of postponing implementation of such structuring without consequences. Moreover, it opens a chance to use markup language (such as the Lexical markup framework (LMF) described in the ISO/DIS 24613-3, standard which is still under development [27]). All abstract toponyms are collected in one `abstract_toponyms` table. An `abstract_toponym_type` column helps to differentiate explicated and compared toponyms, and `complement` column helps to store additional, complementary text. The meanings of the other columns are the same as in the `concrete_toponym` table. The source and destination of comparisons appearing in referring entries are collected in `comparates` table.

In the original dictionary, every concrete toponym has a location that, according to the editorial rules, can be an administrative affiliation and/or descriptive information, depending on the object's nature. Due to the fact that an administrative affiliation can be structured, this part of the information was extracted to the separate `administrative_affiliation` table associated with `concrete_toponyms` through the `administrative_affiliation_assign-`

ments table (which includes attributes introduced to increase data certainty: former—informing that the concrete toponym was known to have this administrative affiliation previously; verification—indicating verification status, which can be, for example: "reconstructed but not verified in the official documents on administrative division").

By assumption, the primary source of information on administrative affiliation was the original dictionary. There was no requirement to manage administrative division of the whole country. Thus, the administrative_affiliation table has the following columns:

- administrative_affiliation_id—a primary key;
- affiliation_description—an attribute merging all parts of administrative affiliation into one string (potentially encoded with the use of markup language) for display on the user interface;
- commune, district, municipality, and voivodeship—attributes used to represent names of administrative units at different levels of administrative division;
- year—an attribute indicating the year in which the administrative affiliation was valid (usually the year of the volume's publication).

Other parts of information related to concrete toponyms are kept in the associated tables: object_characteristics and toponym_classes (both facilitate classification in Polish and English), as well as external_links (introduced to facilitate data linking).

In the original dictionary, references to historical sources were interestingly handled. Their bibliographic metadata were introduced with abbreviations in the reference list. Then, the abbreviations, supplemented by the relative position of the quoted content, were placed inside the entries in the sections with historical materials related to the concrete toponyms. This approach was reflected in the historical_resources table, but with one significant extension. The bibliographic metadata of the historical sources and their electronic versions, subject to availability, have been deposited and published in AZON under unique URIs. This led to the design of the historical_resources table with the following columns:

- historical_resource_id, concrete_toponym_id—primary and foreign keys;
- name—a historical name of a concrete toponym, often in a form presented in a historical source;
- language—the language of the source;
- historical_resource_number—a number representing a position in a chronological list;
- position_in_resource—a relative position of the quoted content (like page, sheet number, etc.);
- resource_abbreviation—an abbreviation assigned to the historical source;
- uri—a link to the record of the historical source deposited and published in AZON (an extension);
- year—year of appearance.

The maintenance of bibliographic data and implementation of digital repositories are a challenge. Delegating these tasks to the external system solved many potential problems.

3.2. Graph Model

The graph model is much more flexible, expressive, and fine-grained than the relational model. It is not limited by the rigid schemas or preliminary assumptions about modeled domains. It can be augmented at any time according to needs. In addition, the supported mechanisms of data exploration are better suited to the use cases characterized by the existence of numerous associations than mechanisms based on the relational model.

The graph model can be built in several ways. One can use RDF [28], which defines a directed labeled graph consisting of triples: subject (entity), predicate (relationship or property), and object (entity or value). The nodes in this graph can be URI references, b-nodes (that have no external URIs) or literals. Arcs are URI references. The model can be serialized in various formats, like RDF/XML, N-triples, or Turtle.

The RDF vocabulary is limited to the terms for describing resources. However, the extension allowing the declaration of taxonomies of classes or properties and restrictions is possible. This is the purpose of RDF Schema (RDFS) [29]. RDF and RDFS together provide basic elements for the description of simple ontologies. More complex ontologies can be declared with the use of the OWL/OWL2 vocabularies [30].

At the beginning of the model's design, a dilemma arose—whether to use existing ontologies, such as, for example, the Conseil International des Musees Conceptual Reference Model (CIDOC CRM) [31], or to design a new one. Due to pragmatic reasons, the second option was chosen and the following explanation should clarify it.

Reasoning in ontologies relies on deriving facts that are not expressed explicitly under the entailment regimes of vocabularies used. This can include, among other things, instance retrieval and taxonomic reasoning or model consistency checks. The RDFS entailment recognizing some set of datatype IRIs defines applicable rules and also tells which queries and graphs are well formed. There are thirteen patterns of RDFS entailment to derive new statements from known ones [32]. In particular, the use of `rdfs:domain` and `rdfs:range` has strong consequences on semantical reasoning (see the *rdfs2* and *rdfs3* rules). These two constructs define the way the subject and object, linked by a property under consideration, are interpreted. It may happen that, for some instances, the domain or range applies to more than one thing. Then, the subject or object became the intersection of all of the types specified by, respectively, `rdfs:domain` and `rdfs:range`, not the union. To avoid unintended inferences, in some of the modeling approaches, these monomorphic properties are omitted in favor of other polymorphic constructs. For example, schema.org introduced `schema:domainIncludes` and `schema:rangeIncludes` properties in their model (https://schema.org/docs/datamodel.html). These two allow declarations of which properties are applicable to which class. This idea has been followed, for example, in the design of the euBusinessGraph Ontology (http://data.businessgraph.io/ontology) or the Sensor, Observation, Sample, and Actuator (SOSA) Ontology (http://www.w3.org/ns/sosa/). However, `schema:domainIncludes` and `schema:rangeIncludes` are not "semantical". Their occurrences are usually mapped to `rdfs:domain` and `rdfs:range` pointing at an instance defined with the use of the `owl:unionOf` property. This property, defined in OWL, determines the collection of classes or data ranges constituting a union and can be represented in RDF as a list [33].

