support

# Materiały szkoleniowe
# Transformacja danych w Hale Studio
# (poziom średniozaawansowany)

MINISTERSTWO
KLIMATU

Sfinansowano ze środków
Narodowego Funduszu
Ochrony Środowiska
i Gospodarki Wodnej

*Materiały szkoleniowe zostały przygotowane w oparciu o źródła umieszczone na portalu https://ekoportal.gov.pl/*

*https://ekoportal.gov.pl/fileadmin/user_upload/Materialy_szkoleniowe_HALE_Studio.pptx_w2.pdf*

*https://ekoportal.gov.pl/fileadmin/user_upload/Zeszyt_cwiczen_-_Hale_Studio_w2.pdf*

# Omówienie zasad przejścia ze struktury źródłowej do struktury docelowej, w szczególności w kontekście harmonizacji danych przestrzennych na potrzeby inicjatywy INSPIRE

## Cele budowy INSPIRE

**Cel główny:** stworzenie zdolności współdziałania (**interoperacyjności**) w zakresie informacji przestrzennej w Europie.

Cele szczegółowe:

- Wsparcie polityk Wspólnoty odniesionych do środowiska
- Zniesienie barier w dostępie do informacji przestrzennej – open data
- Wielokrotne wykorzystanie raz pozyskanej informacji
- Decentralizacja
- Budowa społeczeństwa informacyjnego Funkcjonalność
- Umożliwienie łączenia w jednolity sposób danych przestrzennych pochodzących z różnych źródeł we Wspólnocie i wspólne korzystanie z nich przez wielu użytkowników i wiele aplikacji.

Sposobem na realizację i osiągnięcie powyższego celu jest **harmonizacja** – działania o charakterze technicznym, organizacyjnym i prawnym, mające na celu doprowadzenie do wzajemnej spójności zbiorów danych przestrzennych i usług geoinformacyjnych.

## Kluczowe pojęcia

**Dyrektywa INSPIRE**
Dyrektywa z dnia 14 marca 2007 ustanawiająca jednolitą infrastrukturę informacji przestrzennej dla Unii Europejskiej i państw EFTA.

**Specyfikacja danych**
Dokument ustalający wymagania techniczne, które powinien spełniać wyrób, proces lub usługa.

**Interoperacyjność**
W kontekście infrastruktury informacji przestrzennej oznacza możliwość łączenia zbiorów danych przestrzennych bez powtarzalnej interwencji manualnej. Interoperacyjność oznacza zdolność różnych elementów funkcjonalnych systemów informatycznych do komunikacji, uruchamiania programów lub przesyłania danych pomiędzy nimi w sposób nie wymagający od ich od użytkownika żadnej wiedzy lub wymagający od niego wiedzy minimalnej na temat unikalnych właściwości tych elementów. Zgodnie z dyrektywą INSPIRE to „możliwość łączenia zbiorów danych przestrzennych oraz interakcji usług danych przestrzennych bez

powtarzalnej interwencji manualnej, w taki sposób, aby wynik był spójny, a wartość dodana zbiorów i usług danych przestrzennych została zwiększona"

**Integracja danych**
Działania mające na celu umożliwienie jednoczesnego łączenia wielu zbiorów, usunięcie z nich powtórzeń i elementów zbędnych.

**Harmonizacja danych**
Proces mający na celu zapewnienie dostępu do danych przestrzennych umożliwiającego łączenie ich w sposób spójny z innymi danymi.

Zgodnie z ustawą IIP niezbędnym krokiem w procesie harmonizacji jest opracowanie danych referencyjnych / i ich szerokie i nieodpłatne wykorzystanie.
Zgodnie z Ustawą Prawo geodezyjne i kartograficzne: „harmonizacja zbiorów danych – rozumie się przez to działania o charakterze prawnym, technicznym i organizacyjnym, mające na celu doprowadzenie do wzajemnej spójności tych zbiorów oraz ich przystosowanie do wspólnego i łącznego wykorzystywania”
Zgodnie z dyrektywą INSPIRE „na szczeblu krajowym i wspólnotowym podejmuje się wiele inicjatyw mających na celu gromadzenie, harmonizację lub organizację rozpowszechniania lub wykorzystania informacji przestrzennych” (pkt 11 preambuły).

**Mapowanie**
Proces przyporządkowania obiektów i ich atrybutów ze zbioru wejściowego do schematu danych wyjściowych.

**Specyfikacje techniczne**
Specyfikacje techniczne dzielą się na dwie podstawowe grupy:

1. Dokumenty ramowe, dotyczące ogólnego schematu budowy danych INSPIRE oraz szczegółowo objaśniające sposób konstrukcji klas i atrybutów podstawowych.
2. Szczegółowe specyfikacje dla specjalistycznych tematów zebranych w załączniki.

## Oprogramowanie HALE Studio

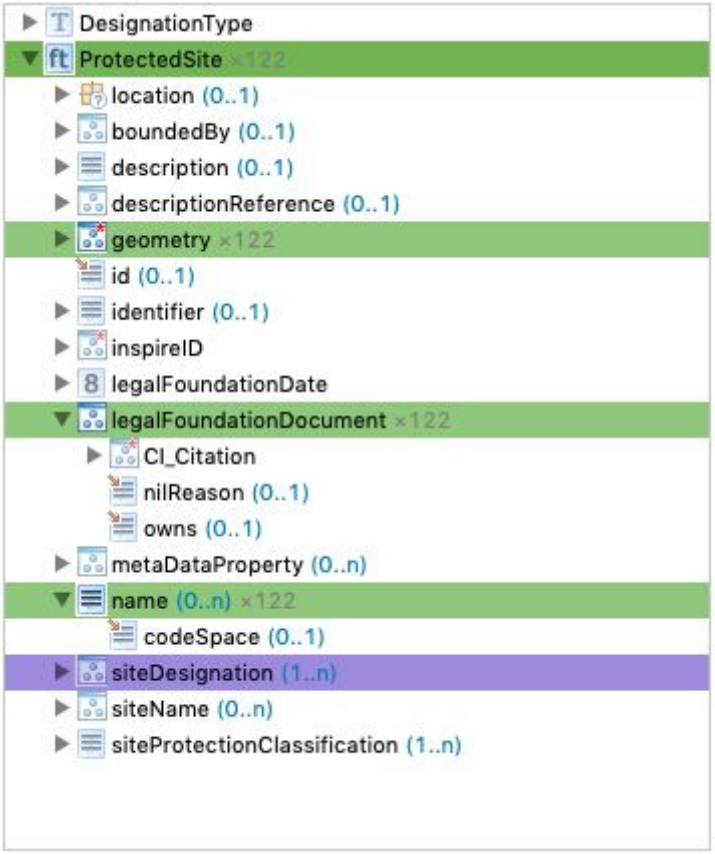
- Oprogramowanie utworzone w ramach projektu HUMBOLDT - pierwotna nazwa **H**UMBOLDT **Al**ignment **E**ditor - w celu ułatwienia przekształcania danych do schematów INSPIRE
- Napisane w języku Java, możliwe do uruchomienia we wszystkich głównych systemach operacyjnych (Mac, Linux, Windows)
- Aktualnie rozwijane przez firmę WeTransform (http://wetransform.to)
- Kod jest dostępny na licencji open source - GNU Lesser General Public License (LGPL) v3.0
- Dostępne jest wsparcie komercyjne od WeTransform
- Pobieranie wersji instalacyjnej - https://www.wetransform.to/downloads/, kodu źródłowego - https://github.com/halestudio/hale
- Dokumentacja http://help.halestudio.org/latest/index.jsp
- Szkolenie jest przygotowano dla wersji 4.0.

## Schemat w HALE

Schemat (**Schema**) jest opisem struktury danych. Wyróżniane są 2 typy schematów:

- Logical Schema - schemat konkretnego zbioru danych (pliku, dokumentu XML, bazy danych)
- Conceptual Schema - model koncepcyjny, opisuje dane bez wskazania konkretnej implementacji (np. model UML)

Do przeprowadzenia transformacji (harmonizacji) wymagane jest posiadanie Logical Schema.




W procesie transformacji w HALE Studio schemat może być:

- zaimportowany z pliku definicji (np. XSD)
- zaimportowany z gotowych szablonów - **presets** - HALE posiada wbudowane gotowe definicje dla schematów aplikacyjnych INSPIRE
- odtworzony na podstawie istniejących danych (np. pliku SHP, bazy danych)

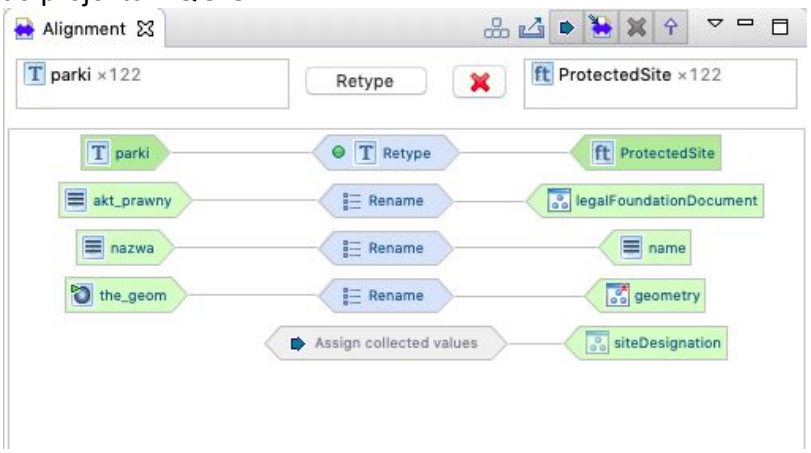
Dla projektu HALE musi zostać zdefiniowany schemat dla danych wejściowych (Source Schema) i wynikowych (Target Schema) by można było dokonać transformacji.

Parametry transformacji są określane jako **Alignment** i definiują zestaw przekształceń pomiędzy schematem źródłowym a wynikowym.

Pojedynczy **Alignment** składa się ze zbioru obiektów typu **Mapping cell**, które definiują przejście pomiędzy typami lub atrybutami wejściowymi i wynikowymi.

Zestaw parametrów transformacji oraz wskaźników do danych wejściowych i wynikowych

nazywa się **Alignment project** i może być zapisany na dysku w formie pliku - analogicznie do projektów QGIS.




