The problem of using persistent identifiers for historical geographical objects

Grzegorz Myrda
https://orcid.org/0000-0002-2756-8654
Tadeusz Manteuffel Institute of History, Polish Academy of Sciences

Tomasz Panecki
https://orcid.org/0000-0003-3483-2035
Tadeusz Manteuffel Institute of History, Polish Academy of Sciences

Abstract: The authors describe the usage of persistent identifiers (PIDs) for historical geographical objects. They provide PIDs' definition and scope of use as well as characterise the process of data harmonisation and PIDs' creation. The article describes and assesses certain approaches used in different projects. Most often, internal identifiers are used, although their stability is not guaranteed. References are also made to external data stores such as Geonames and Wikidata.

Keywords: identifier, harmonisation, PID, historical geographical object

Zarys treści: Autorzy podejmują problem stosowania trwałych identyfikatorów (PID) odnośnie do historycznych obiektów geograficznych. Przytaczają definicję i zakres wykorzystywania PID-ów, omawiają kwestię harmonizacji danych historyczno-geograficznych oraz zasady tworzenia PID-ów. W artykule omówione zostały dotychczasowe rozwiązania stosowane w różnych projektach. Najczęściej używa się wewnętrznych identyfikatorów, których stabilność nie jest określona. Mamy także odwołania do zasobów zewnętrznych (Geonames, Wikidata).

Słowa kluczowe: identyfikator, harmonizacja, PID, historyczny obiekt geograficzny

Data harmonisation

The resource identifiers we use when conducting research play a key role in harmonising data from different sources, such as national and academic, contemporary and historical sources. The identifiers make it possible to establish direct and numerical links to the objects introduced into the system based on references made to them in different historical sources covering different time periods. It is actually time (meaning the diachronic approach) that causes the most difficulties in identifying objects. However, we should not forget about the issues resulting from the synchronic approach (similar time, different source). The identified objects, meaning those assigned the same identifier, are the same objects regardless of their attributes that are changing over time. For instance, attributes of a settlement may include its name, type, location, area occupied, etc.[1]

The identity of geographical objects, and the ascertainment of their continuity over time, becomes a key issue from the perspective of conducting historical and geographical research using GIS tools. It is also the subject of an interdisciplinary discussion of historians, geographers, philosophers and computer scientists. Identification as such is an attempt at a compromise between the intuition of a researcher, who points to the same objects in different sources, and the need to identify these objects numerically in information systems. Based on the research

[1] P. Garbacz, A. Ławrynowicz, B. Szady, *Identity criteria for localities*, in: *Formal ontology in information systems*, ed. S. Borgo, P. Hitzler, O. Kutz, Amsterdam–Berlin–Washington 2018 (Frontiers in Artificial Intelligence and Applications, 306), pp. 47–54.

http://dx.doi.org/10.12775/SG.2020.08

conducted so far on historical and contemporary settlements, we can indicate four main constitutive properties of geographical objects that show their identity over time: proper name, location, type and mereological relations. The first two are taken into account most frequently.[2]

The issues relating to a proper name include its unchangeability, its varieties, variants, but also the problem of different languages. While there is no doubt that the names *Grodzisko*, *Gratez* and *Grodzisk Wielkopolski* refer to the same object, the names *Kąty* and *Winkel* may arouse doubts, even though it is the same expression rendered in different languages. However, if we take the names *Zielonka* and *Przyłęk*, we must use an additional criterion for their identification. In this case, this criterion will be the identical location of the objects. We can assume that geographical objects which are located in the same place according to different sources (assuming that their object classes are compatible) are the same objects, even if they have different proper names. However, this assumption may turn out to be incorrect, as it is difficult to define the compatibility of location which is affected by, among others, the level of generalisation used (if a settlement is seen as a point or an area), the accuracy of the map and the spatial development or even actual relocation of a settlement (for example Nieszawa[3]) which does not affect its identity. When identifying an object, we also sometimes consider the compatibility of object types, such as towns, villages, grain milling hamlets and smithery settlements in the case of a settlement. Of course, the nature of a settlement may have changed over time, but if we have references from similar periods, the difference in the type of a settlement with the same name and location should raise doubts as to its identification.[4] Another example is the diachronic identification of objects from different classes, such as a physiographic object and a settlement in this case. In Nowy Tomyśl Poviat, in the place where the former settlement called Bobrówka was located, there is now a swamp bearing the same name.[5] Therefore, we have a change of type here (and even of object class), but not of name and location, so we can say that it is the same object. The identification is also significantly influenced by mereological relations, especially in the diachronic context. There are many cases where, during the settlement process, a settlement was divided into two or more parts or became part of another settlement. If a reference resource provides an identifier for two equivalent settlements (for example with an annotation *-Dolny* – 'Lower', *-Górny* – 'Upper'), and earlier the settlement was mentioned as one entity, we have a problem with its identification. An example may be Psary (Będzin Poviat): its diachronic identification is ambiguous because of complicated mereological relations.[6]

Even though much experience has been gathered in historical and geographical research and certain solutions for numerical identification of objects have been developed, misidentification scenarios may occur and lead to negative consequences for semantic coherence of data. Two such scenarios are: (1) same settlements, different identifiers and (2) different settlements, same identifiers. The correct scenario is (3) same settlements, same identifiers. In order to achieve it, a conceptually correct data harmonisation is needed.

[2] Ibidem.

[3] W. Duży, *Powiat m. Toruń*, in: *Metodologia tworzenia czasowo-przestrzennych baz danych dla rozwoju osadnictwa oraz podziałów terytorialnych*, ed. B. Szady, [Warszawa 2019] (project report, 10.5281/zenodo.3751266), pp. 377–378.

[4] T. Panecki, *Powiat nowotomyski*, in: *Metodologia tworzenia*, pp. 257–258.

[5] Ibidem, p. 274.

[6] W. Duży, *Powiat będziński*, in: *Metodologia tworzenia*, pp. 45–49.

What is a PID (persistent identifier) and what is it used for?