The SENGŚ graph model was built on the base of the RDF model, taking into account a set of constructs supported by the RDFS++ reasoning. The set of constructs used includes the full list of RDFS and part of the OWL constructs, particularly: `rdf:type`, `rdfs:domain`, `rdfs:range`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `owl:inverseOf`, `owl:sameAs`, `owl:SymmetricProperty`, and `owl:TransitiveProperty` (as well as `owl:hasValue`, `owl:someValuesFrom`, and `owl:allValuesFrom`). This selection was made because of the features offered by the graph database targeted for the implementation of SENGŚ and because of the popularity of such solutions in the semantic web. Additionally, some annotation properties were chosen to document the ontology and the `owl:unionOf` property to declare domains of some of the designed properties. None of them are supported by RDFS++ reasoning (there was no intention to use them for that purpose in SENGŚ).

The resulting SENGŚ ontology (version 2.07) contains (excluding parts of ontology metadata): 17 classes (5 base classes and 12 subclasses), 46 properties (19 object properties, including 4 subproperties and 6 inverse properties, and 27 data properties), and several entities used as qualifiers. These statistics cover only elements directly involved in the representation of etymological vocabulary content. The $\mathcal{DL}$ (description logics) expressivity of this ontology is $\mathcal{ALUHI}(\mathcal{D})$, where: $\mathcal{AL}$ means attributive language (base language that allows: atomic negation, concept intersection, universal restrictions, and limited existential quantification); $\mathcal{U}$—concept union (allowed unions); $\mathcal{H}$—role hierarchy (allowed subproperties); $\mathcal{I}$—inverse properties (allowed properties that are declared as inverse to the other properties); $(\mathcal{D})$—use of datatype properties, data values, or data types (allowed declaration of custom datatypes or properties with values being objects of literals). The core

classes and properties of the SENGŚ ontology are presented in Figure 6. Documenting annotations and datatypes are not shown in order to simplify the view. For the same reason, instead of depicting the uses of `owl:unionOf` constructs in the declaration of domains of properties, these properties were repeated and linked directly to the appropriate type.

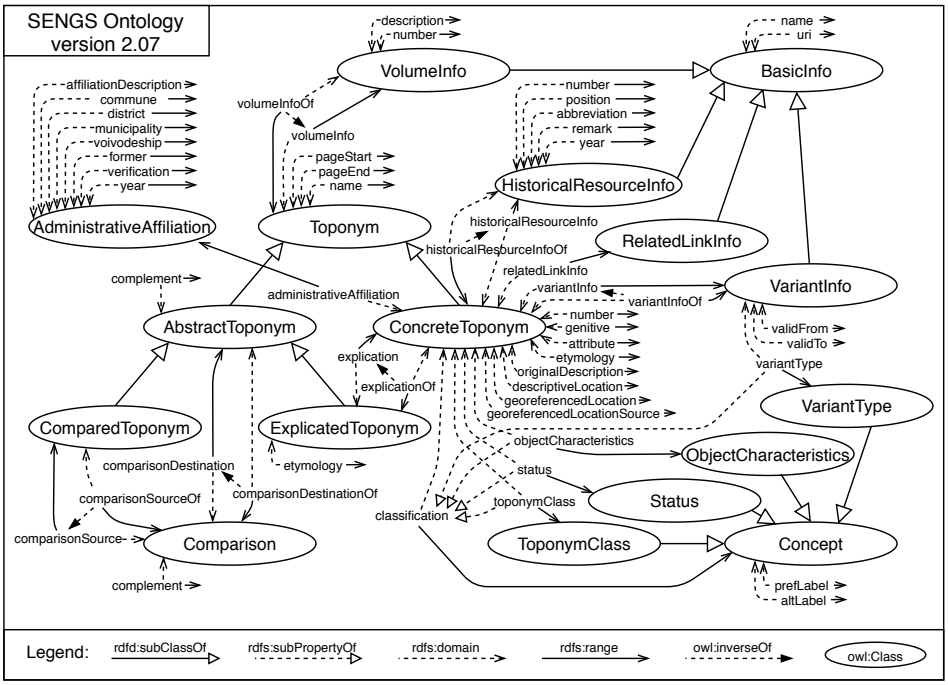


Figure 6. Diagram representing the core classes and properties of the SENGŚ ontology.

In most cases, the names of classes and properties are self-explanatory and correspond to the relational model. However, the modeling paradigm differs.