**Mapping cell** składa się z bytu wejściowego, bytu wyjściowego oraz funkcji przekształcenia.

**Przykłady**:

Funkcja **Retype** przenosi typ wejściowy "parki" na typ wyjściowy "ProtectedSite"
(funkcja **Retype** przeznaczona jest do mapowania **typów danych**)

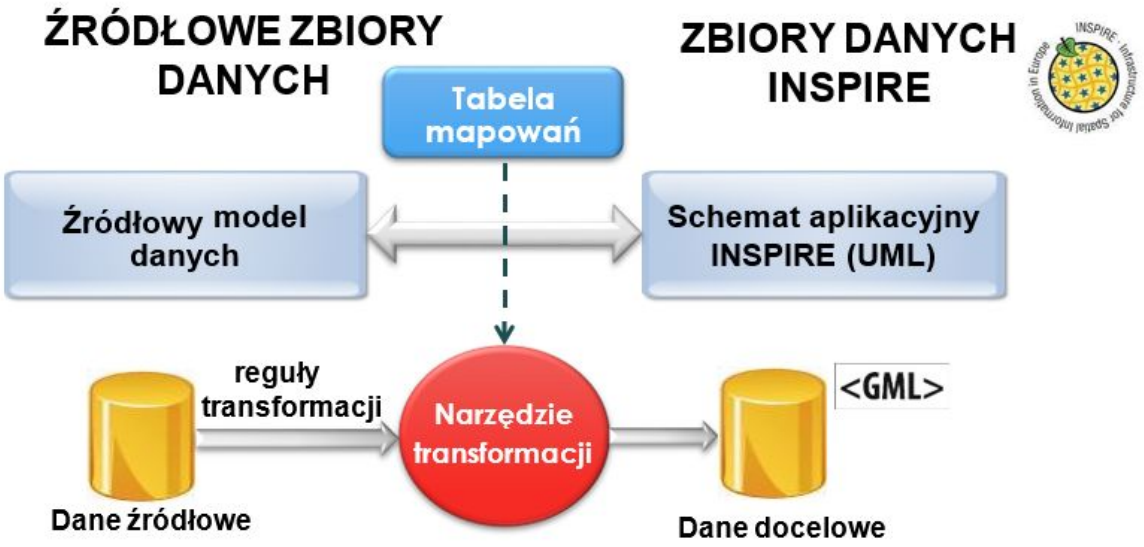

Funkcja **Rename** przenosi wartość atrybutu "akt_prawny" na atrybut "legalFoundationDocument"
(funkcja **Rename** jest przeznaczona do mapowania **atrybutów**)


Proces harmonizacji danych w HALE składa się z następujących etapów:

- Import schematu danych źródłowych
- Import schematu danych wynikowych
- Import próbki danych źródłowych
- Identyfikacja typów danych wynikowych, które mogą być utworzone z danych źródłowych
- Stworzenie relacji pomiędzy typami wejściowymi a wynikowymi
- Identyfikacja i konfiguracja relacji pomiędzy wejściowymi a wynikowymi atrybutami
- Weryfikacja poprawności transformacji
- Eksport danych wynikowych

# Projektowanie, implementowanie procesów ETL

Uproszczony schemat harmonizacji danych i miejsca ETL w tym procesie




Etapy Harmonizacji

1. Identyfikacja zbiorów danych przestrzennych, która polega na wyborze zasobów danych przestrzennych, które powinny podlegać procesowi harmonizacji zgodnie z wytycznymi INSPIRE

2. Przekształcenie danych. Etap który polega na dostosowaniu wybranych danych źródłowych do wymagań wynikających z dyrektywy INSPIRE. Obejmuje zadania mające na celu przemodelowanie danych ze struktur stosowanych w schematach implementacyjnych danych źródłowych do struktur wymaganych przez model wymiany danych zgodnie ze specyfikacją. Składa się z następujących etapów:
a) przygotowanie danych do transformacji, obejmujące stworzenie tabeli mapowań pomiędzy schematem źródłowym a schematem docelowym INSPIRE,
b) transformacja danych ze schematu źródłowego do schematu docelowego INSPIRE,
c) transformacja danych z lokalnego układu współrzędnych do układu wspólnego dla danych publikowanych w INSPIRE,
d) pozyskanie dodatkowych informacji, które nie są dostępne w zbiorze źródłowym, a konieczne są do spełnienia wymogów narzuconych w specyfikacji dla konkretnego tematu

3. Publikacja danych w usługach sieciowych

## Co to są dane?

Zacznijmy od tego, czym w ogóle są dane. Słownik języka polskiego odwołuje się do dwóch znaczeń.

Dane to
1. «fakty, liczby, na których można się oprzeć w wywodach»

2. «informacje przetwarzane przez komputer»

W rządowym Programie otwierania danych publicznych (dalej: Program) znajdziecie z kolei definicję danych publicznych, czyli takich, które są w posiadaniu urzędów, niezależnie od tego, kto je wytworzył:

«Dane publiczne – liczby i pojedyncze wydarzenia lub obiekty na możliwie najniższym poziomie agregacji, które nie zostały poddane przez administrację publiczną przetworzeniu do postaci raportów, wykresów itp. oraz nie został im nadany odpowiedni kontekst lub interpretacja. »

## Co to znaczy, że dane są otwarte?

Dane, aby efektywnie służyły do realizacji wskazanych wyżej celów, powinny być:

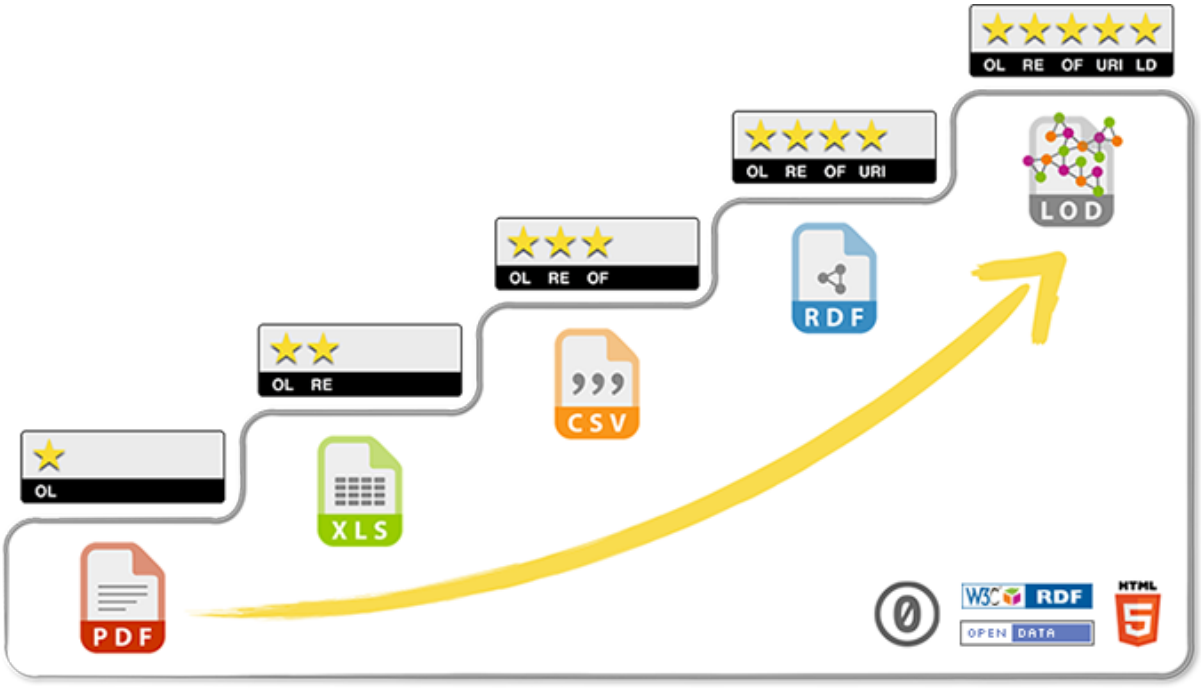
1. przygotowane w sposób przyjazny dla użytkownika/ użytkowniczki – kluczowy jest komunikatywny, zrozumiały dla odbiorcy język;

2. spełniające wszystkie filary otwartości, w których jest odniesienie do kwestii prawnych, udostępnione w formacie przeznaczonym do odczytu maszynowego i w otwartych formatach do ponownego wykorzystywania.

Takie dane powinny być:

1. dostępne – udostępnione bez żadnych ograniczeń szerokiemu gronu użytkowników/użytkowniczek (w szczególności: obywatele, firmy, uczelnie, instytucje) do dowolnych celów;
2. upublicznione w wersji źródłowej – dostępne w oryginalnej i niezmienionej formie, nie w postaci analiz, podsumowań, skrótów czy streszczeń, tak aby możliwe było np. łączenie danych z różnych źródeł;
3. kompletne – udostępnione w całości;
4. aktualne – udostępnione na tyle szybko, aby zachować wartość tych danych;
5. odczytywalne maszynowo – udostępnione w formatach przeznaczonych do odczytu maszynowego i formacie otwartym. Przykładem takich formatów jest CSV, XML, arkusz kalkulacyjny. Zazwyczaj trudne do maszynowego odczytu są formaty PDF, HTML czy pliki tekstowe, które stają się użyteczne do ponownego wykorzystywania dopiero po przełożeniu ich na jeden z formatów otwartych;
6. udostępnione niedyskryminująco – dostępne dla każdego bez konieczności rejestracji, weryfikacji tożsamości poprzez podawanie hasła, loginu czy podpisywania jakichkolwiek umów;
7. dostępne bez ograniczeń prawnych – dane nie są przedmiotem praw autorskich, patentów, znaków towarowych lub tajemnicy handlowej i mogą być wykorzystywane w dowolnych celach bez konieczności ubiegania się o jakąkolwiek zgodę na ich używanie;
8. niezastrzeżone – dostępne w formacie powszechnie stosowanym, który nie jest kontrolowany przez żaden podmiot.

## 5 poziomów dostępności




*Rysunek 1: Pięć poziomów otwartości danych, źródło: https://5stardata.info/*

gdzie poszczególne gwiazdki oznaczają:

* Udostępnienie danych w sieci Web (w dowolnym formacie) na warunkach otwartej licencji

** Udostępnienie danych w formie ustrukturyzowanej (np. arkusz kalkulacyjny zamiast zeskanowanego obrazu tabeli)

*** Używanie formatów otwartych (np. CSV zamiast arkusza kalkulacyjnego)

**** Używanie URI do oznaczania zasobów, aby możliwe było ich wyszukiwanie

***** Łączenie danych, aby zapewnić kontekst

Przykłady poziomów dostępności

* Udostępnienie danych w formie tabeli w pliku PDF

TABL. 4. **NAJDŁUŻSZE RZEKI**
THE LONGEST RIVERS

| Rzeki Rivers | Recypient[a] Recipient[a] | Długość w km Length in km | | Powierzchnia dorzecza w km² Drainage basin area in km² | |
|---|---|---|---|---|---|
| | | ogółem total | w tym w Polsce of which in Poland | ogółem total | w tym w Polsce of which in Poland |
| Wisła | Morze Bałtyckie Baltic Sea | 1022 | 1022 | 193960[b] | 168868[b] |
| Odra | Morze Bałtyckie Baltic Sea | 840 | 726 | 119074 | 106043 |
| Warta | Odra | 795 | 795 | 54520 | 54520 |
| Bug | Narew | 774 | 590 | 38712 | 19239 |
| Narew | Wisła | 499 | 443 | 74527 | 53846 |
| San | Wisła | 458 | 457 | 16877 | 14426 |
| Noteć | Warta | 391 | 391 | 17302 | 17302 |
| Wieprz | Wisła | 349 | 349 | 10497 | 10497 |
| Pilica | Wisła | 333 | 333 | 9258 | 9258 |
| Bóbr | Odra | 279 | 276 | 5874 | 5830 |

*Rysunek 2:Dane na pierwszym poziomie, źródło: Mały Rocznik Statystyczny Polski 2018, Główny Urząd Statystyczny*

** Udostępnienie danych w formie ustrukturyzowanej (np. arkusz kalkulacyjny zamiast zeskanowanego obrazu tabeli)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Nazwa rzeki | Recypient | Długość w km ogółem | Długość w km w Polsce | Powierzchnia dorzecza w km2 ogółem | Powierzchnia dorzecza w km2 w Polsce |
| 2 | Wisła | Morze Bałtyckie | 1022 | 1022 | 193960 | 168868 |
| 3 | Odra | Morze Bałtyckie | 840 | 726 | 119074 | 106043 |
| 4 | Warta | Odra | 795 | 795 | 54520 | 54520 |
| 5 | Bug | Narew | 774 | 590 | 38712 | 19239 |
| 6 | Narew | Wisła | 499 | 443 | 74527 | 53846 |
| 7 | San | Wisła | 458 | 457 | 16877 | 14426 |
| 8 | Noteć | Warta | 391 | 391 | 17302 | 17302 |
| 9 | Wieprz | Wisła | 349 | 349 | 10497 | 10497 |
| 10 | Pilica | Wisła | 333 | 333 | 9258 | 9258 |
| 11 | Bóbr | Odra | 279 | 276 | 5874 | 5830 |
| 12 | | | | | | |