We have all probably encountered the situation when at a given URL (Uniform Resource Locator) instead of the expected content we see the message: "404 – Page not found." As long as such a situation concerns data that are not very important, it does not matter much. Since the Internet is developing quickly and is being used everywhere, for everything and by everyone, such situations will become increasingly common in everyday life. However, they should not occur in cases for which the World Wide Web was invented – for disseminating scientific achievements and coordinating research carried out in different research centres. The 404 Error may be a result of not using persistent identifiers (PIDs) which identify resources in a unique and persistent manner. A lot of online and offline data do not have such identifiers or are given identifiers that only have some of the characteristics of persistent identifiers. That is why it is difficult to use them in any context other than the project under which they were created.

Gradually, more and more data types within published resources are being assigned persistent identifiers. The ISBN is one of the oldest, most common and most frequently cited examples of identification systems. Its basic feature (just like in the case of a persistent identifier for any other resource) is that it is assigned once and for all. Different identification systems may be used depending on the type of resource, whether it is digital or exists in the real world (or whether its virtual equivalent is concerned). Just as various types of resources (for example books, museum objects, records in a database) differ, so do the needs for their identification. The system used to identify them must take their individual characteristics and the resulting needs into account.

There is one common, general need, independent of the type of resource: when we share data or pieces of information that may need to be referenced (quotation) or referred to by external computer systems (harmonisation), it is very important to share these resources together with identifiers that make a long-term and stable reference to the specific resource.

We can give a few examples that show how identifiers may look, without determining whether these are persistent identifiers that meet all the pertinent criteria. Examples of identifiers from the National Register of Geographical Names (Polish: *Państwowy Rejestr Nazw Geograficznych* – PRNG),[7] Geonames[8] and Wikidata[9] concerning the same settlement are set out below. These identifiers meet most of the criteria for persistent identifiers. Yet in fact, we can only state that something is a persistent identifier when it stands the test of time. Before that, it is only its aim.

Table 1. Identifiers of Adamowice in different systems

System	Identifier
PRNG	100
Geonames	776782
Wikidata	Q4680202

Source: authors' own elaboration

Such identifiers are usually shared on the Internet and converted via resolver systems into content related to a given resource.

The following resolvers exist for the above identifiers:

– PRNG: https://pzgik.geoportal.gov.pl/prng/Miejscowosc/PL.PZGiK.204.PRNG.00000000-0000-0000-0000-000000000100-67,

[7] Regulation of the Minister of Administration and Digitisation of 14 February 2012 on the National Register of Geographical Names: *Rozporządzenie Ministra Administracji i Cyfryzacji z dnia 14 lutego 2012 r. w sprawie państwowego rejestru nazw geograficznych*, Dz.U. 2012, poz. 309. The data are available in various formats on the website of the Head Office of Geodesy and Cartography (Polish: Główny Urząd Geodezji i Kartografii): *Dane z państwowego rejestru nazw geograficznych – PRNG*, "Główny Urząd Geodezji i Kartografii" (http://www.gugik.gov.pl/pzgik/dane-bez-oplat/dane-z-panstwowego-rejestru-nazw-geograficznych-prng; accessed on: April 29, 2020).

[8] "Geonames" (https://www.geonames.org/, accessed on: April 29, 2020).

[9] "Wikidata" (https://www.wikidata.org/, accessed on: April 29, 2020).

– Geonames: https://www.geonames.org/776782/adamowice.html,
– Wikidata: https://www.wikidata.org/wiki/Q4680202.

Ensuring the persistence of these identifiers (as well as their translation into relevant URLs) is a key factor enabling further knowledge development. It is often tempting to consider something as a persistent identifier based on only one criterion: a URL under which the resource is available. This is not always enough. A URL can (and often does) undergo changes that are independent of the data manager. These can include, for example, organisational changes or modifications to the IT infrastructure (changing the server or the system for data publication). However, above all, it is not up to the recipient of the data to decide whether something is a persistent identifier. The data manager must clearly express such an aim and make it credible to the outside world that the identifiers it publishes will be persistent identifiers. Moreover, if there are many infrastructural resources, it is also important to provide access to full information about the resource identified by the persistent identifier. Geographical data are an example of infrastructural resources and they often constitute spatial context for other data.

Each of the three examples above represents a different type of data manager, each with its advantages and disadvantages. The National Register of Geographical Names (PRNG) is a state-run resource governed by law, which means that there is relatively little risk that it will be discontinued, but it is limited to a certain area. Moreover, this register does not cover historical data since the law does not prescribe it, so objects that cease to exist in the contemporary world also disappear from the database. Geonames is a worldwide database, but it operates within a commercial model on a relatively small scale. Everyone can assess its future prospects on their own. Wikidata is not only a database, but also a system enabling its functioning. Its development has been sustainable and it is difficult to imagine that it could be stopped. All the more so since it does not only cover geographical data. The choice depends, as we can see, on many factors which are not always objective. However, it seems that the most universal identifier that meets the most criteria is the one provided by Wikidata.

Notably, it is good practice to supplement persistent identifiers with a service that makes it possible to explore the relevant data under a given identifier. This is a service that is independent of the infrastructure used for assigning and hosting persistent identifiers. It enables resolving, which means converting PID values into the presentation of characteristics of a given resource. Unlike persistent identifier management infrastructure which is not always implemented using Internet technologies (for instance ISBN), resolvers are usually based on one of the Internet protocols, such as HTTP. A URL, meaning the location of a resource description, is often treated as a kind of a two-in-one solution – it acts as both an identifier and a resolver. However, a URL as such, in a very narrow sense (with no link to a suitable infrastructure for handling persistent identifiers), should not be seen as a persistent identifier. Even though the URI[10] standard may be the basis for one of possible implementations of a persistent identifier system, this is not enough since a PID should have some additional features.

What are the features of a good PID?

An identifier for a resource should meet many additional criteria compared to a normal identifier in order to be considered as persistent. It is acknowledged that

[10] T. Berners-Lee, R. Fielding, L. Masinter, *Uniform Resource Identifier (URI): Generic Syntax*, IETF, January 2005 (DOI: 10.17487/RFC3986, OCLC: 943595667).

a good identifier, together with the system enabling its functioning, should have the following main features[11]:

Uniqueness: one identifier identifies only one resource.

Stability: an identifier will never point to any other resource.

