DOI 10.4467/21995923GP.18.001.9158

Artur Krawczyk[1], Agnieszka Garguła[2]

HARMONIZATION OF POLISH NATURA2000 DATA SETS WITH THE PROTECTED SITES DATA SCHEMA OF INSPIRE DIRECTIVE IN THE ENVIRONMENT OF HUMBOLDT ALIGNMENT EDITOR (HALE)

1 AGH University of Science and Technology; artkraw@agh.edu.pl;
2 AGH University of Science and Technology;

Keywords: Natrure2000, harmonization, HALE, ProtectedSites, thematic databases INSPIRE

Abstract

The main purpose of this work is to test the process of harmonization of the national data schema of natural protected areas in the Natura2000 system in accordance with the European Protected Sites data framework developed under the INSPIRE directive. The analysis of the harmonization process was carried out using the open source Humboldt Alignment Editor (HALA) software and the open source QGIS application. As a result of the analysis, these elements of the Natura2000 data schema have been identified which need to be supplemented as part of the shared data sets by the General Directorate for Environmental Protection. The final stage of the work was to formulate relevent conclusions and summarize the work.

HARMONIZACJA POLSKICH ZBIORÓW DANYCH NATURA2000 ZE SCHEMATEM DANYCH PROTECTED SITES DYREKTYWY INSPIRE W ŚRODOWISKU HUMBOLDT ALIGNMENT EDITOR (HALE)

Słowa kluczowe: Natura2000, harmonizacja, HALE, ProtectedSites, tematyczne bazy danych INSPIRE

Abstrakt

Podstawowym celem niniejszej pracy jest przeprowadzenie testu procesu harmonizacji krajowego schematu danych o terenach chronionych przyrodniczo w systemie Natura2000, zgodnie z europejskim schematem danych Protected Sites opracowanym w ramach dyrektywy INSPARE. Analizę procesu harmonizacji wykonano przy pomocy otwartego oprogramowania Humboldt Alignment Editor (HALE) oraz otwartej aplikacji QGIS. W wyniku wykonanej analizy zidentyfikowano te elementy krajowego schematu danych systemu Natura2000, które wymagają uzupełnienia w ramach udostępnionych zbiorów danych przez Generalną Dyrekcję Ochrony Środowiska. Ostatnim etapem pracy było sformułowanie odpowiednich wniosków i podsumowanie wyników.

1. INTRODUCTION

As a result of the introduction and application of INSPIRE regulations (Directive, 2007) regarding the interoperability of spatial data sets and services as well as the emergence of relevant regulations in our country, the need arose to adapt existing sets to EU regulations. The basic objective of the directive is to support activities related to the planning and implementation of environmental policy, and in its main assumptions, defines the principle of interoperability of national spatial information infrastructures on a regional basis and at the level of the Member States of the European Union. This rule applies to all national datasets, the achievement of which is not an easy task. The component part of the interoperability is the possibility of interactions between different data sets. Such interaction is only possible by using data harmonization and transformation. In the domain of computer science, the basic category of software intended for data transformation is ETL (Extract, Transform, Load) which additionally has to be supplemented with the possibility of permanent recording of relations between components of data schemas. As part of the conducted research, the authors used the specialized open source HALE software (Fichtinger A., et al. 2011).

2. CHARACTERISTICS OF DATA SUBJECTS AND THE HARMONIZATION PROCESS

2.1. The problems of harmonization of Polish thematic data sets

The harmonization of data sets makes it possible to combine different data sets into one coherent data set. A good example is the data set schemas prepared in different national languages, which we want to combine into one data set using English-language names. If the sets are harmonized then we can change the local language of the data set schema to the English-language data set schema. Thanks to this solution, individual countries can maintain their own national data schemas, which can be expanded and modified according to their own needs, while maintaining only a set of elements that have been harmonized with respect to the reference data schema – eg a selected schema based on the INSPIRE directive.

Of course, you can also take standardization activities and by working out one common data set schema, achieve data consistency without the need to harmonize schemas. This approach was preferred by the participants of the “Karkonosze in INSPIRE – Common GIS for nature protection” project (Andrzejewska et all, 2011). The authors have completely rejected the possibility of building two separate but harmonized national data schemas for national parks due to the great difficulty in the schema substitution of the already acquired and saved data. Currently, however, there are tools that allow such operations to be carried out, saving time and limiting the workload to maintain a single data schema for two national parks.