*Rysunek 3: Przykład danych udostępnionych w formie ustrukturyzowanej*

*** Używanie formatów otwartych (np. CSV zamiast arkusza kalkulacyjnego)

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Nazwa rze | Recypient | Długość w | Długość w | Powierzch | Powierzchnia dorzecza w km2 w Polsce | | | | |
| 2 | Wisła | Morze Bał | 1022 | 1022 | 193960 | 168868 | | | | |
| 3 | Odra | Morze Bał | 840 | 726 | 119074 | 106043 | | | | |
| 4 | Warta | Odra | 795 | 795 | 54520 | 54520 | | | | |
| 5 | Bug | Narew | 774 | 590 | 38712 | 19239 | | | | |
| 6 | Narew | Wisła | 499 | 443 | 74527 | 53846 | | | | |
| 7 | San | Wisła | 458 | 457 | 16877 | 14426 | | | | |
| 8 | Noteć | Warta | 391 | 391 | 17302 | 17302 | | | | |
| 9 | Wieprz | Wisła | 349 | 349 | 10497 | 10497 | | | | |
| 10 | Pilica | Wisła | 333 | 333 | 9258 | 9258 | | | | |
| 11 | Bóbr | Odra | 279 | 276 | 5874 | 5830 | | | | |
| 12 | | | | | | | | | | |

*Rysunek 4: Przykład formatu CSV zobrazowany w tabeli arkusza kalkulacyjnego*

```
Nazwa rzeki;Recypient;Długość w km ogółem;Długość w km w Polsce;Powierzchnia dorzecza w km2 ogółem;Powierzchnia dorzecza w km2 w Polsce
Wisła;Morze Bałtyckie;1022;1022;193960;168868
Odra ;Morze Bałtyckie;840;726;119074;106043
Warta ;Odra;795;795;54520;54520
Bug;Narew;774;590;38712;19239
Narew;Wisła;499;443;74527;53846
San;Wisła;458;457;16877;14426
Noteć;Warta;391;391;17302;17302
Wieprz;Wisła;349;349;10497;10497
Pilica;Wisła;333;333;9258;9258
Bóbr;Odra;279;276;5874;5830
```

*Rysunek 5: Przykład formatu CSV zobrazowany w postaci pliku tekstowego*

## Wytyczne do pracy z danymi

Najlepsze praktyki związane z harmonizacją zbiorów są zgodne z programem otwierania danych publicznych

1. Otwieranie danych. Podręcznik dobrych praktyk https://dane.gov.pl/media/ckeditor/2018/11/22/otwieranie-danych-podrecznik-dobrych-praktyk.pdf
2. Standardy otwartości danych. Standard techniczny https://dane.gov.pl/media/ckeditor/2018/10/04/standard-techniczny_X9RAc3r.pdf

## Główne zasady formatowania danych

- Niezależnie od wyboru formatu pliku danych wymaga się stosowanie właściwego formatowania w szczególności dla danych typu: liczba, data, godzina, wartość logiczna.

- Na poziomach otwartości 4 i 5 każda właściwość ma określony przez URI typ, który wyznacza oczekiwany format wartości (zob. szczegółowe opisy poniżej w punktach dot. CSV, JSON, XML). Przykładowo, jeśli mamy do czynienia z właściwością schema:datePublished, to jednoznacznie ustalony zostaje typ wartości na schema:Date, który z kolei jest datą w formacie zgodnym z normą ISO 8601.
- Na tych poziomach wymagane jest określenie typu danych w ramach definicji właściwości.
- Na poziomach otwartości do 3 zaleca się również stosowanie zapisu daty i czasu w formacie zgodnym z normą ISO 8601, oraz odpowiednio podstawowych typów danych ze standardu XML Schema. Oznacza to w szczególności: stosowanie literalnych wartości true, false jako wartości logicznych, stosowanie kropki dziesiętnej (a nie przecinka) w zapisie ułamków dziesiętnych, bez żadnych dodatkowych separatorów (np. oddzielających tysiące); dopuszczalny jest tzw. zapis naukowy, zapis dat w postaci yyyy-mm-dd, a łącznie dat i godzin w postaci yyyy-mmddThh:mm lub yyyy-mm-dd hh:mm:ss (spacja między datą a czasem).
- Dane zaleca się udostępniać w możliwie najwyższym stopniu granulacji (rozdrobnienia), tzn. nie łączyć kilku danych w jednym polu.

Przykłady formatowania danych

| Data | Liczba | Wartość logiczna |
|---|---|---|
| 1970-01-01 | 42.01 | true |
| 2018-12-31 | 1.05e3 | false |

*Rysunek 6:Rysunek 6: Przykład prawidłowo sformatowanych danych*

| Data | Liczba | Wartość logiczna |
|---|---|---|
| 1/1/1970 | 45,01 | T |
| 1.1.1970 | 1,234.56 | 0 |
| 1 I 1970 | 1 234,5 | NIE |

*Rysunek 7: Rysunek 8: Przykład nieprawidłowo sformatowanych danych*

| Nazwa | Kod PNA | Miejscowość | Cecha | Nazwa ulicy | Nr budynku | Nr lokalu | Nr telefonu |
|---|---|---|---|---|---|---|---|
| CSIOZ | 00-184 | Warszawa | ul. | Stanisława Dubois | 5A | | 225970927 |
| KPRM | 00-583 | Warszawa | al. | Aleje Ujazdowskie | 1/3 | | 226946000 |
| Biuro | 51-152 | Wrocław | pl. | Marsz. Józefa Piłsudskiego | 4 | 75 | 712558765 |

*Rysunek 8: Przykład prawidłowo sformatowanych danych*

| Nazwa | Adres | Nr telefonu |
|---|---|---|
| CSIOZ | 00-184 Warszawa, ul. Dubois 5A | +48 22 597-09-27 |
| KPRM | 00-583 Warszawa, Aleje Ujazdowskie 1/3 | (22) 694-60-00 (cent.) |
| Biuro | 51-152 Wrocław, J. Piłsudskiego 4/75 | 71-25-58-765 |

*Rysunek 9: Przykład nieprawidłowo sformatowanych danych*

## Transformacja danych

Faza przekształcenia (transformacji) obejmuje przetwarzanie danych za pomocą funkcji i reguł w celu uzyskania pożądanej struktury modelu danych. Typowe działania na tym etapie to:

- Obliczanie nowych wartości
- Tłumaczenie zaszyfrowanych wartości
- Zmiana nazw pól
- Łączenie tabel
- Agregacja wartości
- Tworzenie tabel przestawnych
- Weryfikacja danych

W przypadku przetwarzania danych przestrzennych wyróżnia się dodatkowo:

- Zmiana odwzorowania: konwersji danych przestrzennych z jednego systemu, na drugi współrzędnych.
- Analizy przestrzenne: modelowanie zależności przestrzennych i obliczenia na ich podstawie produktów wynikowych
- Transformacje topologiczne: tworzenia relacji między odrębnymi zbiorami danych topologicznych
- Resymbolizacja: zdolność do zmiany kartograficznych cech funkcji, takich jak kolor linii lub stylu (w przypadku transformacji danych)
- Geokodowanie: konwersja atrybutów danych tabelarycznych do postaci danych przestrzennych.

W momencie, kiedy mamy już dane z systemu źródłowego, powinniśmy je przekształcić do wybranej przez nas formy. W sytuacji, gdy mamy dane z wielu różnych źródeł, musimy umieścić je w jednej, wspólnej strukturze. Na tym etapie wykonywane są różne transformacje, takie jak mapowanie pól, skracanie nazw, ustalanie wspólnego formatu łańcuchów znaków czy nadawanie identyfikatorów. Chodzi o to, aby różne struktury połączyć w jedną. Zwykle ten krok wymaga od użytkownika wiele pracy i mnóstwa pomysłów. Często relatywnie proste transformacje, przy dużym wolumenie danych, potrafią być niewydajne i konieczne jest poszukiwanie innych metod.

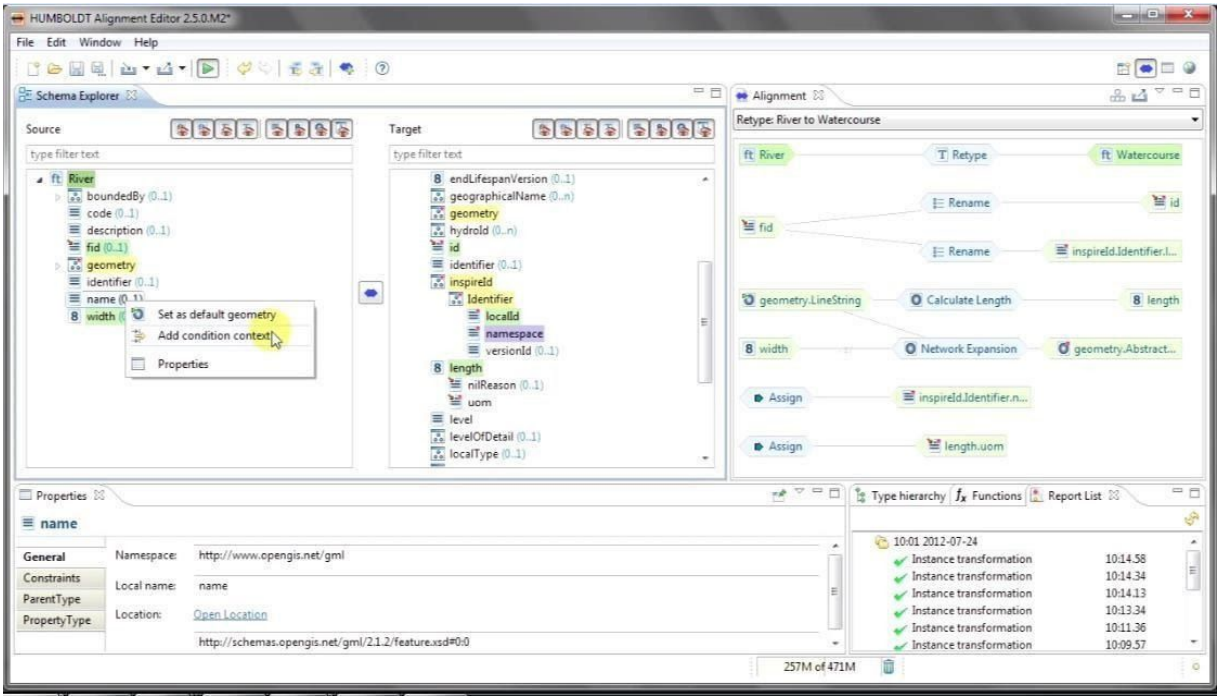
Punktem wyjścia niezależnie od typu zbioru jest modelu danych źródłowych (z przykładowymi wartościami) i modelu danych docelowych (z przykładowymi wartościami) w formie tabeli np. w arkuszu kalkulacyjnym (przykład poniżej).