Persistence: an identifier is assigned once and for all and will always be available, just like the resource it identifies. The assessment of what the chances are that the organisation enabling the functioning of an identifier (sometimes together with basic metadata) will be able to ensure its continuous functioning in the same form is somewhat subjective, but it needs to be taken into account.

Ability to provide the right granularity: the infrastructure for handling identifiers makes it possible to version resources and identify parts of them. For example, it is possible to identify a whole set of settlements or just one of them.

URI specification compliance: the methods for handling identifiers at the level of abstract specification should be independent of technological solutions to ensure that identifiers will be able to function with the use of any technical means that may be developed in the future.

Metadata handling: it should be possible to read the metadata describing a resource before moving to that specific resource.

Lack of semantics: an identifier should not contain any semantic information that may become outdated and need to be changed.

Scalability: the identifier handling system must be able to work efficiently as the number of identifiers handled increases, and be able to work 24 hours a day.

Costs adapted to capabilities: maintaining each identifier (just like maintaining an Internet domain name) often involves an annual cost. If there are many identifiers, costs increase rapidly. At one time, the Max Planck Institute wanted to assign DOIs to 500,000 objects that it had in its database. Since such a solution would cost USD 30,000 a year, this method was not implemented.[12]

Rules for creating persistent identifiers

When creating a persistent identifier, firstly you have to determine the identity of the resource which is to receive a persistent identifier. This resource can later change its various characteristics and further characteristics can be added, but it cannot point to any other object. A resource that is once assigned a specific identifier can no longer change it, so it cannot lose its identity. Depending on the type of resource, different elements determine its identity. It is also not easy to establish the criteria that define it. Therefore, the assignment of persistent identifiers should be well thought-out. When it comes to identifying geographical objects, we should be able to determine whether we are still dealing with the same object despite its changes and the passage of time. Many elements may undergo changes, for example proper name (for instance renaming a settlement following a partition of a country), location (relocation of a settlement as a result of a change in hydrographic conditions), type (granting municipal rights). There is practically no feature that can never change. Usually, only a change of several features at the same time can be considered as a change of identity and create a need for a new identifier.

In the case of geographical data, there is no single manager who owns information on this type of resource. That is why we have to bear an additional aspect in mind:

[11] *Persistent identifiers best practices*, CEOS Data Stewardship Interest Group, ver. 1.1, pp. 9–11 (https://earth.esa.int/documents/1656065/2265358/CEOS-Persistent-Identifier-Best-Practices, accessed on: December 17, 2019); *Persistent and unique Identifiers*, ed. D. Broeder et al., ver. 4, Common Language Resources and Technology Infrastructure 2009, pp. 7–8 (https://office.clarin.eu/pp/D2R-2b.pdf, accessed on: February 20, 2020).

[12] Ibidem, p. 10.

many institutions and systems must be able to identify such a resource simultaneously. This should be taken into account when designing an identification system.

After preparing the data (and their metadata) which are to receive a persistent identifier, we can proceed with the technical part, namely requesting such an identifier from a system that assigns persistent identifiers.

Internal or external system

Theoretically, it is possible to create the necessary infrastructure in your own institution to join one of the existing global systems for managing persistent identifiers. In return for hosting a fragment of a larger IT infrastructure, you are given the ability to assign an unlimited number of identifiers. This infrastructure works like the infrastructure that assigns domain addresses. However, it is much more common to purchase a certain number of identifiers from a service provider. In such a case, handling persistent identifiers consists in hosting the description of a resource available at a specific URL on your own server – while the provider's server hosts up-to--date infrastructure that redirects from the identifier's address to the address on your own server – and updating the metadata about the resource on the provider's server. According to the *Persistent identifier best practices*[13] report, the system using digital object identifiers (DOIs) is currently the most popular system on a global scale.

The DOI is a consortium of publishing houses that offer a commercial infrastructure for assigning and managing persistent identifiers called DOIs. Even though it is the most popular system, the DOI is seen as a commercial solution[14] with a business model that is suitable for individual use. In other cases, using the Handle System directly is recommended (the DOI at the technological level also uses the Handle System).

PIDs and coordinates

When creating persistent identifiers, it may also be tempting to use geographic coordinates as an identifier (just like in the case of a URL mentioned above). Can the coordinates of a geographical object constitute its identifier? As previously stated, any feature of an object can change over time. This also applies to coordinates that describe a location, for example of a settlement. Therefore, just like the proper name, they cannot constitute an identifier. What we want to achieve is the identification of a geographical object, not a place.

Standards and systems

Currently, there are many standards and systems that regulate assigning persistent identifiers and linking them to a specific digital resource. Among these, the following systems are the most important:

- URI-URN Standard – IETF/W3C (example: urn:isbn:0451450523),
- Handle System – Corporation for National Research Initiatives Virginia (example: hdl:2381/12775),
- DOI System – International DOI Federation based on the Handle System (example: doi:10.1186/2041-1480-3-9, https://dx.doi.org/),
- ePIC – consortium of European partners for the European Research Community (example: https://hdl.handle.net/20.500.12434/8d621959),
- PURL – redirection service (example: https://purl.fdlp.gov/GPO/gpo112620),
- UUID – universally unique number (example: 123e4567-e89b-12d3-a456-426655440000).

In addition to the most common Handle System (the DOI at the technical level also uses the Handle System), the following

[13] C. Ferguson et al., *Survey of current PID services landscape*, May 2018, p. 12 (https://zenodo.org/record/1324296, accessed on: April 24, 2020).

[14] D. Van Uytvanck, *PID policy summary*, ver. 1, Common Language Resources and Technology Infrastructure 2014, p. 2 (https://www.clarin.eu/sites/default/files/CE-2013-0340-PID-policy-summary.pdf, accessed on: April 24, 2020).

systems constitute an universal way of providing identifiers that meet some of the criteria for PIDs: the above-mentioned PURL – which is in fact a system of redirections from URLs (treated as stable addresses) to the addresses where the data or pieces of information related to a given resource are actually located (i.e. variable addresses) – and UUID – which is an identifier based on a special algorithm using, among other things, the passage of time to generate a globally unique character string that does not need to be additionally described, for example with the name of a dataset, in order to ensure that it will not be repeated in the context of another dataset.