The use of class inheritance in ontology design brings certain benefits. Thanks to the inference mechanisms, instances of child classes will automatically be included in the extensions of the base classes (which follows from the entailment patterns *rdfs9* and *rdfs11*). Therefore, according to the proposed model, all instances belonging to the `ComparativeToponym` and `ExpandedToponym` classes will appear in the results of inquiries about instances of the `AbstractToponym` class. The `Toponym` class is a base class of the `AbstractToponym` and `ConcreteToponym` classes. The declaration of inverse properties is also important. Because of this, the two-way transition between graph nodes can be modeled without the need for declaring both directions explicitly. The property inheritance (*rdfs5*, *rdfs6*, *rdfs7*) opens up other possibilities. It is especially helpful in qualifier modeling. In common understanding, a qualifier is a word that limits or enhances another word's meaning. In the semantic web domain, qualifiers are used to partition the set of instances of a given class into subsets to improve search accuracy or assure better understanding. The principles of qualification can be defined in various ways. For example, the Dublin Core Metadata Initiative (DCMI) recognizes two broad classes of qualifiers: element refinement (implemented by the use of the mentioned `rdfs:subPropertyOf`, which has intentional semantics) and encoding scheme (implemented by the use of typed literals). The first was used, for example, in the DCMI Metadata Term Recommendation [34] in the declaration of the `dcterms:isPartOf` property. This property is a subproperty of `dcelements:relation` and `dcterms:relation`, and by *rdfs7*, every pair of resources related by `dcterms:isPartOf` is also related by `dcterms:relation`. The second class can be applied when the interpretation of an element value becomes crucial. The encoding scheme merely qualifies a value as a token or typed. The DCMI also specifies a Dumb-Down Principle: "A client should be able to ignore any qualifier and use the information as if it were unqualified" [35].

Due to the presence of qualifiers in the SENGŚ dictionary, it was decided to define custom classes and properties to model them. In effect, the `Concept` class has been declared (with four subclasses: `ToponymClass`, `Status`, `ObjectCharacteristics`, and `VariantType`) in addition to the `classify` property (with four subproperties: `toponymClass`, `status`, `objectCharacteristics`, and `variantType`). Instances of the mentioned `Concept` class and its subclasses form a local, controlled vocabulary. Used in "object" position of the `classify` property, or any of its subproperties, they qualify the "subject" according to the meaning of the underlying term. All qualifiers identified in the original dictionary and required by the extension have been modeled and included in this local vocabulary. Thanks to the rdfs entailment, all vocabulary members used as qualifiers can be retrieved by querying the graph for the "objects" of the `classify` property. Such qualifiers, once used in the "object" position of statements with the `classify` property or its subproperty as a "predicate", become members of the proposed classification schema (as a logical consequence of `rdfs:domain` declaration). Of course, there are other methods of modeling enumerated categories—for example, by the use of OWL constructs of extensional semantics—but these are not supported in RDFS++ reasoning and were not considered.

The SENGŚ model allows application of terms from external dictionaries as qualifiers. However, this should be done with caution. For example, the Library of Congress offers in its Linked Data Service (http://id.loc.gov) a set of 270 relator terms representing the relationship between a name and a bibliographic resource. Every term has its own URI and is declared in compliance with the SKOS specification [18]. By definition, at the same time, it is an instance of a `skos:Concept` and `owl:objectProperty` and is a `rdfs:subPropertyOf relators:role` and `dcelements:contributor`. In general, relators should be used as "predicates". However, the RDF specification states that a predicate URI reference may also be a node in the graph. This means that these terms can be used as "subject" or "object" in RDF statements (so they can be used for classification in SENGŚ). However, this can influence semantical reasoning.

A separate matter to be discussed is the use of blank nodes. A blank node (also called a bnode) is a node in an RDF graph representing a resource for which a URI or literal is not given. Such a node may appear only as "subject" or "object" in a statement, and the scope of it is local [32,36]. A major use of blank nodes is to encode n-ary relations (applicable when modeling complex attributes), but overcoming some syntactical limitations of RDF and encoding constructs of other languages like OWL are also important. Most of the instances declared according to the SENGŚ ontology will get their own URI. However, some of them might be bnodes, such as, for example, instances of the `RelatedLinkInfo` class. From a technical perspective, the possibility of declaring complex attributes using an auxiliary bnode that binds them together may be tricky. This is especially true if such an auxiliary node is going to be reused. Then, insertion of triples into the graph meets the problem of bnode merging and triple removal (see the "Shared blank nodes, unions, and merges" section in [32]), and complicate data linking.

3.3. Georeferences

The management of information about geographical objects in the long term causes many problems. These objects arise, divide, disappear, transform, are replaced with other objects, etc. Their surroundings also evolve. Thus, modeling and maintaining knowledge about the course of all these changes are not easy. In particular, keeping track of renaming is challenging. The same names can be transferred over time to different objects, and with the change of the nature of the objects, they can change the meaning (e.g., the name of a building may transfer to the name of a settlement).

The authors of the original dictionary tried not only to explain the origins of names, but also to show them in a historical context. Therefore, the description of the entries contains references to the resources based on which the previously valid names were identified through historical analyses, along with the year of introduction. The analyses were carried out to check whether a given name really corresponded to the place or object

under consideration. This required finding the descriptive attributes of the objects and their probable locations—all from written text.

A fairly accurate location can be discovered from the description of administrative affiliation. However, it would be much easier to analyze name changes based on the exact geographic location. However, it is difficult to find reliable data in historical resources. When the resource is a text document, the cause is lost by definition. Old maps created without any coordinate reference system do not help much. Even maps edited using some geographic projections do not guarantee success due to difficulties in matching names found in old documents to names presented on these maps.