In the last few years, several research works have been undertaken in Poland aimed at harmonizing national thematic data sets. Most of the works, however, were limited to the conceptual or design phase without going into the implementation (technological) phase. The first such work was undertaken in the field of harmonization of the thematic databases GUGIK and PIG (Sikorska-Maykowska M, Olszewski R., 2005). In the field of harmonization of international databases of the two national parks located in the Czech and Polish Giant Mountains, work has been undertaken to harmonize the definitions of concepts themselves in order to achieve a unified database (Andrzejewska et all, 2011). In the field of data sets on building objects, only data set analyzes were carried out without the technical implementation of the proposed solutions (Buśko M., 2017).

Advanced work has been undertaken to adapt the Hydrographic Map of Poland scale data model on a scale of 1:10 000 to the topic: hydrography of the EU INSPIRE directive (Borzuchowski J, Olszar M., 2013). The authors presented a complete approach to the problem of harmonization, which they realized both in a conceptual form and in the form of implementation (with a technological layer). In the technological layer of harmonization of hydrographic data, they used ArcGIS environment with the Data Interoperability overlay to transform the schemas. This overlay is equipped with a graphical interface that greatly facilitates the management of the harmonization process of two different data schemas. The created data harmonization model can then be used to transform any amount of hydrographic data.

In the literature one can encounter a broader interpretation of the concept of harmonizing data sets. This concept may also apply to conceptual models and network services (Kuczyńska J., 2009). While the use of the concept of harmonization with conceptual models is justified, in relation to network services, the concept of orchestration of network services should be used interchangeably rather than the concept of harmonization.

2.2. Characteristics of data schemas

The final data harmonization schema is the ProtectedSite schema, which, according to the INSPIRE documentation, defines a designated or managed area under international, Community and Member States’ legislation in order to achieve specific conservation objectives

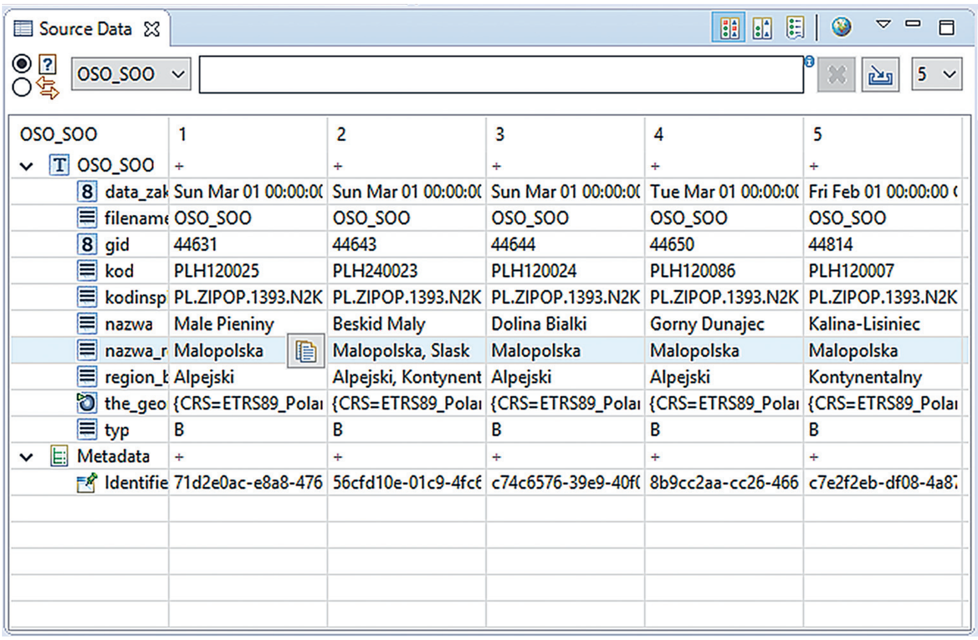
(ProtectedSite, 2014). Each area is a protected area if its boundaries has been defined in a legal act that is formally valid in the territory of a given EU Member State. Establishing a protected area is usually backed up by legislation, so it is important in decisions regarding land use change and spatial planning. In each place it is usually chosen as a representative an example of a wider resource and chosen by a formal approach based on a given criterion. A protected area can be the adjacent land / sea range or a set of separate areas that together form one formal protected area.