A report[15] describing the current status of different infrastructures for handling persistent identifiers lists three PID systems that are used so far for *Temporal period & historical place* and describes the level of maturity of their infrastructures as immature. These systems are: ARK (Archival Resource Key), URI (Uniform Resource Identifier) and accession number.

According to the authors of a Finnish concept[16] of persistent identifiers for geographical objects which has been developed in accordance with the INSPIRE Directive, the identifier consists of a unique namespace of the data source, resource type, dataset identifier, local identifier and version identifier. Local identifiers are generated by the data manager. They are published in the HTTP URI format. The format of the publication address is as follows:

http://{domain name}/{URI type}/{dataset identifier}/{local identifier}[/{version identifier}]

The resource type may be one of the following:

"id" – a real-world object,

"so" – a geographical object,

"def" – a definition of a geographical object,

"doc" – documentation related to different forms of presentation.

However, it is difficult to apply a similar approach to historical data which are not in the scope of the INSPIRE Directive because they are not covered by national legislation and thus do not have a predefined manager. Therefore, it is difficult to ensure that the namespace of the data source is unique.

What should a PID indicate?

The material scope of identifiers is another issue worth considering. In the system consisting of the so-called source and resultant data, the PID of a given geographical object, for instance a settlement, is a link between these two components and makes it possible to refer the interpretation to the sources.[17] While in the resultant component a PID acts as a unique identifier and primary key, in the case of source data a PID is a foreign key and may be repeated, as a source may contain many references to a given settlement. The question is whether the scope of assigning persistent identifiers should not also apply to the identifiers for source data. In such a case, a reference could be made not only to a critical object, but also to individual source proofs that have stable identifiers. As we can see, persistent identifiers for geographical objects can be accompanied by persistent identifiers for any other data that can be referenced. All the more so if there is a chance that these identifiers can be used in an entirely different context.

Unfortunately, persistent identifiers do not solve the age-old problem of data redundancy. Theoretically (and in practice this often happens), the same object can be described with two different identifiers by

[15] C. Ferguson et al., *Survey*, p. 10.

[16] *JHS 193 unique identifiers of geographic data*, "JHS-suositukset" (http://docs.jhs-suositukset.fi/jhs-suositukset/JHS193_en/JHS193_en.html, accessed on: April 30, 2020).

[17] B. Szady, *Spatio-temporal databases as research tool in historical geography*, "Geographia Polonica," 89 (3), 2016, pp. 359–370.

two different organisations independent of each other. In such a case, a PID indicates one of the versions of the object description.

Importantly, objects should be described with terms which have definitions that can also be identified with persistent identifiers. Persistent identifiers can thus point not only directly to objects (geographical data), but also to definitions that determine the semantic framework for these data.

Existing solutions

So far, there is no coherent methodology for harmonising and identifying historical topographic objects and for creating their persistent identifiers. We can describe a number of different approaches based on several examples ("Geohistorical Data," "GBU," "GASID," "GB 1800," "HistoGIS," "Pleiades" and "NHGIS"), but we have to bear in mind that in the absence of project documentation or of a clear statement from the data manager, the stability of any identifiers should be treated with caution.

The authors of the first project, Geohistorical Data, have developed a geoportal and provide data from the 18th-century Cassini map.[18] The downloadable data include the road network and natural landscape elements, but also borders represented linearly (*limites administratives*), buildings (*taches urbaines*), settlements shown as areas and points (*chefs-lieux*) and other objects (*lieux ponctuels*). The data on the settlements are in fact part of the last three tables, which results in data redundancy. The role of an identifier for external systems is played by the URL of objects from the website of the "Des villages de Cassini aux communes d'aujourd'hui" project. This project deals with the critical processing of the data from the Cassini map, namely settlements and administrative divisions of 18th-century France.[19] However, no reference is made to external reference datasets, such as Geonames and Wikidata, and there are several fields in the database that have other identifiers whose use is unknown due to the lack of documentation.

The Beauplan's Ukraine project is a digital edition of special maps of Ukraine that were made by Guillaume Le Vasseur de Beauplan.[20] The project has a form of a digital gazetteer which contains elements of map content located according to Google Maps.[21] In the attribute table (*.shp file) of these data we have access to a lot of information, including the URL of each object which constitutes its unique identifier. We also have another identifier (*gazbu-id*) which is different from that of the URL, and a proper name, (contemporary) coordinates and the URL of the object identified in Geonames (*geonames-id*).

The GASID project (Galicia and Austrian Silesia Interactive Database 1857–1910) aims to make statistical and cartographic information on 19th-century Galicia and Austrian Silesia available to a wide range of researchers.[22] The data collected will be prepared as digital maps and presented as a geoportal in 2020, but a part of them has already been published, namely roads and settlements according to the second topographic map (the so-called Franciscan land survey).[23] The available data concerning settlements only show their names taken from the map and the contemporary names. No identifier is included, except for the internal, numerical one.

[18] J. Perret, M. Gribaudi, M. Barthelemy, *Roads and cities of 18th century France*, "Scientific Data," 2, 2015, p. 150048; "Geohistorical Data" (http://www.geohistoricaldata.org/, accessed on: April 21, 2020).

[19] C. Motte, M.-C. Vouloir, *Le site cassini.ehess.fr un instrument d'observation pour une analyse du peuplement*, "Bulletin du Comité français de cartographie," 191, 2007, pp. 68–84.

[20] B. Olszewicz, *Polska kartografia wojskowa*, Warszawa 1921, pp. 17–19.

[21] M. Polczynski, M. Polczynski, *Beauplan's Ukraine: open access georeferenced databases for studies of early modern history of Central and Eastern Europe*, "Miscellanea Geographica," 23 (3), 2019, pp. 185–193; *Beauplan's Ukraine*, "Harvard Dataverse" (https://dataverse.harvard.edu/dataverse/BU, accessed on: April 21, 2020).

[22] "GASID" (http://gasid.pl/, accessed on: April 21, 2020).

[23] D. Kaim, M. Szwagrzyk, K. Ostafin, *Mid-19th century road network dataset for Galicia and Austrian Silesia, Habsburg Empire*, "Data in Brief," 28, 2020, p. 104854.