A famous case of using a descriptive location to find a real object was the finding of the mythical Troy by Heinrich Schliemann based on Homer's works [37]. In SENGŚ, similar examples exist. One of them is *Zawisna*. The etymology of this name, according to the authors of the dictionary, translated into English, is: the cultural name Zawisna: adjective: envious (pol. zawisny), noun: envy (pol. zawiść), i.e., "a settlement that causes envy". This name appears in SENGŚ several times:

- Zawisna (1, vol. 4)—smelter and settlement, now part of the town of Praszka (name transferred from the name of the settlement)—with the oldest name dating back to 1845 according to the historical resources.
- Zawisna (Sowisna, Eichelhof) (1, vol. 16)—premises/buildings to Ciecierzyna, Byczyna municipality, Kluczbork district (pol. *powiat*), Opole voivodeship—with multiple name forms, of which the oldest one dates back to 1783. Etymology: *eichelhof* (ger.)—"oak mansion".
- Zawisna (Grenzwiese) (2, vol. 16)—Part of Nowa Wieś Oleska, Gorzów Śląski municipality, Olesno district (pol. *powiat*), Opole voivodship—with multiple name forms, of which the oldest one dates back to 1834. Etymology: *grenzwiese* (ger.)—"border meadow"—an artificial German name from 1936;.
- Zawisna (Grenzmühle) (3, vol. 16)—carding mill with buildings to Zawisna, Olesno district (pol. *powiat*), Opole voivodship. Etymology: grenzmühle (ger.)—"border mill" (compare *die Grenze* (ger.)—"border", *die Mühle* (ger.)—"mill") has been replaced with the Polish name Zawisna. The historical resource list includes the map *Messtischblatt 1:25,000 (symbol 4976)* published in 1940.

Based on the administrative affiliation determined by the authors, it can be concluded that the names in (1, vol. 4), (2, vol. 16), and (3, vol. 16) refer to objects located close to each other. The Topographic Card of the Kingdom of Poland (pol. *Topograficzna Karta Królestwa Polskiego*) from 1843 could help determine their original location (Kol.I Sek.V., https://academica.edu.pl/reading/readSingle?page=6&uid=3742198). This map covers the geographical area bordering the Silesia and luckily also these objects. Figure 7 shows a fragment of this map. Marked on it are places that can be associated with the name Zawisna. This is where the problem arises—it is not known exactly which object to match to which name in the dictionary.

The carding mill (3, vol. 16) should probably be associated with the mill C). On the other hand, part of *Nowa Wieś Oleska* (2, vol. 16) could be associated with the settlement *Neydorf* A) or *Zawisno* B). The *Zawisna grange* (pol. *Folwark Zawisna*) D) should probably be rejected because it is too far from the place matching the description. The smelter (1, vol. 4) is a bit more difficult to locate. By studying the history of the region, one can find out that low-percentage iron ores were mined in the Prosna river valley and its tributaries, and charcoal from the surrounding forests was used to smelt pig iron. It would probably be possible to find information in the traditional archives about where the smelter was actually located, and to check this information against current maps or in the field. However, without the need to make long trips, the already compiled data published on websites can be used, such as a portal that collects old and current photos of historical objects (Zawisna border crossing, https://polska-org.pl/5719736,Praszka,Zawisna_czesc_Praszki.html) or a website that collects genealogical data (a historical place of Zawisna, http://genealogia.mrog.org/Zawisna.html).

Figure 7. Fragment of the Topographic Card of the Kingdom of Poland with marked objects potentially related to *Zawisna*: (**A**) *Neydorf*—a settlement (now the village of *Nowa Wieś Oleska*), (**B**) *Zawisno*—a settlement near the village of *Praszka* (currently non-existent), (**C**) a mill that lies on the border of the former Kingdom of Poland (currently non-existent), and (**D**) *Folwark Zawisna* (currently, there are buildings in its place).

In many cases, when taking into account the names of large objects, e.g., city names, it is possible to obtain the geolocation automatically (directly from gazetteers, or indirectly after georectifying old maps). In particular, the following official Polish date sources [38] are useful for the area of Silesia:

- *Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju* (TERYT: "National Official Register of the Territorial Division of the Country");
- *Państwowy Rejestr Nazw Geograficznych* (PRNG: "National Register of Geographical Names");
- *Państwowy Rejestr Granic* (PRG: "National Register of Boundaries");
- *Baza Danych Obiektów Ogólnogeograficznych* (BDOO: "General Geographic Database");
- *Baza Danych Obiektów Topograficznych* (BDOT: "Database of Topographic Objects");
- *Komputerowa mapa podziału hydrograficznego Polski* ("Computer map of hydrographic division of Poland");
- The outcomes of institutions responsible for names standardization, such as *Komisja Nazw Miejscowości i Obiektów Fizjograficznych* (KNMIOF: "Commission on Names of Localities and Physiographic Objects") and *Komisja Standaryzacji Nazw Geograficznych poza Granicami Rzeczypospolitej Polskiej* (KSNG: "Commission on Standardization of Geographical Names Outside the Republic of Poland").

One may also consider unofficial data sources:

- Geonames (http://www.geonames.org/export/web-services.html);
- OpenStreetMap (OSM; https://gis-support.pl/openstreetmap-jak-pobrac-dane/);
- Getty Thesaurus of Geographic Names® Online (https://www.getty.edu/research/tools/vocabularies/tgn/);
- Mapster (http://igrek.amzp.pl/search.php?range=short);
- Wikipedia (https://www.wikipedia.org/).