The output schema is a collection of Natura 2000 areas data made available by GDOŚ in the form of a WFS service. It is worth recalling that the first Natura 2000 areas were established in Poland in 2003, long before the implementation of the INSPIRE Directive. During their implementation and establishment, spatial databases were created containing basic information about the boundaries of these areas. In Poland, the Natura 2000 program covered 849 habitat areas. Spatial data sets were created in GIS (Geography Information Systems) according to national rules using attribute names in Polish. The data schema created then was also made in Polish. For this reason, the national schema should be harmonized with the schema compliant with INSPIRE for protected areas.

3. THE COURSE OF THE HARMONIZATION PROCESS

We begin the harmonization process by downloading sample data Using QGIS, we collect the SPAs and SACs data sets and combine them into one layer. These two sets of data are spatially interdependent by using a type identifier that identifies the type of polygon. Each polygon is classified either in the areas of SPAs type A or in the SACs areas as type B or is a common part of these areas and is then marked as type C. To harmonize we need data from one voivodship, Lesser Poland Voivodship was selected for the study. Therefore, we extract from the obtained data set only those areas that are in the selected voivodship. After importing data from WFS to the QGIS program, we export it to the * .shp file.

With the prepared data in the format of shp we can proceed to further work. The latest version of HALE allows you to use files in the format shp, gml, wfs, both to read the data itself and to read the schema of this data. Earlier versions of this program needed special preparation of the xsd file for the data schema and the gml file for the source data. Currently, however, you can use the same shp file to download only the data and the schema of this data. The program will independently



OSO_SOO	1	2	3	4	5
OSO_SOO	+	+	+	+	+
data_zak	Sun Mar 01 00:00:0(	Sun Mar 01 00:00:0(	Sun Mar 01 00:00:0(	Tue Mar 01 00:00:0(	Fri Feb 01 00:00:00 (
filename	OSO_SOO	OSO_SOO	OSO_SOO	OSO_SOO	OSO_SOO
gid	44631	44643	44644	44650	44814
kod	PLH120025	PLH240023	PLH120024	PLH120086	PLH120007
kodinsp	PL.ZIPOP.1393.N2K	PL.ZIPOP.1393.N2K	PL.ZIPOP.1393.N2K	PL.ZIPOP.1393.N2K	PL.ZIPOP.1393.N2K
nazwa	Male Pieniny	Beskid Maly	Dolina Bialki	Gorny Dunajec	Kalina-Lisiniec
nazwa_r	Malopolska	Malopolska, Slask	Malopolska	Malopolska	Malopolska
region_b	Alpejski	Alpejski, Kontynent	Alpejski	Alpejski	Kontynentalny
the_geo	{CRS=ETRS89_Polai	{CRS=ETRS89_Polai	{CRS=ETRS89_Polai	{CRS=ETRS89_Polai	{CRS=ETRS89_Polai
typ	B	B	B	B	B
Metadata	+	+	+	+	+
Identifie	71d2e0ac-e8a8-476	56cfd10e-01c9-4fc6	c74c6576-39e9-40f(	8b9cc2aa-cc26-466	c7e2f2eb-df08-4a87

Fig. 1. HALE program tab – viewing imported data
Rys. 1. Zakładka programu HALE – widok zaimportowanych danych

generate a data schema from it and load the data of all objects with attributes and their values. Figure 1 shows the schema of the national Natura 2000 data set after importing into the HALE program.

In the program tab in the first column, you can view the attributes of the national data schema. The following columns contain attribute values for individual objects. The number of columns in this tab is equal to the number of geometric objects imported from the data file.

The next stage of harmonization is importing the target data schema. In our research, we will use the ProtectedSites.xsd file, which is an INSPIRE data schema for protected areas. This schema was created around 2014 and is available on the INSPIRE websites.

We begin the harmonization with the implementation of relations binding the elements of the source schema with the elements of the target schema. There are several relationships between elements. The construction of the schematic projection begins with the construction of the relationship using the function retype. This is the simplest type of relationship that expresses the source and target data as semantically equal. We use this relationship when we want to combine one type of element of a given schema with a specific element of the target schema. In this case, we combine elements of the SPAs_SACs schema (file containing our input data) with the ProtectedSite.xsd schema. This allows you to assign one type of source schema to the second target data schema. The retype command applies to the main element of the schema (root). The retype command assumes that "mapping", that is projecting all sub-elements will be implemented automatically, provided identical elements names occur in two harmonized schemas. Due to language differences between schemas, the name of any element from the national schema does not appear in the target schema, hence this function could not automatically combine any attributes.