It is interesting to examine how "A Vision of Britain through Time" platform[24] identifies objects. It is a comprehensive geoportal concerning the history of Great Britain in spatial terms which consists of many modules: statistical (quantitative thematic maps), cartographic (old maps) and semantic (search engine for places).[25] On the geoportal, each settlement is described by an identifier in the form of a URL and is also linked to other resources, such as Geonames and Wikipedia.

HistoGIS is a platform for collecting, creating and compiling geographical historical data, as well as for sharing them with other researchers. It operates under the Linked Open Data and uses SKOS to organise information.[26] The platform shows political and administrative divisions of Europe which were developed based on vectorising old and historical maps and were afterwards transformed into a common data model. The platform makes it possible to download data (in *.json format) or view them on the geoportal. The information includes an internal Permalink identifier (URL), an ID from Wikidata and mereological relations with higher and lower-level units.

Pleiades is a gazetteer of ancient places that uses semantic networks. It makes it possible to search for ancient places, display them on a map and in a graphical form, and download them.[27] The data are stored in an object database which links the following components: "places," "names" and "locations." Each resource, meaning a specific place with a specific location and name, is given a fixed and unchanging URI (Uniform Resource Identifier) as well as references to Geonames.

The American National Historical Geographic Information System (NHGIS)[28] makes it possible to download US statistical data (also in spatial data formats) for years from 1790 to the present day. The system includes both areal data (statistical areas at different aggregation levels) and point data (settlements) in different time frames. Areal data from different years are only partly linked to one another due to changes in geometry. As for point data, NHGIS uses a system of stable identifiers that indicate the same settlement in different years regardless of changes in name and even in census identifiers.

The above review shows that the issue of identifiers is treated in various ways, both as regards the stability of the internal ID and as regards the links to external systems (table 2). Two solutions (Geohistorical Data, GASID) include no such link and three solutions provide a link to reference datasets: Geonames (GBU, GB 1800) and Wikipedia / Wikidata (HistoGIS, GB 1800). HistoGIS, Pleiades and NHGIS claim to have stable identifiers.

Model used in the Department of Historical Atlas

Within historical and geographical research currently carried out in the Department of Historical Atlas of the Tadeusz Manteuffel Institute of History, Polish Academy of Sciences, we generally use two external systems and we treat their identifiers as stable. These systems are: the National Register of Geographical Names (Polish: *Państwowy Rejestr Nazw Geograficznych* – PRNG) for point data (mainly settlements) and the National Register of Boundaries (Polish: *Państwowy Rejestr Granic* – PRG) for the divisions of secular and ecclesiastical administration. It should be noted that these are national and official

[24] *Historical maps*, "A Vision of Britain through Time" (http://www.visionofbritain.org.uk/maps/, accessed on: November 21, 2019).

[25] H. Southall, *Constructing a Vision of Britain through Time: Integrating old maps, census reports, travel writing, and much else, into an online historical atlas*, in: *Historical atlas: Its concepts and methodologies*, ed. PK. Bol, Seoul 2016, pp. 133–151; H. Southall, P. Aucott, *Expressing history through a geo-spatial ontology*, "ISPRS International Journal of Geo-Information," 8 (8), 2019, p. 362.

[26] "HistoGIS" (https://histogis.acdh.oeaw.ac.at/, accessed on: April 24, 2020).

[27] "Pleiades" (https://pleiades.stoa.org/, accessed on: April 29, 2020).

[28] "IPUMS NHGIS" (https://www.nhgis.org/, accessed on: July 21, 2020).

Table 2. Features of selected geoportals

Geoportal name	Data type	Internal ID	External ID	Downloadable data
"Geohistorical Data"	settlements, administrative borders, roads, hydrography	numerical; no stability declaration	none	*.shp
"GBU"	settlements, toponyms	numerical; URL; no stability declaration	Geonames	*.shp
"GASID"	settlements, roads	numerical; no stability declaration	none	*.shp
"GB 1800"	settlements, administrative divisions	numerical; no stability declaration	Geonames, Wikidata	none
"HistoGIS"	administrative divisions	URI; stable	Wikidata	*.json
"Pleiades"	places	URI; stable	Geonames	*.json, *.csv, *.kml, *.rdf
"NHGIS"	statistical data	alphanumeric; stable	none	*.shp

Source: authors' own elaboration

registers, which is important for the stability of their identifiers. The harmonisation of resources included in the National Register of Geographical Names (PRNG) is based on methodologies developed in the Department for the Identification of Settlements, primarily on the basis of location identity and similarity of the name. Of course, some data do not have PRNG identifiers and concern settlements that disappeared, were absorbed or divided. The identification of such settlements would be vitiated by a too serious error. The data on settlements could also be harmonised with the BDOT10k database and the SIMC and TERYT registers, but the PRNG was chosen due to its scope, among others (it also includes physiographic objects[29]). In addition, PRNG is a source register for the BDOT10k database. Currently, work is also underway to harmonise administrative divisions within the cadastral precincts from the PRG and to use their identifiers to diachronically identify secular and ecclesiastical units from different periods. The methodology consists in adding appropriate attributes of administrative affiliation to cadastral precincts based on historical settlements located in their territory. If a cadastral precinct has settlements with mutually exclusive attributes (for example they belonged to different poviats in the 16th century), this cadastral precinct should be divided so that only points with coherent attributes remain in each polygon.

However, a disadvantage of the official registers mentioned above is that they only cover the territory of contemporary Poland, while in the past Polish borders extended further, especially to the east. Hence the idea of using identifiers from other resources (such as Wikidata, Open Street Map and Geonames) and treating them as references. The above-mentioned databases (or rather systems enabling their functioning) make it possible to handle data that are located anywhere in the world (we are not limited to a specific area as in the case of national resources). Moreover, these databases provide data that are ready to use. However, the Open Street Map data model does not include historical data and object identifiers can

[29] Topographic objects related to hydrography (hydronyms) and collected during the work on the "Historical atlas of Poland" also have identifiers from the electronic dictionary of Polish hydronyms ("Elektroniczny Słownik Hyrdonimów Polski", https://eshp.ijp.pan.pl/, accessed on: April 29, 2020), but we cannot be sure if these identifiers will be stable and unchangeable.