However, even when working with these sources, unexpected problems may be encountered. This can be illustrated by the example of *Altkemnitz—Stara Kamienica*. A request for the location of `Stara Kamienica` can be sent to PRNG via the SPARQL endpoint (http://semantic.geoportal.gov.pl/):

```
select ?l {
?a <https://pzgik.geoportal.gov.pl/ontologies/prng/nazwaGlowna>\linebreak "
    Stara Kamienica" .
?a <http://www.opengis.net/ont/geosparql#hasGeometry> ?g .
?g <http://www.opengis.net/ont/geosparql#asWKT> ?l} limit 10
```

The response will be a WKT string: `POINT (15.5720119127338 50.9200692737104)`. This point is located on the site of the ruins of the Stara Kamienica castle (an important building identified with the village). However, from the Wikipedia article (https://pl.

wikipedia.org/wiki/Stara_Kamienica), the following location can be retrieved: 50°55′04″N 15°33′49″E (which is `POINT (50.9177778 15.5636111)`). By navigating further, this point can be visualized on a map with the aid of OpenStreet Map (https://www.openstreetmap.org/?mlat=50.917778&mlon=15.563611&zoom=11#map=16/50.9172/15.5641). It turns out that the marker is placed outside the town (see Figure 8a). In addition, the list of previous town names published on Wikipedia only partially coincides with the names mentioned in SENGŚ and does not offer any references to the historical resources.

Figure 8. Map of Stara Kamienica: (**a**) Retrieved from the OpenStreet Map service (in EPSG:4326). (**b**) Retrieved from the Mapster website and georectified (in EPSG:2180). The point obtained from Wikipedia is shown with a marker in (**a**) and is labeled with A in (**b**). The point labeled with B in (**b**) represents the correct location as delivered by *Państwowy Rejestr Nazw Geograficznych* (PRNG: "National Register of Geographical Names").

This simple experiment showed that data obtained from unofficial sources should be treated with great caution. Perhaps the location given on Wikipedia is calculated as an administrative area centroid. Incidentally, the use of a centroid can cause a discrepancy between the original and computed locations, but also errors in topology. It happens that the center of gravity of a city located on a river shifts from one side to the other as the borders change.

Some entries in SENGŚ hold references to maps or atlases. However, other resources might be considered as well. Many old maps have been digitized and can be used to determine toponyms' geographic locations that are not found in official sources. It is enough to georectify these maps and read the coordinates of recognized objects. However, selecting a proper coordinate reference system and applying correct, accurate conversion when integrating various data sources is crucial.

In Poland, for example, the 1992 National Projected Coordinate Reference System, PL-1992 (ETRS89/Poland CS92, EPGS:2180), is used for topographic maps at a scale of 1:10,000 and smaller. This plain orthogonal coordinate system was officially introduced through the Council of Ministers' decree in 2000 [39] (with further updates). It resulted from works started at the beginning of the 1990s, including Poland's territory in the European system of spatial references (ETRS) with the ETRF'89 system and the GRS-80 ellipsoid. It is based on a geocentric ellipsoid, GRS-80 (practically insignificantly different from the WGS-84 ellipsoid), and a Gauss–Krüger projection in one ten-degree zone (enough for the whole territory of Poland), with a meridian axial $L_o = 19°$ and a scale factor $m_o = 0.9993$. The linear distortions range from -70 cm/km in the central meridian to about +90 cm/km on the country's edges. The EPSG:2180 definition is publicly available in WKT format (https://epsg.io/2180), and many GIS (Geographic Information System) tools support algorithms for coordinates conversion.

For illustrative purposes, an old map published in 1939 with the name Altkemnitz was found on the Mapster website (http://mapy.amzp.pl/tk25_list.cgi?show=5059;sort=w) and processed. Figure 8b shows the fragment of that map after georectification using the QuantumGIS software. This software was able to do the conversion between different coordinate systems (EPSG:4326 and EPSG:2180) on the fly. The points obtained from

Wikipedia (A) and PRNG (B) are depicted. As confirmed, using an old map for geolocation discovery can end up with the same results as delivered from PRNG (the distortions caused by coordinate conversion are negligible). This example shows another essential issue. The castle presented on the map has survived to modern times. It can be found on a modern map, so the results of geolocation discovery can be verified. However, quite often, the objects of interest presented on old maps no longer exist.

4. Discussion

Designing dictionaries of geographical names is a challenge. The same place can be the origin of several topographic/geographic features. Geographic features may merge and bifurcate in time (which happens when partitioning or uniting cities). Names may appear in various languages. Official names may change over time due to administrative decisions, while unofficial names (name variants) may be preserved for longer in local communities. The authors of the original printed work tried to face these problems. They proposed a standardized structure for dictionary entries. However, due to the use of natural language, many exceptions have emerged. Thus, all the assumptions about the data structure had to be revised. The existing ontologies (designed for linguistic research) and ISO standards (which describe the way of data presentation for printed dictionaries) or traditional lexicographic methods did not help much. Therefore, a new model has been proposed that encompasses the printed dictionary as much as possible.

The project started with a relational model proposal. The technology stack used (Python, Django) and other factors—security audits, access permission, code review, and others—enforced this. Next, the following approach was successfully applied in the implementation of the AZON platform, and the relational model was mapped onto the semantic one.

During the construction of the SENGŚ prototype, a difficult task was to convert dictionary scans into a form that would comply with the proposed models. The overworked solution combined several processing steps. At first, with the aid of commercial OCR (Optical Character Recognition) software, the scans were converted into text documents. These documents underwent manual corrections, as the number of errors was quite large. The text processing that was applied next involved the use of regex (regular expression) rules. It resulted in structured data stored in JSON files. The regex used captured as many detectable features as possible. Finally, the conversion of these JSON files ended up with JSON files ready to feed the relational database directly as well as to serve as a base for generating triples compliant with the designed ontology. There were considerable difficulties in synchronizing data of both models. However, due to editorial limitations, these issues are not discussed.