Therefore, we have to approach each element of the schema individually and perform a manual connection (matching) of the appropriate type. By using the rename connection, we can copy the values of any attributes from the output schema to the target schema. In this way, we change the attributes:

Data_zakla – legalFoundationDate

Data_zakla – under this attribute there are dates informing us when the area was qualified as an SPAs or SACs area.

legalFoundationDate – the documentation describes this as the date when the protected areas were created in accordance with the law. It is the day of creation of the habitat protection facility, not the date of entering into the IT system. In the case of Natura 2000 sites, the protected area can go through several decision stages as part of the admission to protected areas of different categories). The next attributes to which the rename function was applied are:

kod – id

kod – assigned to each area when it is qualified for the Natura 2000 program in accordance with the assumption:

"kod: PLB and number – Special Protection Areas established formally in Poland until the end of October 2008 by ordinances of the Minister of the Environment,

kod: PLH and number – Special Areas of Conservation approved by the end of December 2008 by the European Commission – the so-called areas of Community importance,

kod: PLH and the number separated by the low dash – new Special Areas of Conservation proposed formally until the end of October 2009 by the Government of the Republic of Poland (sent as a Polish proposal to the European Commission)

kod: PLH and the number separated by the low dash – additional Special Areas of Conservation consulted in the summer of 2009 by the Ministry of the Environment, and not included by the end of October 2009 into lists of areas formally proposed by the Government of the Republic of Poland,

kod: PLTMP and number – additional Special Areas of Conservation proposed by natural non-governmental organizations, and by the end of October 2009 not approved by the authorities of the Republic of Poland."

ID – The object identifier. (ID of the object)

kodinspire – InspireID.Identifier.localld

kodinspire – The id code assigned to each object from the list of INSPIRE themes.

InspireID.Identifier.localld inspireID – The object identifier from the INSPIRE directive. The object's external identifier is a unique identifier published by the responsible authority that can be used by external applications to refer to a spatial object. It is the identifier of the spatial object and not the address or inventory number of this phenomenon in the real world.

nazwa – siteName.GeographicalName.spelling.SpellingOfName.text

The last parameter to which we used the rename function is the Name attribute. It contains information about the exact and full name of a protected area.

The next step is to specify the geometry of the objects. To do this, we use the Network Expansion function. We select GeometricComplex as targets because the objects in the study area do not have regular shapes. Documentation of Natura2000 areas describes geometry as the definition of the boundary of the protected area. In the past, these boundaries have been defined in many ways, including geodetic, digitized cartographic materials or visual identification in the field by reference to physiographic or anthropogenic traits. This boundary should be clearly defined by a legal document that creates protected areas based on the national reference databases of the Spatial Information Infrastructure.

The Classification function is more complicated, in which the sources are to be mapped to the properties specified in the target schema. This operation was performed using the *typ* attribute (having the values A, B, C), thanks to which it was possible to classify three parameters in the target schema.

typ – siteProtectionClassification

siteProtectionClassification – Classification of the protected area based on the protection purpose. An area can have more than one classification. In the Natura-2000 program, we can distinguish the division into:

- natureConservation
- archeological
- cultural
- landscape
- geological

In the case of our data, that is: protection of habitats and protection of birds, each area, regardless of the type it has, is classified as a site counted for nature conservation. In this classification, the only appropriate attribute from the list is natureConservation.

typ – description

description – Description of a given area, it has been divided according to the following guidelines. According to the rules of qualification, the SPAs and SACs areas were divided into A, B, C types and were qualified accordingly to:

A – SPAs Special Protection Areas,
B – SACs Special Areas of Conservation
C – Common area of SPAs and SACs.

typ – siteDesignation.Designationtype.designation

designation – Real area designation. The values of the designation attribute that can be taken for the Natura 2000 classification are:

specialAreaOfConservation: The area of the protected site is designated as a Special Area of Conservation (SAC) under Natura 2000.

specialProtectionArea: The area of the protected area is designated as a Special Protection Area (SPA) under the Natura 2000 program.

siteOfCommunityImportance: A protected area proposed as an area of Community Importance under Natura 2000.

proposedSiteOfCommunityImportance: A protected area proposed as a Special Protection Area (SPA) under the Natura 2000 program.