Identifiers
83131

(a) Variable Settlement

Identifiers	Names	VariableSettlementIdentifiers	StartsAt	EndsAt
155657	Stara Nieszawa	83131	1460-09-25	1554-12-31
83131	Podgórz	83131	1555-01-01	2016-12-31

(b) Manifestation of Name

Identifiers	SettlementTypeIdentifiers	VariableSettlementIdentifiers	StartsAt	EndsAt
155657	2	83131	1460-09-25	1611-11-06
155656	3	83131	1611-11-07	1833-03-26
155655	2	83131	1833-03-27	1924-12-31
155654	3	83131	1925-01-01	1938-03-31
83131	61	83131	2016-01-01	2016-12-31

(c) Manifestation of Type

Identifiers	the_geom	VariableSettlementIdentifiers	StartsAt	EndsAt
155660	POINT(18.5932311869783 52.9988146722835)	83131	1460-09-25	1554-12-31
83131	POINT(18.5916356118256 52.9921219054298)	83131	1555-01-01	2016-12-31

(d) Manifestation of Location

Identifiers	PartIdentifiers	WholeIdentifiers	StartsAt	EndsAt
7	83131	112602	1938-01-01	2016-12-31

(e) Manifestation of Mereology

Identifiers	Names	VariableSettlement-Identifiers	Types	Mereology	Geometries	Name-Identifiers	Type-Identifiers	Location-Identifiers	Mereology-Identifiers	StartsAt	EndsAt
83363	Stara Nieszawa	83131	2	Null	POINT-(18.5932311869783 52.9988146722835)	155657	155657	155660	Null	1460-09-25	1554-12-31
83364	Podgórz	83131	2	Null	POINT-(18.5916356118256 52.9921219054298)	83131	155657	83131	Null	1555-01-01	1611-11-06
83365	Podgórz	83131	3	Null	POINT-(18.5916356118256 52.9921219054298)	83131	155656	83131	Null	1611-11-07	1833-03-26
83366	Podgórz	83131	2	Null	POINT-(18.5916356118256 52.9921219054298)	83131	155655	83131	Null	1833-03-27	1924-12-31
83367	Podgórz	83131	3	Null	POINT-(18.5916356118256 52.9921219054298)	83131	155654	83131	Null	1925-01-01	1937-12-31
83368	Podgórz	83131	3	112602	POINT-(18.5916356118256 52.9921219054298)	83131	155654	83131	7	1938-01-01	1938-03-31
83369	Podgórz	83131	3	112602	POINT-(18.5916356118256 52.9921219054298)	83131	0	83131	7	1938-04-01	2015-12-31
83370	Podgórz	83131	61	112602	POINT-(18.5916356118256 52.9921219054298)	83131	83131	83131	7	2016-01-01	2016-12-31

(f) Aggregated Manifestation

Fig. 1. Information about the 83131 settlement in the database.
Source: authors' own elaboration

change (it is a very rare, but potentially possible situation).[30] If we compare the communities gathered around Geonames and Wikidata and their operating models, Wikidata seems to be the best choice for handling the data harmonisation process, even though (or maybe this is actually the reason behind it) location is not the main priority of this database.

If we have an attribute in the form of an identifier for an external dataset (which is useful at the time of harmonisation) and we want to link it to something, we have to create an appropriate data model. The Department of Historical Atlas has created a data model for the settlement network and administrative divisions which takes not only the current reality, but also the entire history into account. Its underlying principle is that each geographical object exists independently of time only in the form of an identifier. However, all its features such as name, location, type and mereology change over time, so they are not elements of the identifier. In figure 1

[30] *Can I depend on the country/city IDs of OSM?*, "Geographic Information Systems" (https://gis.stackexchange.com/questions/279755/can-i-depend-on-the-country-city-ids-of-osm?rq=1, accessed on: April 29, 2020).

you can see tables of the database containing information about a settlement with the 83131 identifier. The settlement was once called *Stara Nieszawa* and later changed its name to *Podgórz*. Similarly, as time passed, other features of this settlement changed. The value of 83131 is a persistent identifier for the settlement within the database managed by the Department of Historical Atlas.

Geographical objects are also linked to information about corresponding contemporary objects which constitute one of the sets of attributes with relevant dates. These are PRNG and PRG identifiers. Thanks to that, it is possible to move from any historical object to its contemporary version and see its entire history.

Discussion and conclusion

The discussion on stable identifiers for historical geographical data seems to concern two aspects: theoretical, related to the harmonisation of resources as such and its substantive correctness, and technical or practical, regarding the choice of a way of designing persistent identifiers and assigning them to resources.

Both aspects are related and cannot be considered independently. Scenarios in which the same settlements have different identifiers or different settlements are described with the same identifiers are unwanted; they may result from a lack of cooperation between historians and computer scientists. Developing and adopting rules for harmonisation and identification of historical data in time and space is crucial to ensure the correct modelling of these data in databases. It is important to include a possibility of expressing uncertainty in the case of such data, as historians are not always able to make a clear and certain interpretation.[31]

Problems of a more technical nature, which are also relevant from a substantive point of view, include the issue of using internal or external identifiers, as well as the scope of referring the data manager's resources to external datasets (PRNG/PRG, Geonames, Wikidata, etc.). When we decide how identifiers will be managed technically, we actually decide how the outside world will assess the credibility of our claim that our identifiers will be persistent. The criteria seem simple: if we have a lot of resources to share and at the same time we have, or can have, the necessary IT infrastructure, then we should decide to use internal identifiers (that are managed within the institution of the data manager). The question remains how much is a lot and whether we are able to provide the appropriate infrastructure. Otherwise, external identifiers are normally used, which means that the process of managing identifiers is entrusted to an external organisation that usually specialises in this type of activity. A typical example of an external identifier is the DOI since it is generated and hosted outside the organisation which is the data manager. If we see that a resource has an external identifier, we do not have to take the data manager's word for whether it is a persistent identifier. As for internal identifiers, in addition to the statement by the data manager arguing that an identifier is persistent, we have to believe this claim based on our knowledge about the basic principles of the functioning of persistent identifiers, the risks associated with the operation of specific IT infrastructures and the internal organisational solutions of the data manager.