At present, the SENGŚ prototype offers user interfaces, on which one can browse through registered toponyms and manage them. The final implementation should offer a list of features similar to those available in the AZON platform. However, the work has not been finished yet. The AZON platform offers: a presentation front-end (where users can view, search, and retrieve various records, https://zasobynauki.pl/); a management front-end (where users can manage records according to their privileges, https://deponuj.azon.e-science.pl/); a semantic interface (manifested by direct access to semantic records through content negotiation); and the SPARQL query tool, including a query editor with some predefined samples and a SPARQL endpoint (https://sparql.e-science.pl/). The same features will be available in SENGŚ (and are already implemented to a great extent).

The AZON platform serves as a repository of resources cited in SENGŚ. The whole idea of such integration can be explained as follows: Every entry in SENGŚ that has a historical part offers links to the proper records in AZON. Records in AZON deliver metadata of cited historical resources and also links to where they were published (if such links are available).

Currently, over 2000 historical resources cited in SENGŚ have been registered in the AZON platform, but have not been published yet (they are still under the review process).

However, the SENGŚ dataset includes their URIs. Proper generation and maintenance of these URIs is crucial. In both solutions (AZON and SENGŚ), all entities have numerical identifiers generated by relational databases. These identifiers appear in the URIs of triples in the graph databases. The relational and graph databases are synchronized. Forms "attached" to the relational database allow the CRUD (create, read, update, and delete) operation, but every "save" action also triggers updates to the graph database. However, to update a triple, it must be removed and then re-inserted into the graph with its property values changed (property values in the graph database are not editable, unlike attributes in the relational database, and the concepts of "identifier" or "primary key" are not working). So, for consistency, both the delete and insert operations must be performed with the same triple URI, which is a bit complicated.

Matured graph databases offer SPARQL endpoints by default. However the implementation of URI dereferencing (through the content negotiation) had to be done separately. Additionally, due to security reasons, all the endpoints should be protected from DDoS (distributed denial of service) attacks. However, these technical details are out of the scope of this manuscript.

Regarding the preparation of data for SENGŚ, there are many issues related to it: inconsistent use of abbreviations in all 17 volumes, mistakes in historical resource citations (which are the essential source of information for historians and other researchers), changes in the administrative division, and the like. Thus, all the data published on a prototype portal (https://sengs.e-science.pl/) need further corrections. As long as the core data are not shaped well to their final form, the persistent and final URIs cannot be offered. Thus, the SENGŚ prototype, at the moment, does not deliver pure semantic data nor an SPARQL endpoint.

The big problem with the integration of different data sources is the correct understanding of what the original resource and its derivatives are. In the case of SENGŚ, the prepared dataset is intended to be treated as reference data. The published URIs should be used in the Linked Open Data world as "original"—the information collected was checked and verified by historians (the cited documents, maps, registers, etc. are sometimes a hundred years old, and their importance is out of the question). Of course, making the references to the other sources of geographical information, such as Geonames, is perfectly right, but the original references are much more significant than those derived. SENGŚ is intended to deliver such references. In that sense, it plays the role of a primary source of data. However, SENGŚ refers to old documents. If such documents were scanned and published in some digital libraries or archives, SENGŚ might deliver links to them via metadata deposited in AZON (in that case, it will play a role of a secondary source of data). Work is currently underway to obtain these documents from publicly available sources.

The information model of SENGŚ includes attributes designed to store geographic location. As the original dictionary does not provide this (there are no geographic coordinates in the printed volumes), the current dataset misses these values. An attempt to discover such georeferenced locations automatically from the official sources (national registries) as well as unofficial ones (Gemet, TGN, Wikipedia) did not succeed, as many of the toponyms could not be found anywhere. It was clear from the beginning that entries discovered in the old archives or that were reconstructed do not exist in these sources. Interestingly, Wikipedia.de ensured the most successful search (local communities are much more interested in promoting their history than big organizations). Another thing that matters is the already-discussed accuracy of the georeferences found.

According to UX (user experience) guidelines, the software systems should deliver information organized into categories to make the exploration friendlier to the user. Therefore, in the SENGŚ model, the `ToponymClass` (a class with entities such as oikonym, oronym, etc.) and `ObjectCharacteristics` (a class with entities such as city, village, etc.) were introduced. The entities of these classes appear in the index of the SENGŚ portal (https://sengs.e-science.pl/indexes/#type).

The SENGŚ dictionary is monolingual in its origin. Its translation is possible, but without a deep understanding of the meaning of words used as toponym names in its original language, it would be cumbersome. SENGŚ is still waiting for such an action. However, despite these imperfections, it can already serve anyone interested in the origin of Silesian geographical names.

5. Conclusions

This article presents issues related to the digitization of the etymological dictionary of Silesian geographical names. Apart from introducing the dictionary itself and discussing the rules that were in force during its compilation, two proposals were given for representing the dictionary's content in a digital form. The first one was based on a relational model that meets the requirements of web application frameworks. The second involved a graph model, which ensures the realization of the idea of linked data and enables integration with the semantic web. Both proposals were successfully applied in the prototype designed to popularize knowledge about Silesian toponyms and to serve as a reference source. The prototype was deployed as a web application, offering over 32,000 geographical names along with etymological descriptions and references to more than 2000 sources.