As the last one, the assign function was used, in which the target value is assigned by the user creating a given harmonization on the basis of general knowledge or according to the information precisely defined by the target element to be included in the given place. This function allows you to add values to those schema elements that are not in the input schema but must be in the target schema. In this situation, it is necessary to add the missing values to the mandatory elements of the ProtectedSites schema. So the identifier inspireID.Identifier.namespace was assigned:

assign inspireID.Identifier.namespace: PL.ZIPOP.1393.N2K.PLB140004

Another attribute to describe was designationScheme, that is the identifier of the European conservation program, which is identified by a special program code. The code values that the designationScheme element can adopt are determined by the documentation regarding the harmonization of the protected areas of INSPIRE and these are:

Nature2000: The protected area is marked both under the Habitat Directive (92/43 / EEC) and the Birds Directive (79/409 / EEC).

emeraldNetwork: The protected area has a designation within the Emerald Network. (The Emerald Network is an ecological network that protects wild fauna and flora and their natural habitats in Europe).

ramsar: The protected area has a designation under the Ramsar Convention. The Ramsar Convention provides a framework for national activities and international cooperation in the application of conservation of wetlands and their resources.

UNESCOWorldHeritage: The protected area has a designation within the UNESCO World Heritage

IUCN: The area of the protected site is classified using the International Union of Nature Conservation classification system.

UNESCOManAndBiosphereProgramme: The protected area is marked as part of the UNESCO Man and Biosphere Programme.

nationalMonumentsRecord: The area of the protected facility is classified using the National Monuments Record system".

In the present case, the Natura2000 value should be added:

assign **siteDesignation.DesignationType.designationScheme: Nature2000**

Next, the value for percentageUnderDesignation is determined, which is the percentage of the place marked with the IUCN category. This is the part of the Natura2000 area that at the same time belongs to any nature conservation category by the IUCN organization of the International Union for Conservation of Nature. Unfortunately, no authority of the Polish state belongs to this organization and this qualification is not applied