There is no doubt that it is important and necessary to include an attribute with an external identifier in a dataset under development, but the question remains as to the extent of such harmonisation: is it enough to include an identifier for one selected resource or should we adhere to the

[31] G. Myrda, B. Szady, A. Ławrynowicz, *Modeling and presenting incomplete and uncertain data on historical settlement units*, "Transactions in GIS," 24 (2), 2020, pp. 355–370 (DOI: https://doi.org/10.1111/tgis.12609).

principle of "the more, the better"? More linked datasets mean, on the one hand, greater analytical capacity but, on the other, difficulty in ensuring coherence between these datasets which increases with the number of linked resources. The ideal solution would be to create a central list of identifiers for historical geographical objects to which the data managers would refer. However, such a solution would require the involvement and agreement of many parties, institutions and projects. ■

Bibliography

Beauplan's Ukraine, "Harvard Dataverse" (https://dataverse.harvard.edu/dataverse/BU, accessed on: April 21, 2020).

Berners-Lee T., Fielding R., Masinter L., *Uniform Resource Identifier (URI): Generic Syntax*, IETF, January 2005 (DOI: 10.17487/RFC3986, OCLC: 943595667).

Can I depend on the country/city IDs of OSM?, "Geographic Information Systems" (https://gis.stackexchange.com/questions/279755/can-i-depend-on-the-country-city-ids-of-osm?rq=1, accessed on: April 29, 2020).

Dane z państwowego rejestru nazw geograficznych – PRNG, "Główny Urząd Geodezji i Kartografii" (http://www.gugik.gov.pl/pzgik/dane-bez-oplat/dane-z-panstwowego-rejestru-nazw-geograficznych-prng; accessed on: April 29, 2020).

Duży W., *Powiat będziński*, in: *Metodologia tworzenia czasowo-przestrzennych baz danych dla rozwoju osadnictwa oraz podziałów terytorialnych*, ed. B. Szady, [Warszawa 2019] (project report, 10.5281/zenodo.3751266).

Duży W., *Powiat m. Toruń*, in: *Metodologia tworzenia czasowo-przestrzennych baz danych dla rozwoju osadnictwa oraz podziałów terytorialnych*, ed. B. Szady, [Warszawa 2019] (project report, 10.5281/zenodo.3751266).

"Elektroniczny Słownik Hyrdonimów Polski" (https://eshp.ijp.pan.pl/, accessed on: April 29, 2020).

Ferguson C., McEntrye J., Bunakov V., Lambert S., Sandt S. van der, Kotarski R., Stewart S., MacEwan A., Fenner M., Cruse P., Horik R. van, Dohna T., Koop-Jacobsen K., Schindler U., Cafferty S., *Survey of current PID services landscape*, May 2018 (https://zenodo.org/record/1324296, accessed on: April 24, 2020).

Formal ontology in information systems, ed. S. Borgo, P. Hitzler, O. Kutz, Amsterdam–Berlin–Washington 2018 (Frontiers in Artificial Intelligence and Applications, 306).

Garbacz P., Ławrynowicz A., Szady B., *Identity criteria for localities*, in: *Formal ontology in information systems*, ed. S. Borgo, P. Hitzler, O. Kutz, Amsterdam–Berlin–Washington 2018 (Frontiers in Artificial Intelligence and Applications, 306).

"GASID" (http://gasid.pl/, accessed on: April 21, 2020).

"Geohistorical Data" (http://www.geohistoricaldata.org/, accessed on: April 21, 2020).

"Geonames" (https://www.geonames.org/, accessed on: April 29, 2020).

"HistoGIS" (https://histogis.acdh.oeaw.ac.at/, accessed on: April 24, 2020).

Historical atlas: Its concepts and methodologies, ed. P.K. Bol, Seoul 2016.

Historical maps, "A Vision of Britain through Time" (http://www.visionofbritain.org.uk/maps/, accessed on: November 21, 2019).

"IPUMS NHGIS" (https://www.nhgis.org/, accessed on: July 21, 2020).

JHS 193 unique identifiers of geographic data, "JHS-suositukset" (http://docs.jhs-suositukset.fi/jhs-suositukset/JHS193_en/JHS193_en.html, accessed on: April 30, 2020).

Kaim D., Szwagrzyk M., Ostafin K., *Mid-19th century road network dataset for Galicia and Austrian Silesia, Habsburg Empire*, "Data in Brief," 28, 2020.

Metodologia tworzenia czasowo-przestrzennych baz danych dla rozwoju osadnictwa oraz podziałów terytorialnych, ed. B. Szady, [Warszawa 2019] (project report, 10.5281/zenodo.3751266).

Motte C., Vouloir M.-C., *Le site cassini.ehess.fr un instrument d'observation pour une analyse du peuplement*, "Bulletin du Comité français de cartographie," 191, 2007.

Myrda G., Szady B., Ławrynowicz A., *Modeling and presenting incomplete and uncertain data on historical settlement units*, "Transactions in GIS," 24 (2), 2020 (DOI: https://doi.org/10.1111/tgis.12609).

Olszewicz B., *Polska kartografia wojskowa*, Warszawa 1921.

Panecki T., *Powiat nowotomyski*, in: *Metodologia tworzenia czasowo-przestrzennych baz danych dla rozwoju osadnictwa oraz podziałów terytorialnych*, ed. B. Szady, [Warszawa 2019] (project report, 10.5281/zenodo.3751266).

Perret J., Gribaudi M., Barthelemy M., *Roads and cities of 18th century France*, "Scientific Data," 2, 2015.

Persistent and unique Identifiers, ed. D. Broeder, M. Dreyer, M. Kemps-Snijders, A. Witt, M. Kupietz, P. Wittenburg, ver. 4, Common Language Resources and Technology Infrastructure 2009 (https://office.clarin.eu/pp/D2R-2b.pdf, accessed on: February 20, 2020).

Persistent identifiers best practices, CEOS Data Stewardship Interest Group, ver. 1.1 (https://earth.esa.int/documents/1656065/2265358/CEOS-Persistent-Identifier-Best-Practices, accessed on: December 17, 2019).

"Pleiades" (https://pleiades.stoa.org/, accessed on: April 29, 2020).