Author Contributions: Conceptualization, co-creation of a relational model design, design of a graph model, improvement of data, delivery of data extensions, shaping of the user interfaces of the SENGŚ prototype, writing, and editing. The author have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by Digital Poland Project Centre (CPPC) grant number POPC.02.03.01-00-0010/16-00. The APC was funded by Wrocław University of Science and Technology.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The team research and contact funding of the original SENGŚ are owed to the Silesian Institute (*Instytut Śląski*), renamed later as the Governmental Research Institute—Silesian Institute in Opole (*Państwowy Instytut Naukowy, Instytut Śląski w Opolu*, PIN-IŚ). The work on the web-based SENGŚ was initiated within the scope of the Aktywna Platforma Informacyjna e-scienceplus.pl, POPC.02.03.01-00-0010/16 project with the PIN-IŚ' permission. The information presented on the web pages and embedded inside their semantic counterparts corresponds fully to the content of the fifth to the seventeenth volumes of *Słownik etymologiczny nazw geograficznych Śląska* published by PINIŚ in the years 1970–2016 with some extensions. The consistency with toponyms described in the first four volumes was preserved. The web-based SENGŚ logically integrates with the AZON platform built within the scope of the same *Aktywna Platforma Informacyjna e-scienceplus.pl* project. The AZON platform serves as a digital repository for historical resources mentioned in the dictionary, and delivers a bibliographic description of the resources and, if possible, instances of these resources in a digital form. The Wrocław Center for Networking and Supercomputing (WCSS), Wrocław University of Science and Technology maintains both the web-based SENGŚ and the AZON platform. The author wishes to thank: Wrocław University of Science and Technology for funding the statutory activity, and for funding the acquisition and conducting of the AZON projects; Wrocław Center for Networking and Supercomputing for providing resources; WCSS Development Team: Adam Włodarczyk, Dobrosław Kowalski, Marcin Stajno, Monika Długosz, Mateusz Bolek, Bartosz Karpiński, Mariusz Uchroński, Agnieszka Kwiecień, Miłosz Białczak, Maciej Olejnik, Stefan Piróg, and Marek Rybak, for participation in the design and development of the SENGŚ platform; Roman Ptak, for his cooperation in information model design, data processing, and other activities that led to the SENGŚ prototype deployment; Professor Grzegorz Strauchold for the fruitful discussions and suggestions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ginsburg, R.; Khidekei, S.; Knyazeva, G.; Sankin, A. *A Course in Modern English Lexicology (Revised and Enlarged, Second ed.)*; VYSŠAJA ŠKOLA: Moscow, Russia, 1979.
2. Harpring, P. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*; Getty Research Institute: Los Angeles, CA, USA, 2010.
3. Kubik, T. Role of Thesauri in the Information Management in the Web-Based Services and Systems. In *Transactions on Computational Collective Intelligence III*; Nguyen, N.T., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 25–49.
4. INSPIRE Registry, Theme Register: Names, Definitions (From the INSPIRE Directive) and Descriptions (Based on the Data Specifications Technical Guidelines), 2020. Available online: http://inspire.ec.europa.eu/theme/gn (accessed on 2 October 2020).
5. Hill, L.L. *Georeferencing: The Geographic Associations of Information*; MIT Press: Cambridge, MA, USA; London, UK, 2006.
6. Cimiano, P.; Chiarcos, C.; McCrae, J.P.; Gracia, J. Linguistic Linked Open Data Cloud. In *Linguistic Linked Data: Representation, Generation and Applications*; Springer International Publishing: Cham, Switzerland, 2020; pp. 29–41.
7. Berners-Lee, T. Linked-Data Design Issues. W3C Design Issue Document, 2009. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 2 October 2020).
8. Babczyński, T.; Kubik, T.; Ptak, R.; Strauchold, G. GIS as a tool to analyze the history of Silesia and the changes in its political (and cultural) geography. *Stud. Geohistorica* **2016**, *4*, 113–141.
9. Sochacka, S. *Słownik etymologiczny nazw geograficznych Śląska. Suplement A–Ż*; Państwowy Instytut Naukowy, Instytut Śląski w Opolu: Opole, Poland, 2016.
10. Semkowicz, W. Podstawy historyczno-geograficzne Śląska. In *Historia Śląska od Najdawniejszych Czasów do Roku 1400*; Spulera, B., Translator; Die Historisch-Geographischen Grundlagen Schlesiens: Berlin, Germany, 1935; Volume 1.
11. Arnold, S. *Terytoria Plemienne w Ustroju Administracyjnym Polski Piastowskiej (w. XII-XIII)*; Prace Komisji dla Atlasu Historycznego Akademii Umiejętności: Kraków, Poland, 1927. (In Polish)
12. Hosák, L.; Šrámek, R. *Místní jména na Moravě a ve Slezsku I, A–L, II, M–Ž*; Academia, Tisk 2, Brno: Praha, Czech Republic, 1970–1980.
13. Howard, J. *Lexicography: An Introduction*, 1st ed.; Routledge: London, UK; New York, NY, USA, 2002.
14. ISO 1951:2007. *Presentation/Representation of Entries in Dictionaries—Requirements, Recommendations and Information*; International Organization for Standardization: Geneva, Switzerland, 2007.
15. Pras, A.; Schoenwaelder, J. On the Difference between Information Models and Data Models. RFC 3444, IETF, 2003. Available online: http://tools.ietf.org/rfc/rfc3444.txt (accessed on 2 October 2020).
16. Burnard, L.; Bauman, S. (Eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.6.0. Last updated on 16th July 2019*; TEI Consortium, 2019. Available online: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (accessed on 2 October 2020).
17. ISO 25964-1. *Information and Documentation—Thesauri and Interoperability with Other Vocabularies—Part 1: Thesauri for Information Retrieval*; Standard; International Organization for Standardization: Geneva, Switzerland, 2011.
18. Miles, A.; Bechhofer, S. SKOS Simple Knowledge Organization System Reference. W3C recommendation, W3C, 2009. Available online: http://www.w3.org/TR/skos-reference (accessed on 2 October 2020).
19. "About WordNet." WordNet. Princeton University. 2010. Available online: http://wordnet.princeton.edu (accessed on 2 October 2020).
20. Maziarz, M.; Piasecki, M.; Rudnicka, E.; Szpakowicz, S.; Kędzia, P. PlWordNet 3.0—A Comprehensive Lexical-Semantic Resource. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; Calzolari, N., Matsumoto, Y., Prasad, R., Eds.; ACL: Stroudsburg, PA, USA, 2016; pp. 2259–2268.
21. McCrae, J.; de Cea, G.A.; Buitelaar, P.; Cimiano, P.; Declerck, T.; Pérez, A.G.; Gracia, J.; Hollink, L.; Montiel-Ponsoda, E.; Spohr, D.; et al. The Lemon Cookbook. 2012. Available online: https://lemon-model.net/lemon-cookbook/index.html (accessed on 2 October 2020).
22. Rospond, S. *Klasyfikacja Strukturalno-Gramatyczna Słowiańskich Nazw Geograficznych*; Prace Wrocławskiego Towarzystwa Naukowego, Seria A, nr 58; Państwowy Wydawnictwo Naukowe: Wrocław, Poland, 1957.
23. Szymanek, B. Compounding in Polish and the absence of phrasal compounding. In *Further Investigations into the Nature of Phrasal Compounding*; Trips, C., Kornfilt, J., Eds.; Language Science Press: Berlin, Germany, 2017; pp. 49–79.
24. Codd, E.F. *The Relational Model for Database Management: Version 2*; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1990.
25. Curé, O.; Blin, G. (Eds.) Chapter Two—Database Management Systems. In *RDF Database Systems. Triples Storage and SPARQL Query Processing*; Morgan Kaufmann: Boston, MA, USA, 2015; pp. 9–40. [CrossRef]
26. ISO/IEC 13249-3:2016. *Information Technology—Database Languages—SQL Multimedia and Application Packages—Part3: Spatial*; Standard; International Organization for Standardization, International Electrotechnical Commission: Geneva, Switzerland, 2016.
27. ISO/DIS 24613-3. *Language Resource Management—Lexical Markup Framework (LMF)—Part 3: Etymological Extension*; Working Draft; International Organization for Standardization: Geneva, Switzerland, 2019.
28. Cyganiak, R.; Wood, D.; Lanthaler, M. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, W3C, 2014. Available online: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ (accessed on 2 October 2020).