| TARGET MODEL<br>INSPIRE feature catalogue v3.0rc3 | | | | | | SOURCE MODEL<br>Cyfrowa mapa glebowo-rolnicza 1:5 000 oraz baza danych profili glebowych dla województwa podkarpackiego | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Feature Name** | **Attribute Name** | **Attribute Type** | **Attribute Cardinality** | **Possible Values** | **Attribute Definition** | **Feature Name** | **Attribute Name** | **Attribute Type** | **COMMENTS** |
| **SoilPlot** | **FeatureType** | | | | | **Stanowisko badań gleby (Działka gleby)** | | | |
| Spot where a specific soil investigation is carried out. | inspireId | Identifier | 0..1 | DataType | External object identifier of the soil profile. | [ODKRYWKA] | ID_ODKR.ID | namespace:<br>PL.PZGiK.000.SO.3.3<br>localID:<br>SoilPlot.[ID_ODKR.ID] | do ustalenia, zbiór danych nie zgłoszony do EZiUDP |
| | soilPlotLocation | Location | 1 | (xlink:href)<br>EXAMPLE reference to a place name, municipality or reference to an exact X,Y location. | A reference to a location on the earth; it can be a point location identified by coordinates or a description of the location using text or an identifier. | | Geometry | Point | opis położenia (pośredni lub bezpośredni) "Stanowisko badań gleby"<br><br>współrzędne położenia profilu |
| | soilPlotType | SoilPlotTypeValue | 1 | CodeList<br>The allowed values for this code list comprise only the values specified in Annex C. | Gives information on what kind of plot the observation of the soil is made on. | | | trialPit | odkrywka |
| | beginLifespanVersion | DateTime | voidable<br>1 | | Date and time at which this version of the spatial object was inserted or changed in the spatial data set. | | X_DATA_UTWORZENIA | | |
| | endLifespanVersion | DateTime | voidable<br>0..1 | | Date and time at which this version of the spatial object was superseded or retired in the spatial data set. | | x | | brak danych |
| | observedProfile | ObservedSoilProfile | voidable<br>1 | FeatureType<br>(xlink:href) | Link to the observed soil profile for which the soil plot provides location information. | | ID_ODKR.ID<br>(ObservedSoilProfile) | referencja na "Badany profil glebowy"<br><br>w przypadku braku wartości:<br>unknown | |
| | locatedOn | SoilSite | voidable<br>0..1 | FeatureType<br>(xlink:href) | Link to the soil site on which the soil plot is located or to which the soil plot is belonging. | | ID_OBR.ID<br>(SoilSite) | referencja na "Obszar badania gleby" | |

Tego typu zestawienie pozwala odnaleźć kluczowe podobieństwa i zależności, które muszą zostać zaimplementowane podczas projektowania procesu transformacji w oprogramowaniu typu ETL.

Zestawienie danych pozwala też odkryć wszelkie niespójności związane ze strukturą danych takimi jak braki w atrybutach, niewłaściwie zapisane dane, niespójne formaty czy błędy w sposobie zapisu samych danych.

Po dokonaniu wstępnego przeglądu i określeniu kierunków transformacji (wymienionych na wstępie rozdziału tj. *obliczanie nowych wartości, tłumaczenie (…)* itd. można przejść do czynności polegających na zaimplementowaniu reguł przejścia w aplikacjach do modelowania procesu ETL, np. Hale Studio. Aplikacje tego typu posiadają stosowny zestaw narzędzi pozwalający na „wyklikanie” reguł transformacji dla dowolnego zbioru wejściowego. Przykład takiego procesu zaprezentowano poniżej.




### Ładowanie danych

Ten etap polega na załadowaniu danych do tabel docelowych, czyli tych, w których będą one składowane i odczytywane. Dane powinny być oczyszczone i ujednolicone.

## Instalacja i konfiguracja narzędzia ETL – Hale Studio

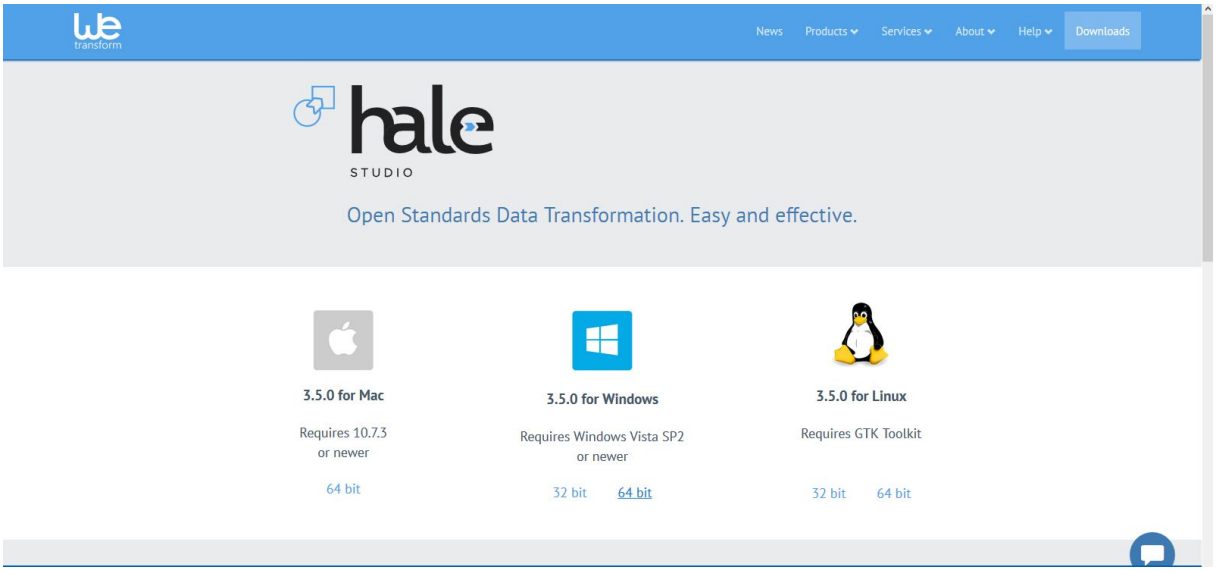
## Konfiguracja środowiska narzędzia i jego omówienie, w tym omówienie możliwości *readerów* i *writerów*, ograniczenia, porównanie do innych narzędzi typu ETL)

HALE Studio jest programem napisanym w Javie i jako taki wymaga wirtualnej maszyny Java do działania. Dystrybucja dla systemu Windows zawiera już wbudowaną maszynę wirtualną, więc nie ma potrzeby instalacji. W systemach MacOS i Linux wymagane jest Java JDK (nie JRE) w wersji min. 8

# Instalacja HALE Studio w Windows

## Pobranie instalatora

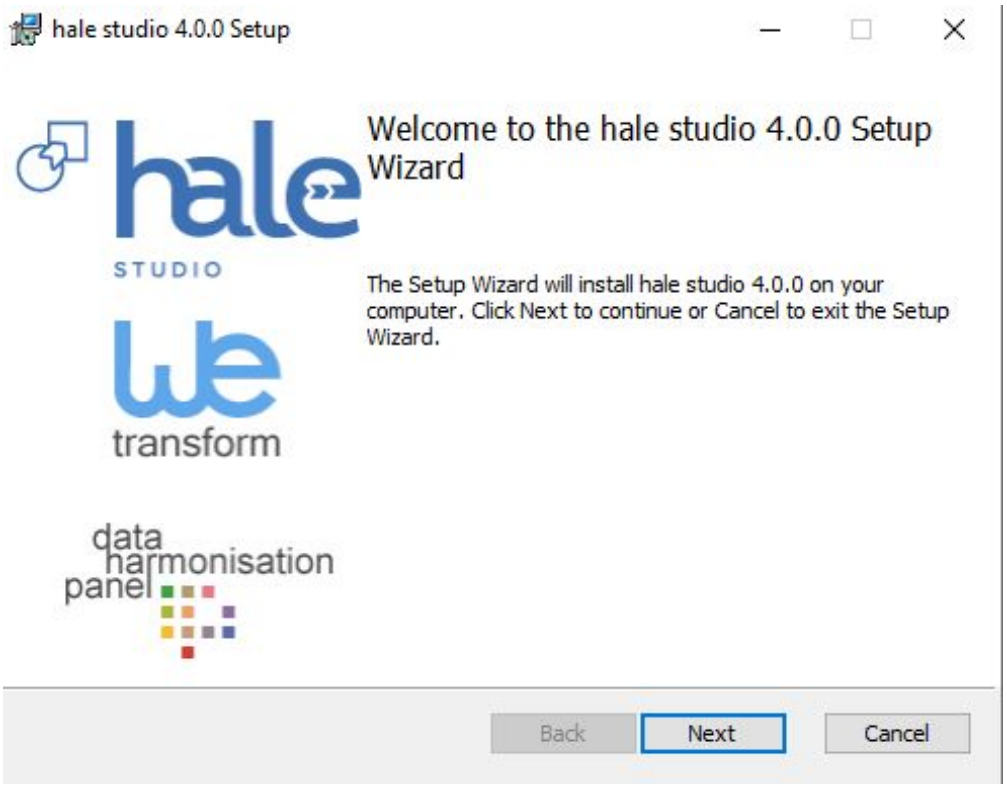
Instalator HALE jest dostępny pod adresem https://www.wetransform.to/downloads/ - należy wybrać wersję odpowiednią dla swojego systemu - obecnie najczęściej 64 bit.



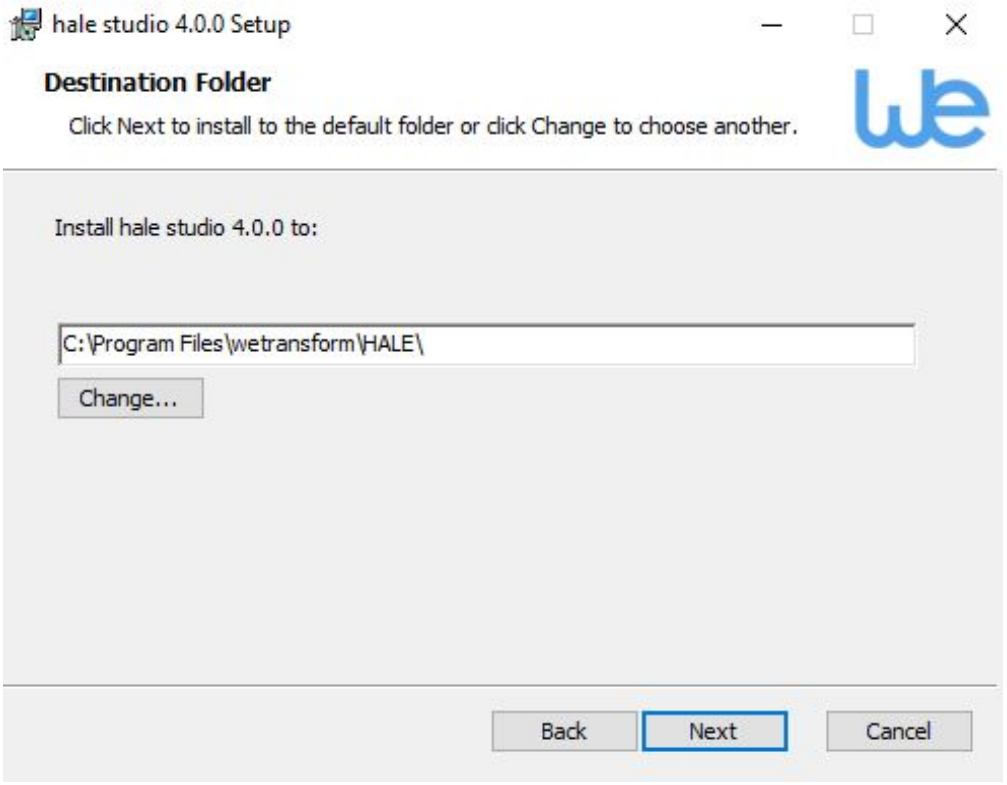

Pakiet instalacyjny dla Windows zawiera wszystkie komponenty potrzebne do uruchomienia programu - nie jest konieczna oddzielna instalacja Javy.