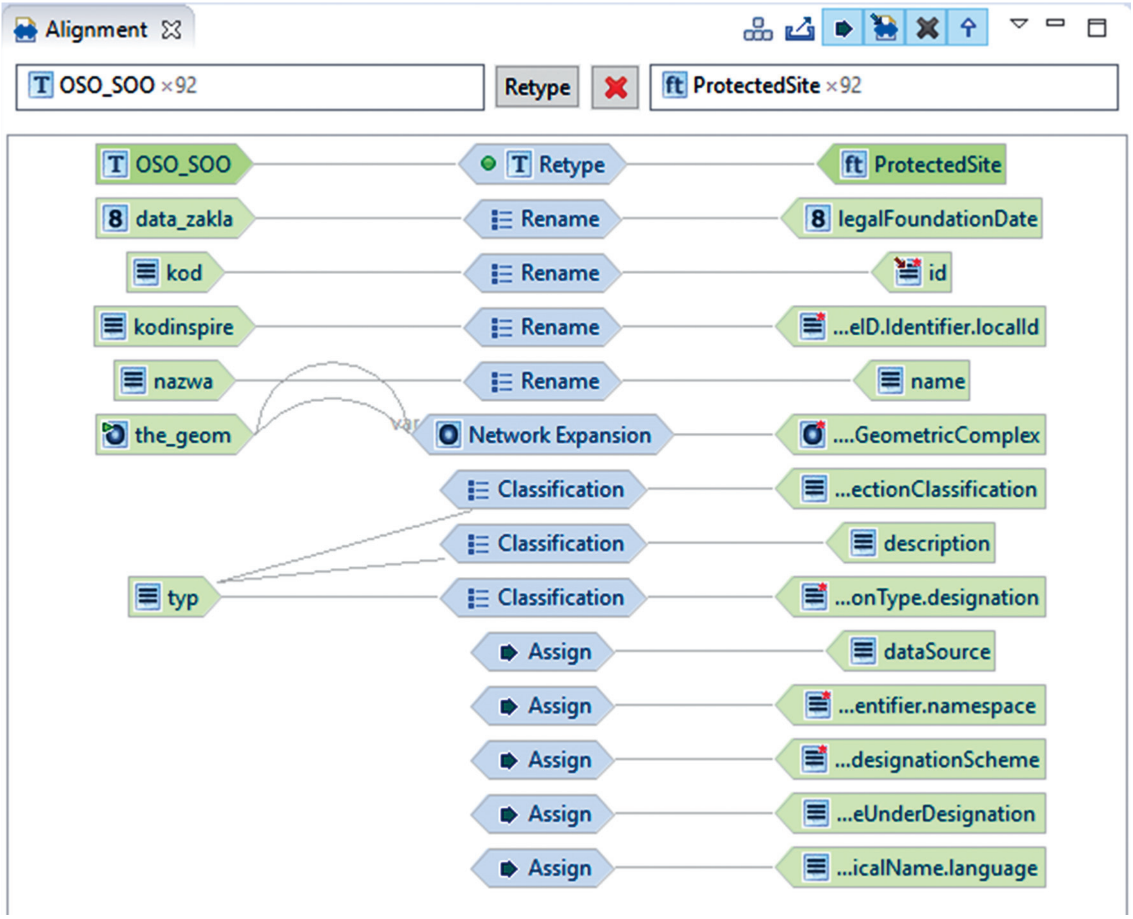


Fig. 2. Graphical interface of the HALE program presenting relations between data schemas
Rys. 2. Graficzny Interfejs programu HALE prezentujący relacje pomiędzy schematami danych

in Poland. Therefore, if the value is not specified for this attribute, it is assumed that it is 100% and this is also specified in this harmonization model.

assign **siteDesignation.DesignationType.**
precentageUnderDesignation 100_percent

The last value on which the described transformation was performed was to determine the language in which the names of protected areas are given.

assign **siteName.GeographicalName.language: PL**

After the mapping process completed in this way, we export the result of harmonization with the "transformed data" option to the gml file, which we can open in the QGIS program and compare with the file containing the same data but not being harmonized.

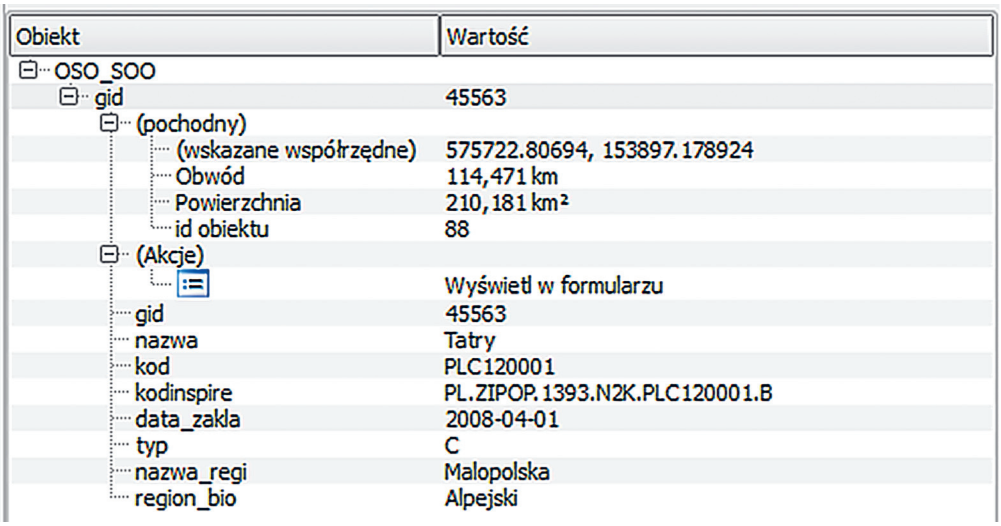
4. COMPARISON OF DATA FILES BEFORE AND AFTER TRANSFORMATION

After completing the configuration of the harmonization of the data set schemas, the geometric data was transformed into a new gml file.The resultant after harmonization, the target file opened in the QGIS program does not graphically differ from the output file in terms of geometry. It means that none of the objects has been modified, all objects that were before the harmonization are in the target file.