Polczynski M., Polczynski M., *Beauplan's Ukraine: open access georeferenced databases for studies of early modern history of Central and Eastern Europe*, "Miscellanea Geographica," 23 (3), 2019.

Rozporządzenie Ministra Administracji i Cyfryzacji z dnia 14 lutego 2012 r. w sprawie państwowego rejestru nazw geograficznych, Dz.U. 2012, poz. 309.

Southall H., *Constructing a Vision of Britain through Time: Integrating old maps, census reports, travel writing, and much else, into an online historical atlas*, in: *Historical atlas: Its concepts and methodologies*, ed. P.K. Bol, Seoul 2016.

Southall H., Aucott P., *Expressing history through a geo-spatial ontology*, "ISPRS International Journal of Geo-Information," 8 (8), 2019.

Szady B., *Spatio-temporal databases as research tool in historical geography*, "Geographia Polonica," 89 (3), 2016.

Van Uytvanck D., *PID policy summary*, ver. 1, Common Language Resources and Technology Infrastructure 2014 (https://www.clarin.eu/sites/default/files/CE-2013-0340-PID-policy-summary.pdf, accessed on: April 24, 2020).

"Wikidata" (https://www.wikidata.org/, accessed on: April 29, 2020). ■

Problem stosowania trwałych identyfikatorów odnośnie do historycznych obiektów geograficznych

Streszczenie

Autorzy tekstu omawiają stosowanie trwałych identyfikatorów (PID) odnośnie do historycznych obiektów geograficznych. Rozpoczynają od zdefiniowania PID-ów jako stabilnych i trwałych identyfikatorów, które pozwalają precyzyjnie ujednoznacznić zasoby. Identyfikatory takie powinny mieć następujące cechy: unikalność (jeden zasób – jeden PID), stabilność (konkretny PID nie będzie nigdy przypisany do innego zasobu), trwałość (PID zawsze będzie dostępny), obsługiwanie różnych stopni rozproszenia (możliwość odesłania do różnych danych w zależności od ich rozproszenia lub generalizacji), zgodność ze specyfikacją URI (niezależność od konkretnych technologii), brak wewnętrznej semantyki (brak cech czy atrybutów zasobu), obsługa metadanych (ważne dla sztucznej inteligencji) i skalowalność (system powinien efektywnie funkcjonować przy wzroście liczby identyfikatorów). Następnie autorzy omawiają kwestię harmonizacji danych na przykładzie osad i jednostek administracyjnych. Obiekty takie można identyfikować diachronicznie lub synchronicznie na podstawie: nazwy własnej, lokalizacji, typu obiektu i relacji mereologicznych (część–całość). Proces ten pozwala przypisać im trwałe PID-y, identyfikujące obiekty z różnych zasobów. W artykule omówiono i poddano ocenie koncepcje wykorzystania PID-ów w różnych projektach. Najczęściej stosowane są identyfikatory wewnętrzne, chociaż nie mają zagwarantowanej stabilności i tylko niektóre z nich spełniają kryteria pozwalające zaklasyfikować je jako PID-y. Pojawiają się również odniesienia do zewnętrznych baz danych, takich jak Geonames lub Wikidata. W ostatniej części pracy autorzy omawiają model wykorzystywany w Zakładzie Atlasu Historycznego Instytutu Historii im. Tadeusza Manteuffla Polskiej Akademii Nauk. Do każdej osady przypisany jest PID pozbawiony semantyki, dlatego wszystkie atrybuty (nazwa, lokalizacja, typ) są zależne od czasu. Odnośnie do identyfikatorów odsyłających do zasobów zewnętrznych wykorzystywane są polskie rejestry nazw i granic, chociaż w przyszłości projekt będzie też korzystał z Wikidata. ■

Grzegorz Myrda, MSc – research assistant in the Department of Historical Atlas (Tadeusz Manteuffel Institute of History, Polish Academy of Sciences). A graduate of the Faculty of Automatic Control, Electronics and Computer Science of the Silesian University of Technology. For over twenty years he has been dealing with issues related to Geographical Information Systems, and recently he has been focusing on the applications of GIS in historical research. Author of books on GIS and publications at the crossways of history and computer science
(grzemy@gmail.com)

Tomasz Panecki, PhD – assistant professor in the Department of Historical Atlas (Tadeusz Manteuffel Institute of History, Polish Academy of Sciences). He obtained a master's degree in history and geography and defended a doctoral thesis in the Faculty of Geography and Regional Studies of the University of Warsaw. He is particularly interested in historical cartography, digital editions of historical maps and cartographic representation of historical data
(tpanecki@uw.edu.pl)

mgr Grzegorz Myrda – asystent w Zakładzie Atlasu Historycznego w Instytucie Historii im. Tadeusza Manteuffla Polskiej Akademii Nauk. Absolwent Wydziału Automatyki, Elektroniki i Informatyki Politechniki Śląskiej. Od ponad dwudziestu lat zajmuje się zagadnieniami związanymi z Systemami Informacji Geograficznej, a ostatnio – zastosowaniem GIS w badaniach historycznych. Autor książek poświęconych GIS oraz publikacji z pogranicza historii i informatyki
(grzemy@gmail.com)

dr Tomasz Panecki – adiunkt w Zakładzie Atlasu Historycznego w Instytucie Historii im. Tadeusza Manteuffla Polskiej Akademii Nauk. Absolwent geografii i historii. Na Wydziale Geografii i Studiów Regionalnych Uniwersytetu Warszawskiego przygotował pracę doktorską pt. *Koncepcja struktury bazy danych historycznych obiektów topograficznych*. Jego zainteresowania badawcze to przede wszystkim kartografia historyczna, cyfrowe edycje dawnych map oraz sposoby historycznej reprezentacji rzeczywistości geograficznej w źródłach pisanych i kartograficznych
(tpanecki@uw.edu.pl)

The research presented in this paper was supported by the *Cartography at the service of political reforms in the times of Stanisław August Poniatowski – a critical elaboration of 'Geographical-statistical description of the parishes in the Kingdom of Poland' and the maps of the palatinates by Karol Perthées* (11H 18 0122 87) grant funded by National Programme for the Development of Humanities.

NARODOWY PROGRAM
ROZWOJU HUMANISTYKI