## Przebieg instalacji

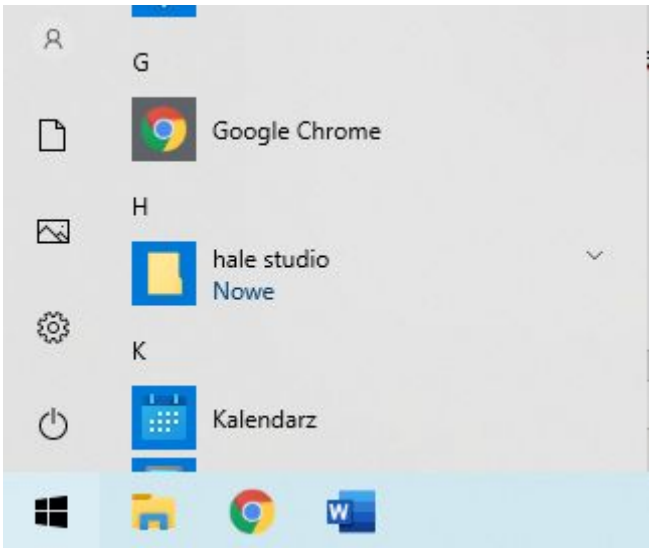
Instalator nie jest rozpoznawany jako "zaufany" przez Windows, więc należy zaakceptować ostrzeżenia bezpieczeństwa. Dalszy proces instalacji przebiega standardowo.




Ważne jest, aby dokonać instalacji w katalogu, w którym użytkownik ma prawo zapisu np.: C:\Users\Nazwa użytkownika\wetransform



Po zakończonej instalacji program będzie dostępny w menu Start pod nazwą "hale studio".




## Definicja ETL

**E**xtract, **T**ransform and **L**oad - proces pozyskania danych dla baz danych, składający się z:

- pozyskania danych ze źródeł zewnętrznych,
- przekształcenia danych,
- załadowania danych do bazy docelowej.

ETL wykorzystywany jest np. do migracji danych pomiędzy systemami, przy aktualizacji, do zasilania systemów klasy Business Intelligence. Wobec faktu, iż proces harmonizacji danych na potrzeby Inspire wymaga wykonywania powtarzalnych czynności prowadzących do uwspólniania formy przekazywanych danych wynikający z różnych sposobów zapisu danych źródłowych przez kraje członkowskie, oprogramowanie typu ETL jest powszechnie stosowane w celu automatyzacji procesu przejścia od zbioru źródłowego do formy wskazanej w rozporządzeniach UE.
Inne oprogramowanie typu ETL, które może być użyte do danych przestrzennych to na przykład:

- GDAL / OGR (open source)
- GeoKettle (open source)
- FME (komercyjny)

Cechy poszczególnych programów zestawiono w tabeli poniżej:

| ***Cecha*** | ***HALE Studio*** | ***GeoKettle*** | ***FME*** |
|---|---|---|---|
| Wbudowane schematy INSPIRE | TAK | NIE | TAK |

| | | | |
|---|---|---|---|
| Możliwość odczytu i zapisu z i do WFS | TAK | NIE | TAK |
| Dostępna usługa chmurowa | TAK - do przechowywania projektów i publikacji danych (hale connect) | NIE | TAK - do przeprowadzania transformacji w chmurze (FME Cloud) |
| Dostępne formaty danych | 10 formatów | 45 formatów | 335 formatów |
| Dostępne funkcje geoprocessingu | 6 funkcji | 27 funkcji | 80 funkcji |
| Cena | 0 | 0 | 2000 € |
| Główne zastosowanie | Harmonizacja danych INSPIRE | Migracja danych przestrzennych | Migracja i przetwarzanie danych przestrzennych |

HALE Studio obsługuje następujące formaty wejściowe **schematów**:

- XSD (w tym WFS DescribeFeatureType)
- własny format Hale Schema Definition (w wariantach JSON i XML)
- CSV
- XLS i XLSX
- MDB
- SQLite
- ESRI Shapefile
- PostGIS
- Microsoft SQL Server

HALE Studio obsługuje następujące formaty wejściowe **danych**:

- XML i GML - także skompresowane programem gzip (to co innego, niż najpopularniejszy ZIP)
- CSV
- XLS i XLSX
- ESRI Shapefile
- MDB
- SQLite
- WFS GetFeature
- PostGIS

- Microsoft SQL Server

HALE Studio obsługuje następujące formaty wynikowe **schematów**:

- deegree
- Hale Schema Definition (XML)
- Hale Schema Definition (JSON)

Podane wyżej formaty służą zapisowi ustawień i reguł przejścia zdefiniowanych w trakcie pracy nad projektem w aplikacji do ich późniejszego wykorzystania.

HALE nie umożliwia obsługi schematów koncepcyjnych, jak UML. Taką funkcjonalność miało narzędzie HUMBOLDT Conceptual Schema Transformer - niestety usługa sieciowa zakończyła działalność i narzędzie już nie jest dostępne.

HALE Studio obsługuje następujące formaty wynikowe **danych**:

- CSV
- PostGIS
- Microsoft SQL Server
- GML
- GeoJSON
- SQLite
- XLS
- XML (Custom root element)
- WFS-T

## Ćwiczenie 1: Przygotowanie zbioru, harmonizacja próbki danych

Ćwiczenie „Harmonizacja próbki danych” będzie polegało na wykonaniu harmonizacji dla danych o obszarach Natura 2000 pochodzących z Geoserwisu GDOŚ.

Ćwiczenie ma na celu przejście procesu:
1. Importu schematu danych wejściowych
2. Importu danych źródłowych
3. Import schematu wyjściowego
4. Ustawienie reguł transformacji
5. Wykonanie transformacji
6. Eksport wyników przetworzenia

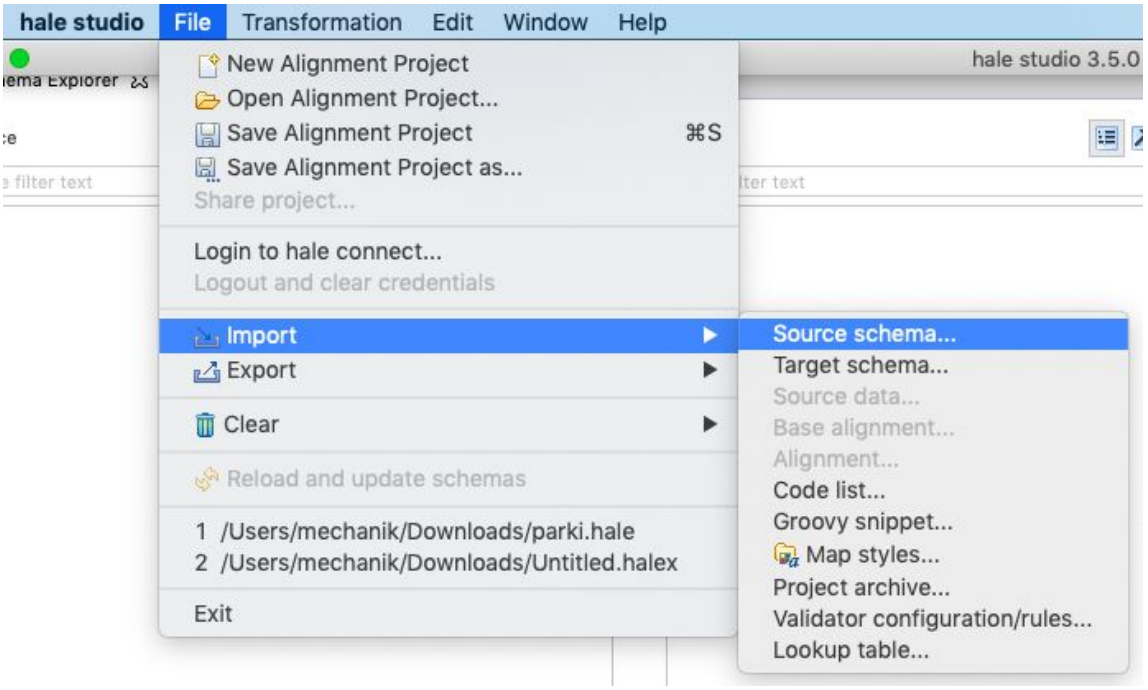
Ćwiczenie składa się z dwóch części – omówienia pełnego procesu przejścia krok po kroku a następnie wykonania transformacji według podanego schematu krok po kroku.

Metodyka harmonizacji została oparta o publikację:

Krawczyk A., Garguła A., 2018, Harmonization of Polish Natura2000 data sets with the protected sites data schema of inspire directive in the environment of Humboldt Alignment Editor (HALE), Geoinformatica Polonica 17:7-15, DOI: 10.4467/21995923GP.18.001.9158

Schemat wejściowy może być zadany poprzez definicję schematu XSD lub istniejący zbiór danych (plik lub baza danych). W przypadku obszarów Natura 2000 rozpowszechnianych w formie plików SHP, właściwą opcją będzie import z pliku - 

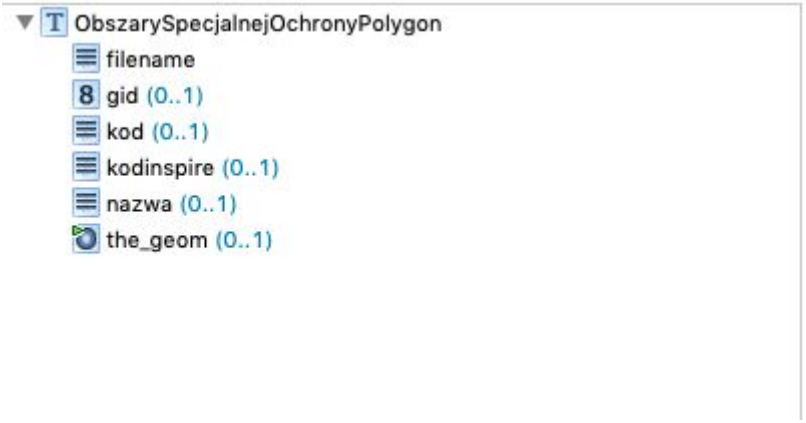

**Import -> Source schema**




28

Inspekcja schematu

Import SHP jako schematu powoduje załadowanie wyłącznie schematu- same dane nie zostaną wczytane.