The key issue is how the data schema changed in these files. The following are two drawings of data and attribute schemas: No. 3 input schema and No. 4 target schema, which show selected names of the schema attributes before and after the harmonization. It can be seen that in the geometry range the first file contains identical attributes and identical values that appear in the file after harmonization. The main differences between the files can be seen in the names of the text attributes of the target data set, such as: gml id, description, localId, namespace, legalFoundationDate, designationScheme, designation, percentageUnderDesignation, language, text, siteProtectionClasification. The attributes whose names and content match the specific values in the regulation to the EU directive. They can be used without any problems to, for example, combine them with also standardized in accordance with the Directive data from other regions or to create international database files.

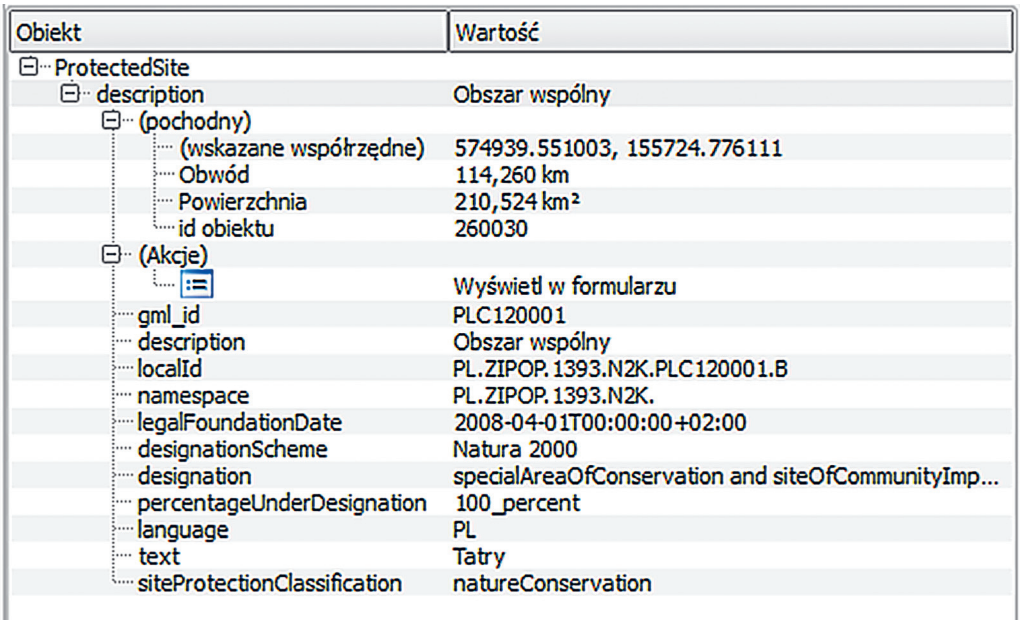
5. CONCLUSIONS

As a result of the analyses carried out, it can be assumed that the scope of the data set schema used by the GDOŚ Directorate contains many data consistent with the conceptual meaning of the ProtectedSites data set schema. Unfortunately, in relation to the INSPIRE schema, a few data are missing and they need to be supplemented as part of the harmonization. Five el-



Obiekt	Wartość
OSO_SOO	
gid	45563
(pochodny)	
(wskazane współrzędne)	575722.80694, 153897.178924
Obwód	114,471 km
Powierzchnia	210,181 km^2
id obiektu	88
(Akcje)	
	Wyświetl w formularzu
gid	45563
nazwa	Tatry
kod	PLC120001
kodinspire	PL.ZIPOP.1393.N2K.PLC120001.B
data_zakla	2008-04-01
typ	C
nazwa_regi	Malopolska
region_bio	Alpejski

Fig. 3. A fragment of the file's schema structure before harmonization
Rys. 3. Fragment struktury schematu pliku danych przed harmonizacją



Obiekt	Wartość
ProtectedSite	
description	Obszar wspólny
(pochodny)	
(wskazane współrzędne)	574939.551003, 155724.776111
Obwód	114,260 km
Powierzchnia	210,524 km²
id obiektu	260030
(Akcje)	
☰	Wyświetl w formularzu
gml_id	PLC120001
description	Obszar wspólny
localId	PL.ZIPOP.1393.N2K.PLC120001.B
namespace	PL.ZIPOP.1393.N2K.
legalFoundationDate	2008-04-01T00:00:00+02:00
designationScheme	Natura 2000
designation	specialAreaOfConservation and siteOfCommunityImp...
percentageUnderDesignation	100_percent
language	PL
text	Tatry
siteProtectionClassification	natureConservation