29. Guha, R.; Brickley, D. RDF Schema 1.1. W3C Recommendation, W3C, 2014. Available online: http://www.w3.org/TR/rdf-schema/ (accessed on 2 October 2020).
30. W3C OWL Working. OWL 2 Web Ontology Language. Document overview (Second Edition). W3C Recommendation, W3C, 2012. Available online: http://www.w3.org/TR/owl2-overview/ (accessed on 2 October 2020).
31. ISO 21127:2014. *Information and Documentation—A Reference Ontology for the Interchange of Cultural Heritage Information*; Standard; International Organization for Standardization: Geneva, Switzerland, 2014.
32. RDF 1.1 Semantics. W3C Recommendation, W3C, 2014. Available online: http://www.w3.org/TR/2014/REC-rdf11-mt-20140225/ (accessed on 2 October 2020).
33. OWL 2 Web Ontology Language Mapping to RDF Graphs (Second Edition). W3C Recommendation, W3C, 2012. Available online: http://www.w3.org/TR/2012/REC-owl2-mapping-to-rdf-20121211/ (accessed on 2 October 2020).
34. DCMI Usage Board. DCMI Metadata Terms. DCMI Recommendation, Dublin Core Metadata Initiative, 2020. Available online: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (accessed on 2 October 2020).
35. DCMI Usage Board. Using Dublin Core—Dublin Core Qualifiers. DCMI recommended resource, Dublin Core Metadata Initiative, 2005. Available online: https://www.dublincore.org/specifications/dublin-core/usageguide/qualifiers/ (accessed on 2 October 2020).
36. Hogan, A.; Arenas, M.; Mallea, A.; Polleres, A. Everything you always wanted to know about blank nodes (but were afraid to ask). *J. Web Semant.* **2014**, *27*, 42–69. [CrossRef]
37. Martínez, O. How archaeologists found the lost city of Troy. *Natl. Geogr. Hist. Mag.* **2018**. Available online: https://www.nationalgeographic.com/history/magazine/2015/12/the-lost-city-of-troy/ (accessed on 2 October 2020).
38. Rymut, K. The Polish Toponymic Guidelines, 1993. Available online: http://ksng.gugik.gov.pl/pliki/the_polish_toponymic_guidelines.pdf (accessed on 2 October 2020).
39. Council of Ministers. Regulation of the Council of Ministers of 8 August 2000 Concerning the National Reference System (pol. Rozporządzenie Rady Ministrów z Dnia 8 sierpnia 2000 r. w Sprawie Państwowego Systemu Odniesień Przestrzennych. In *Dz.U. 2000 nr 70, poz. 821*. Available online: http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20000700821 (accessed on 2 October 2020).