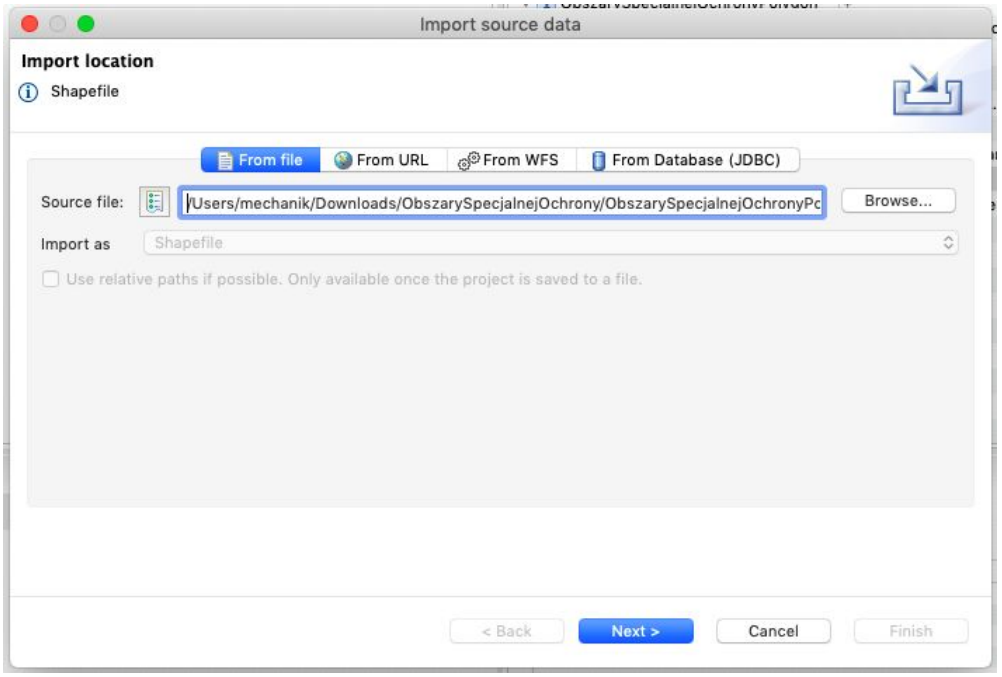



Symbol  oznacza atrybut tekstowy

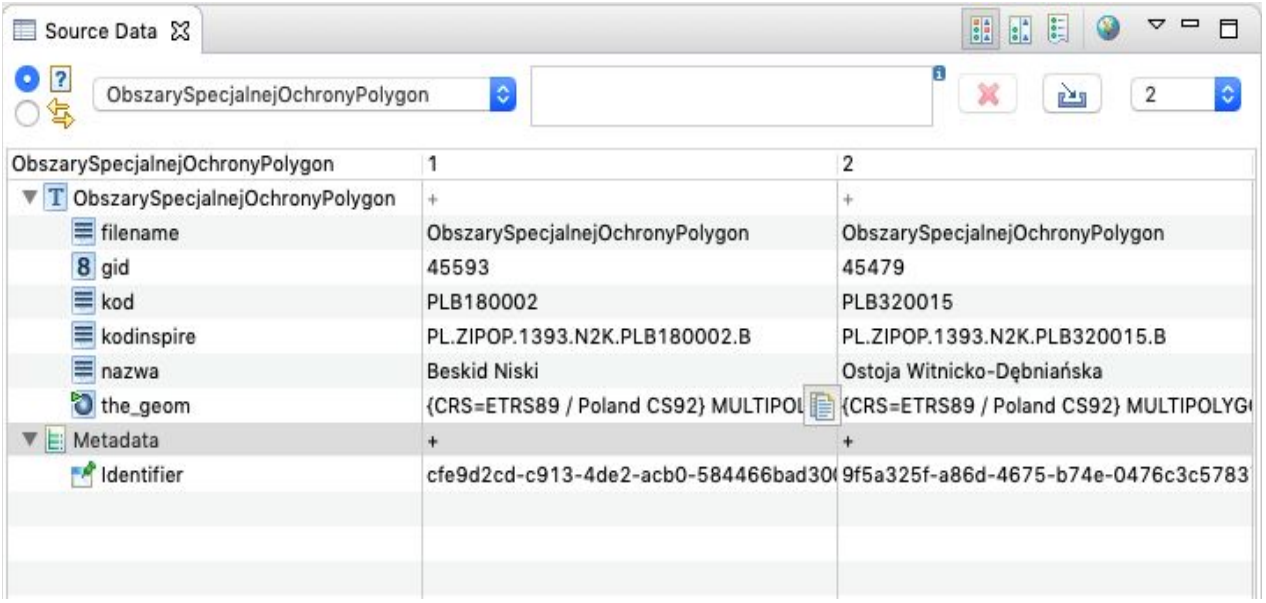
 oznacza atrybut liczbowy

 oznacza geometrię

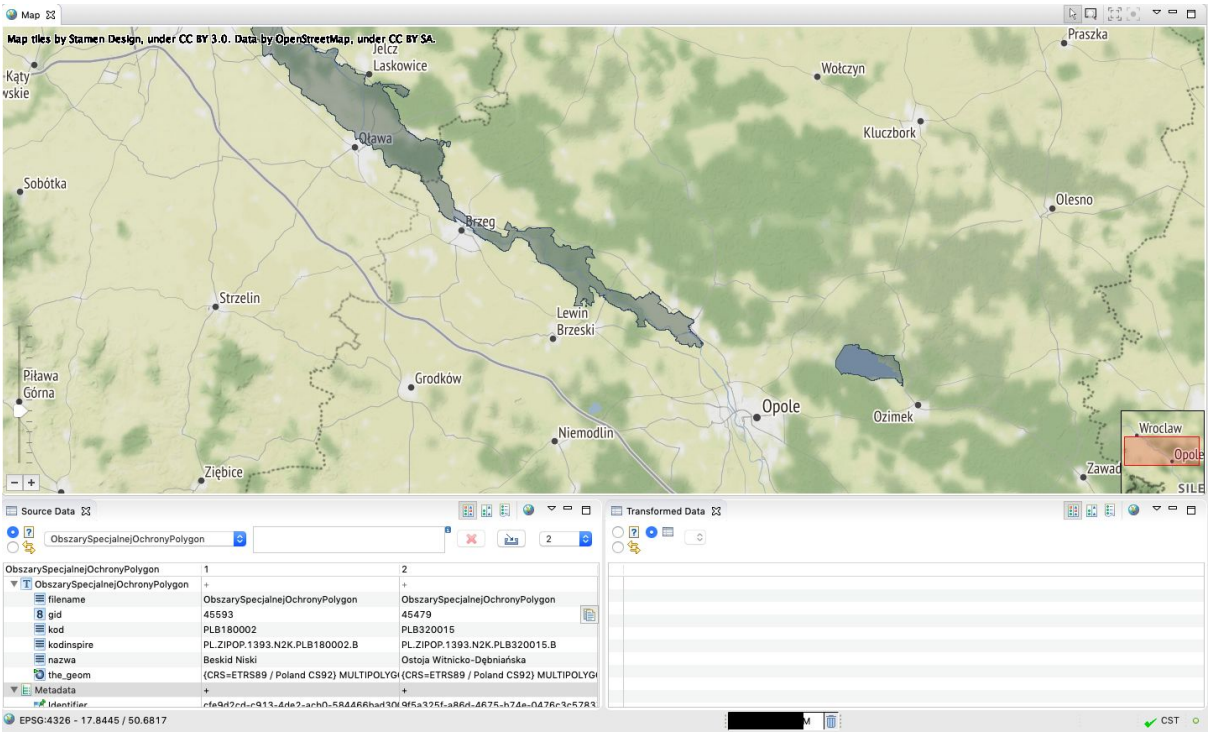
Import danych musi być wykonany po wczytaniu schematu wejściowego. Domyślnie importowana jest próbka 500 pierwszych obiektów. Wielkość próbki można ustawić w oknie Window -> Preferences -> Project -> Source data.



Dostępne są 2 widoki dla danych, widok tabeli




oraz widok mapy  - przełącznik widoków znajduje się w prawym górnym rogu ekranu.



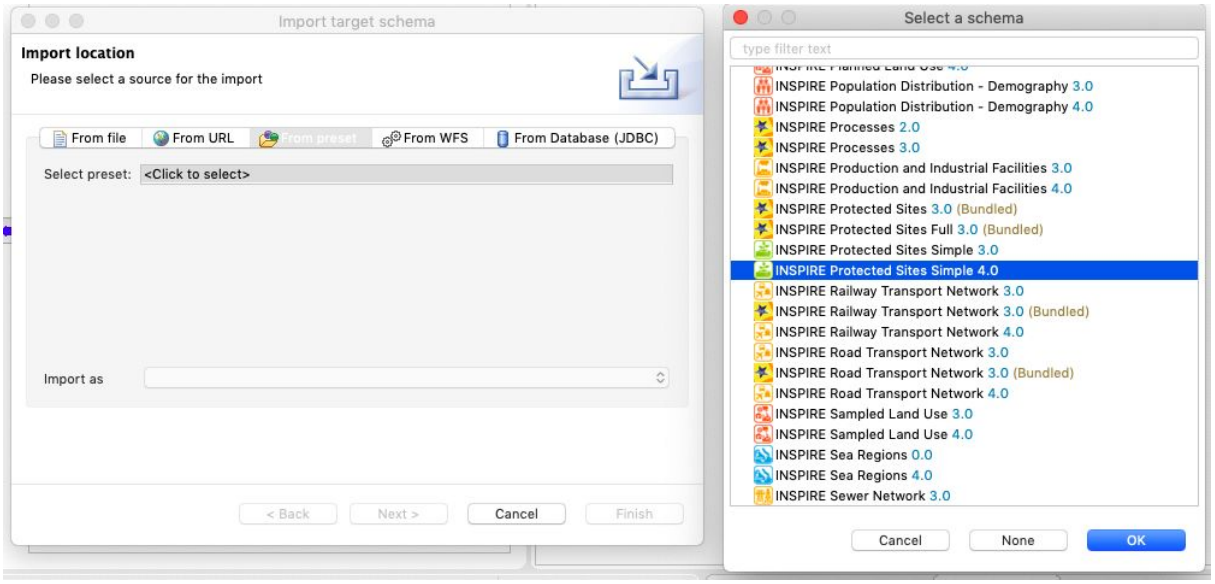

Prezentacja atrybutów przykładowego obiektu źródłowego

| | |
|---|---|
| ObszarySpecjalnejOchronyPolygon | + |
| kod | PLB180002 |
| filename | ObszarySpecjalnejOchronyPolygon |
| gid | 45593 |
| kodinspire | PL.ZIPOP.1393.N2K.PLB180002.B |
| nazwa | Beskid Niski |
| the_geom | {CRS=ETRS89 / Poland CS92} MULTIPOLYGON (((669315.801783 |
| Metadata | + |
| Identifier | cfe9d2cd-c913-4de2-acb0-584466bad300 |

Import schematu wyjściowego INSPIRE

**File -> Import -> Target Schema**

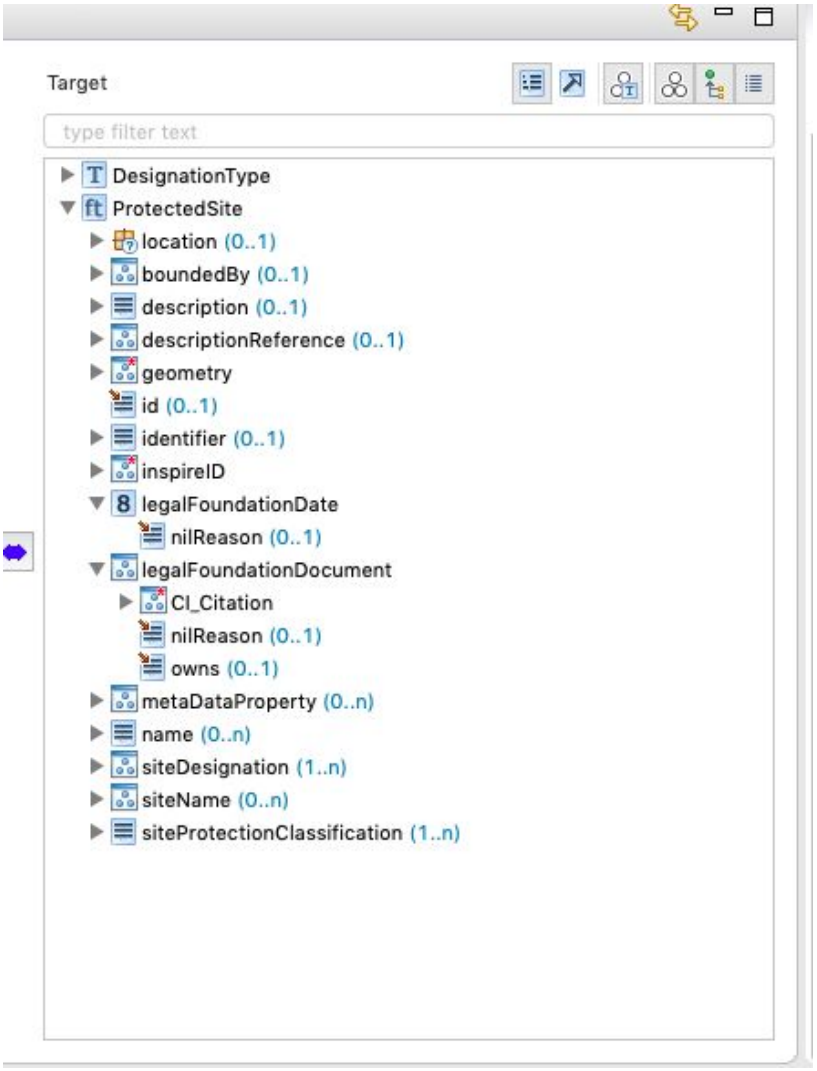
Schematy INSPIRE są dostępne w zakładce 


Schemat "INSPIRE Protected Sites Simple 4.0"

Schematy INSPIRE są schematami złożonymi tzw. Complex Features, co oznacza, że atrybuty mogą być względem siebie zagnieżdżone, tj. atrybut może zawierać swoje własne atrybuty.