Fig. 4. A fragment of the structure of the schema file after harmonization
Rys. 4. Fragment struktury schematu pliku danych po harmonizacji

ements of the INSPIRE schema have been identified for which there is a lack of data in the GDEP schema. Of course, these data do not have to be supplemented in the GDEP schema, because they can be parameters of this harmonization. Certainly there is no need to implement an attribute in the national schema under the name siteName.GeographicalName.language: PL, only when changing the language of the data set schema, then this information must be included. Similarly, the designationScheme element designating the name of the target schema can be completed during harmonization. However, the DataSource attribute should be included in the national data schema. It allows to identify the quality of the geometry of the Natura2000 area defined by the scale of the map (spatial data set) used for the reference location of the border in the field. If you used 10,000 scale maps then the area boundary cannot be used in larger scales (e.g. 1: 2000). Finding and determining the value of this attribute based on historical data is not always possible.

On the other hand, thanks to harmonization, there is no obstacle to maintaining a stock of national data in a data schema containing attributes defined in Polish. The data harmonization process itself – carried out and preserved – can be referred to repeatedly, and the data can be harmonized online. Currently, the HALE software has a geoserver plugin that allows downloading data taking into account their harmonization with the selected data schema.

In the summary of the literature research carried out, we can emphasize a very small number of publications in Poland, regarding harmonization in which harmonization of an exemplary physical data set was performed. This may result from ignorance of available software dedicated to this purpose. HALE software is free and open source, which allows it to be used for commercial and teaching purposes without any problems.

Finally, the term “spatial data interoperability” in the context of harmonization should also be understood as access to spatial data sets through network services that allow interoperability also through data transformation, including data schema transformation, through the publication services used in spatial data infrastructure defined by the INSPIRE Directive.

ACKNOWLEDGMENTS

The present article is a form of publications of results of studies made during the realization of Agnieszka’s Garguła master’s thesis with the help of the promoter of Artur Krawczyk (Garguła A. 2015).

BIBLIOGRAPHY

Andrzejewska M., Jała Z., Rusztecka M., 2011, Harmonisation of spatial data concerning a transboundary protected area on the example of two national parks: The polish karkonoski park narodowy and its czech counterpart, The Krkonoský Národní Park, within the framework of the project entitled Karkonosze in INSPIRE – Common GIS for nature protection. Roczniki Geomatyki, tom IX, z. 4(48).

Buśko M., 2017, Legal and technical aspects of harmonization of databases of buildings, Geoinformatica Polonica, vol. 16, p. 101–113.

Borzuchowski J, Olszar M., 2013, Data harmonization in the context of inspire directive. the hydrographic map of Poland at the scale of 1:10 000, Roczniki Geomatyki, Tom XI, z. 3(60).

Dyrektywa 2007/2/WE Parlamentu Europejskiego i Rady z dnia 14 marca 2007 r. ustanawiająca Infrastrukturę Informacji Przestrzennej we Wspólnocie Europejskiej – INSPIRE, 2007.

Fichtinger A., Rix J., Schäffler U., Michi I., Gone M., Reitz T., 2011, “Data Harmonisation Put into Practice by the HUMBOLDT Project”, International Journal of Spatial Data Infrastructures Research, Vol. 6, 234–260.

Garguła A., 2015, „Proces harmonizacji danych o terenach chronionych przyrodniczo zgodnie z dyrektywą INSPIRE”, Master Thesis, unpublished, AGH-UST WGGIŚ, Kraków.

Hintz D., 2012, *Data Harmonization Principles and Development Approaches as Applied to INSPIRE SDIs* Safe Software.

Kuczyńska J., 2009, Aspects of harmonization and integration of reference data in the process of developing SDI with application of MDA and SOA strategy, Roczniki Geomatyki, Tom VII, z. 4(34).

ProtectedSite “D2.8.I.9 Data Specification on Protected Sites – Technical Guidelines” INSPIRE Thematic Working Group Protected Sites Ispra 2014.

Sikorska-Maykowska M., Olszewski R., 2005, The concept of harmonization of thematic databases of the head office of geodesy and cartography and polish geological institute based on a uniform reference data system, Roczniki Geomatyki, Tom III, z. 2.