Taka struktura jest prosta w realizacji w formatach XML (GML) i JSON (GeoJSON), ale nie pasuje dobrze do plików płaskich np. SHP i relacyjnych baz danych.

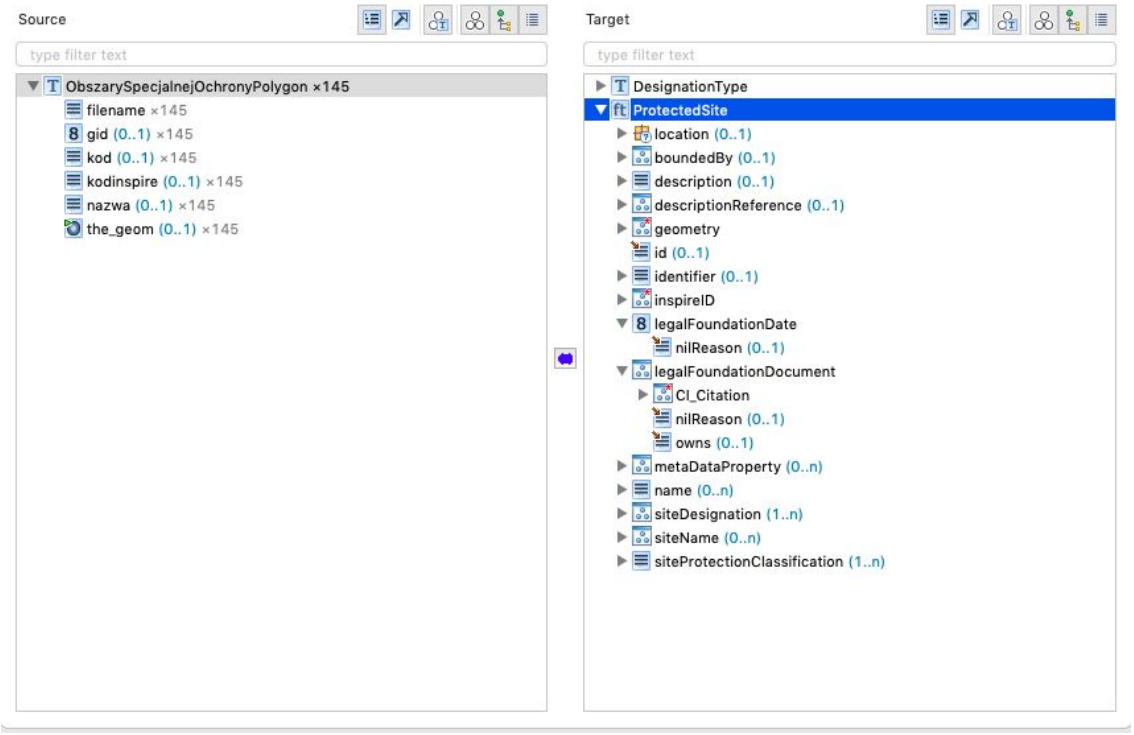



Po wczytaniu schematu wejściowego i wynikowego, można przystąpić do przygotowania mapowania (**Alignment**).

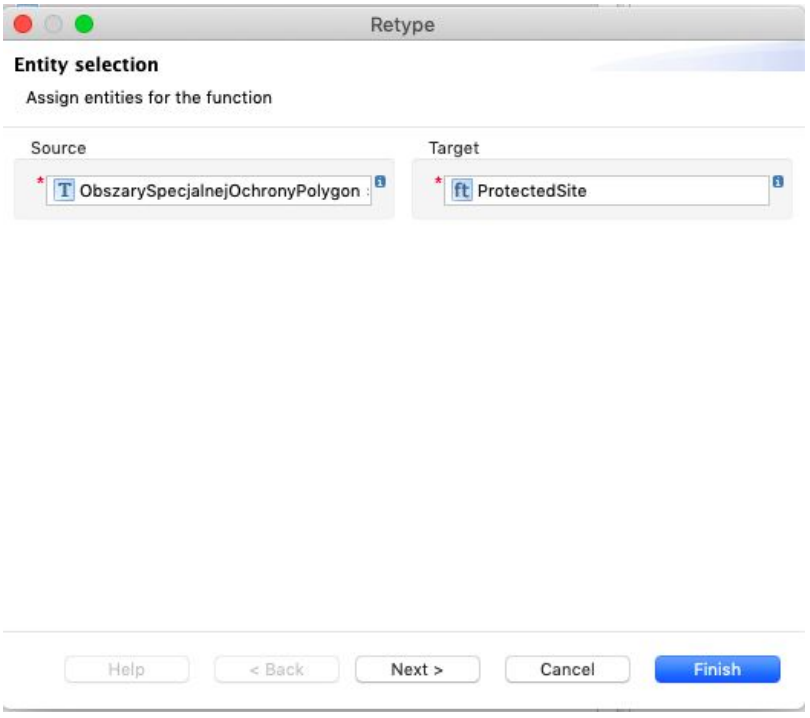
Wczytanie danych źródłowych nie jest obowiązkowe, ale umożliwia podgląd wyników i ewentualnych błędów na bieżąco.

Pierwszym krokiem jest mapowanie typów danych. Najprostszym przykładem jest przeniesienie jednego zbioru wejściowego w jeden zbiór zharmonizowany - zostanie w tym celu użyta funkcja **Retype**.

Mapowanie typów. Należy kliknąć typ źródłowy w panelu **Source**, typ wynikowy w panelu **Target** i kliknąć przycisk  , wybrać **Retype**.

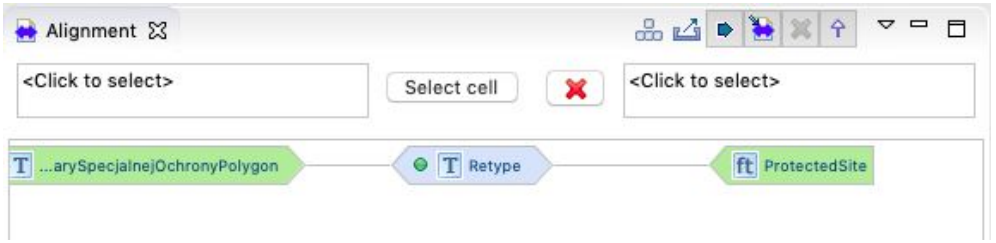



Funkcja **Retype** dokonuje mapowania jednego typu wejściowego na jeden typ wynikowy.

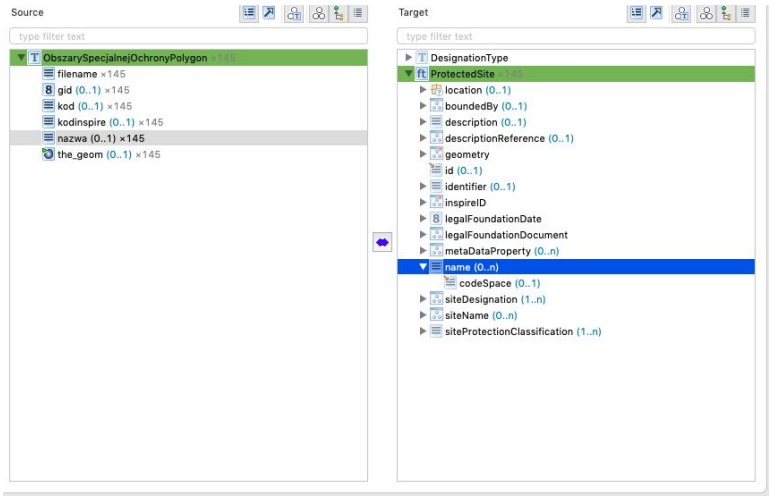


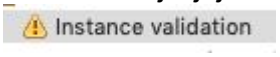

## Mapowanie atrybutów

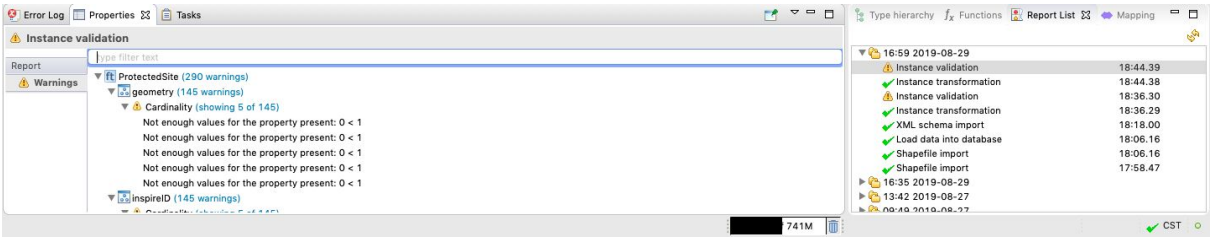
Dodanie funkcji **Retype** dodaje **Mapping Cell** do tabeli mapowań - można wówczas rozpocząć dodawanie kolejnych mapowań dla atrybutów.




Najprostszą funkcją jest **Rename**, czyli przeniesieniu wartości atrybutu A w schemacie wejściowym w niezmienionej formie do atrybutu B w schemacie wynikowym. Przykład - "nazwa" na "name"




Po każdej edycji mapowania następuje wykonanie transformacji i jej walidacja. Błędy walidacji można przeczytać, klikając dwukrotnie w 


Minimalny zestaw mapowań, jaki pozwala na poprawne zaliczenie walidacji to:


Hale Studio pozwala na zastosowanie wielu reguł mapowania wymienionych poniżej:

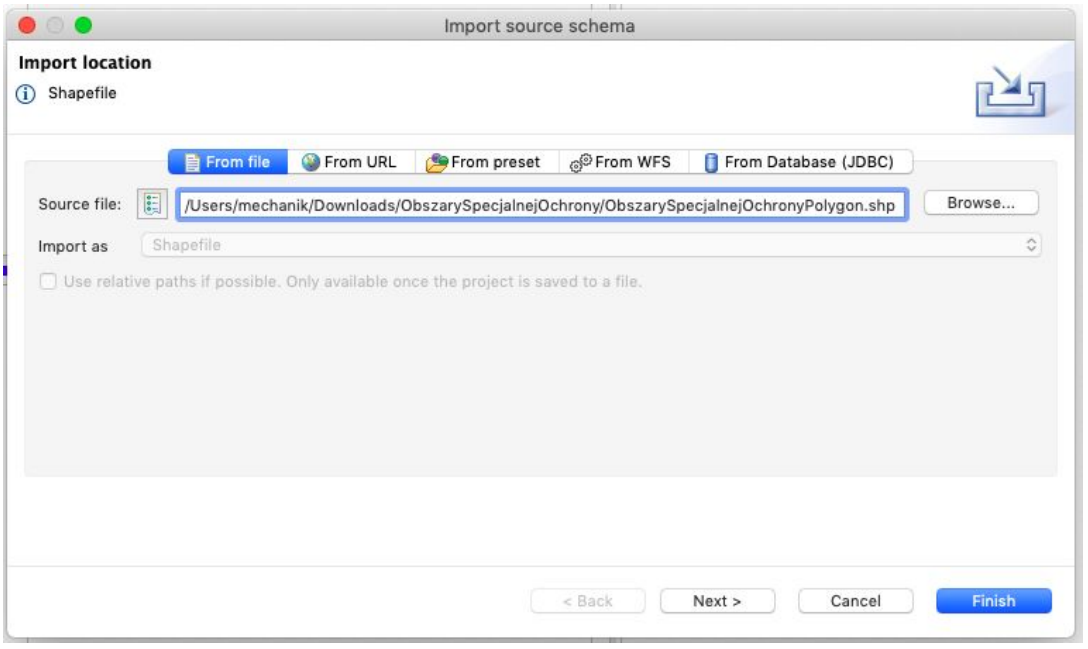
- Retype
  Mapuje atrybut zbioru wejściowego na jeden atrybut pola docelowego
- Merge
  Scala wiele wystąpień różnych typów źródłowych w jedno wystąpienie typu docelowego na podstawie co najmniej jednej pasującej właściwości.
- Join
  Złącza wiele wystąpień różnych typów źródłowych w jedno wystąpienie typu docelowego na podstawie co najmniej jednej pasującej właściwości.
- Create
  Twórz instancję określonego typu schematu
- Date extraction
  Wyodrębnia datę z ciągu tekstowego
- Regex Analysis
  Analizuje ciąg znaków na podstawie wyrażenia regularnego
- Rename
  Kopiuje wartość do atrybutu docelowego
- Assign
  Przypisuje wartość do atrybutu docelowego
- Generate Unique Id
  Przypisuje wygenerowany unikalny identyfikator do atrybutu docelowego
- Classification
  Mapowanie typu słownikowego
- Formatted string
  Tworzy sformatowany ciąg na podstawie wzorca i zmiennych wejściowych
- Inline transformation
  Wbudowana transformacja typów w obrębie właściwości właściwości.
- Assign (Bound)
  Przypisuje wartość do atrybutu docelowego, jeśli atrybut źródłowy występuje
- Spatial Join
  Połączenie przestrzenne obiektów różnych typów, pobiera wartości na podstawie relacji przestrzennej.
- Ordinates to Point
  Tworzy punkt ze zmiennych X, Y i opcjonalnie zmiennej Z
- Network Expansion
  Tworzy bufor wokół geometrii
- Calculate Length
  Oblicza długość geometrii

- Calculate Area
  Oblicza powierzchnię geometrii.
- Centroid
  Oblicza centroid (środek ciężkości) danej geometrii.
- Compute Extent
  Oblicza zasięg geometryczny na podstawie wszystkich geometrii wejściowych. Dostępne są opcje Bounding Box, Convex Hull i Union.
- Aggregate
  Łączy podobne geometrie wejściowe
- Reproject Geometry
  Dokonuje reprojekcji (przeliczenia układu współrzędnych) geometrii.
- Interior Point
  Oblicza punkt wewnętrzny geometrii (w 2D). Punkt wewnętrzny będzie leżeć we wnętrzu geometrii, jeśli możliwe jest dokładne obliczenie takiego punktu. W przeciwnym razie punkt może leżeć na granicy geometrii (np. jeśli geometria jest linią).
- Groovy (Retype, Create, Merge, Join)
  Zestawy funkcji wykonywanych poprzez interpreter języka Groovy
- Mathematical Expression
  Definiuje wartość, używając wyrażenia matematycznego z obsługą zmiennych
- Generate sequential ID
  Generuje liczbowy identyfikator pobierając kolejną wartość z sekwencji (np. 1, 2, 3...n)

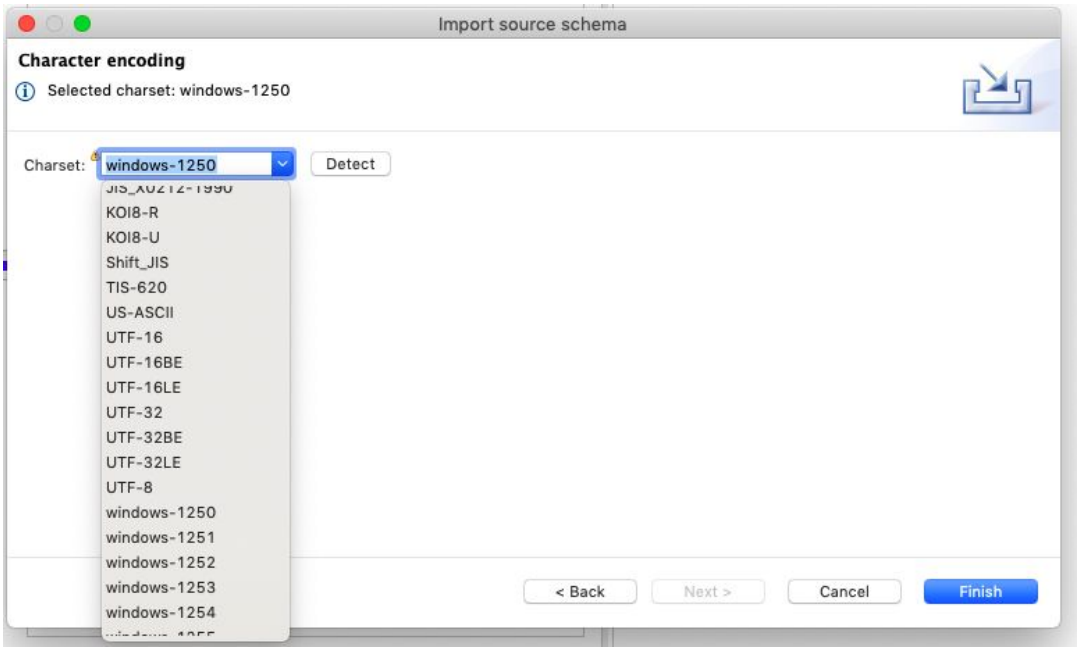
Część praktyczna ćwiczenia:

## Import schematu źródłowego

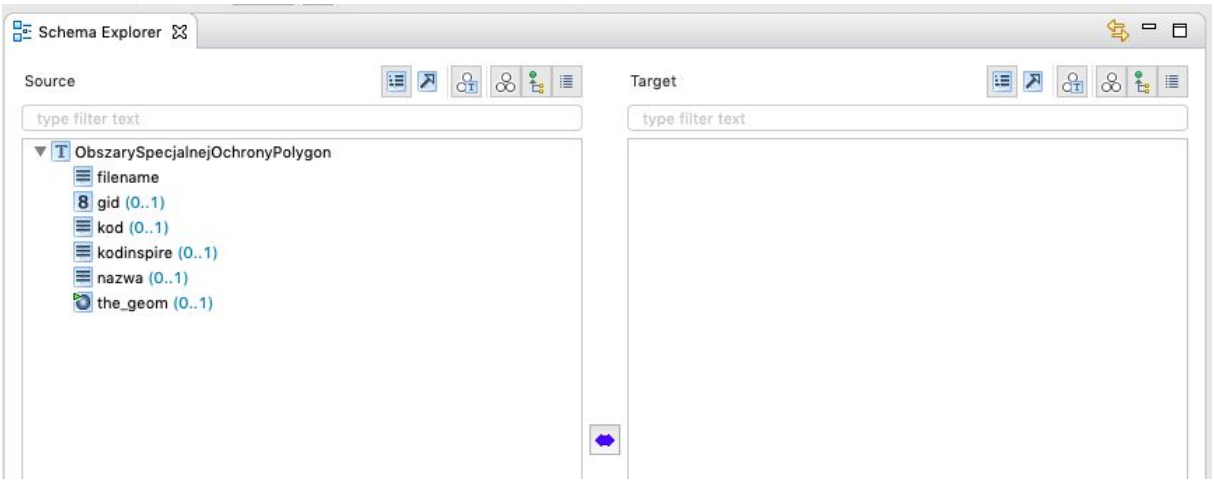
Należy wybrać File -> Import -> Source schema, a następnie jako Source file wybrać plik "ObszarySpecjalnejOchronyPolygon.shp" z katalogu "HALE/masowa_konwersja" z danych do ćwiczeń.



Następnie należy kliknąć **Next,** wybrać kodowanie windows-1250 i zatwierdzić poprzez **Finish.** Kodowanie windows-1250 jest w tym przypadku inicjalnym sposobem kodowania ciągów tekstowych zapisanych w pliku źródłowym. Podanie prawidłowego kodowania pozwoli na właściwą konwersję znaków w dalszych działaniach związanych z transformacją danych. Innym popularnym sposobem kodowania znaków stosowanym w polskich systemach GIS jest format UTF-8.

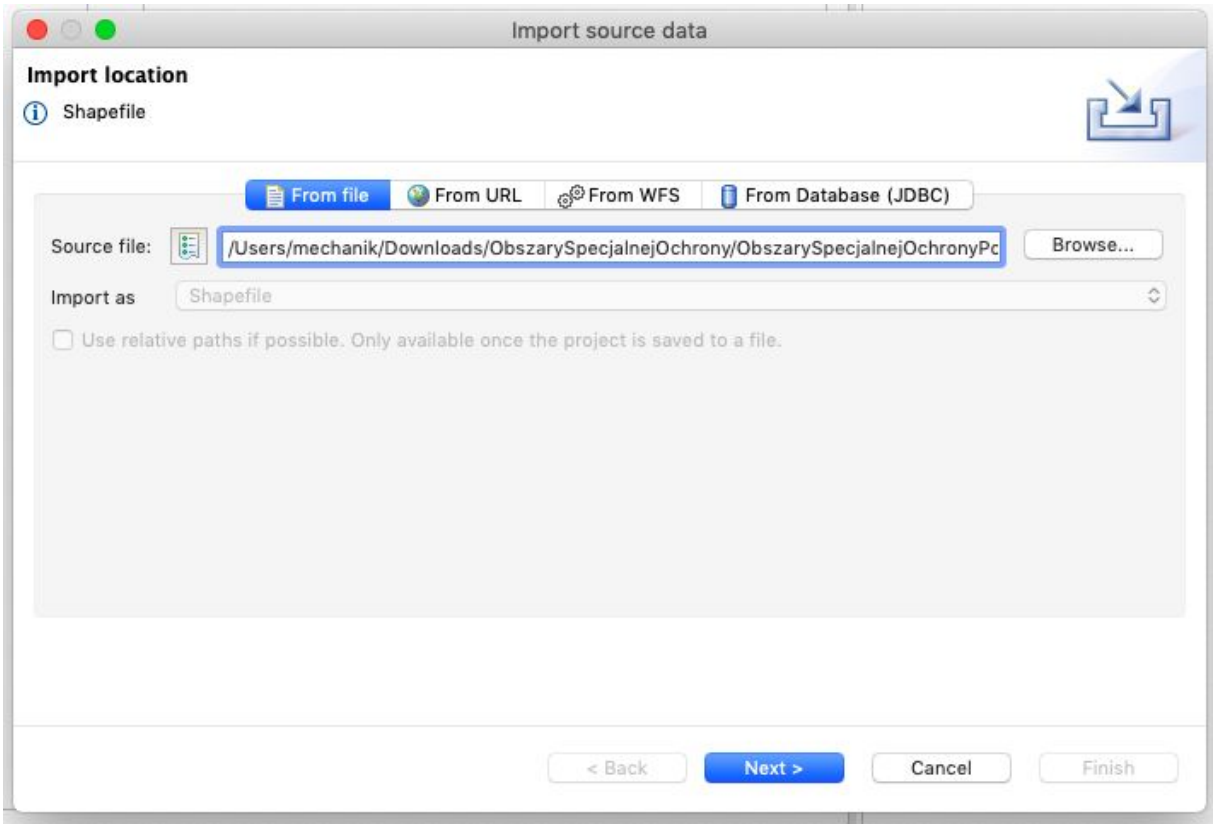


Panel **Schema Explorer** powinien wyświetlić dostępne atrybuty w schemacie źródłowym:

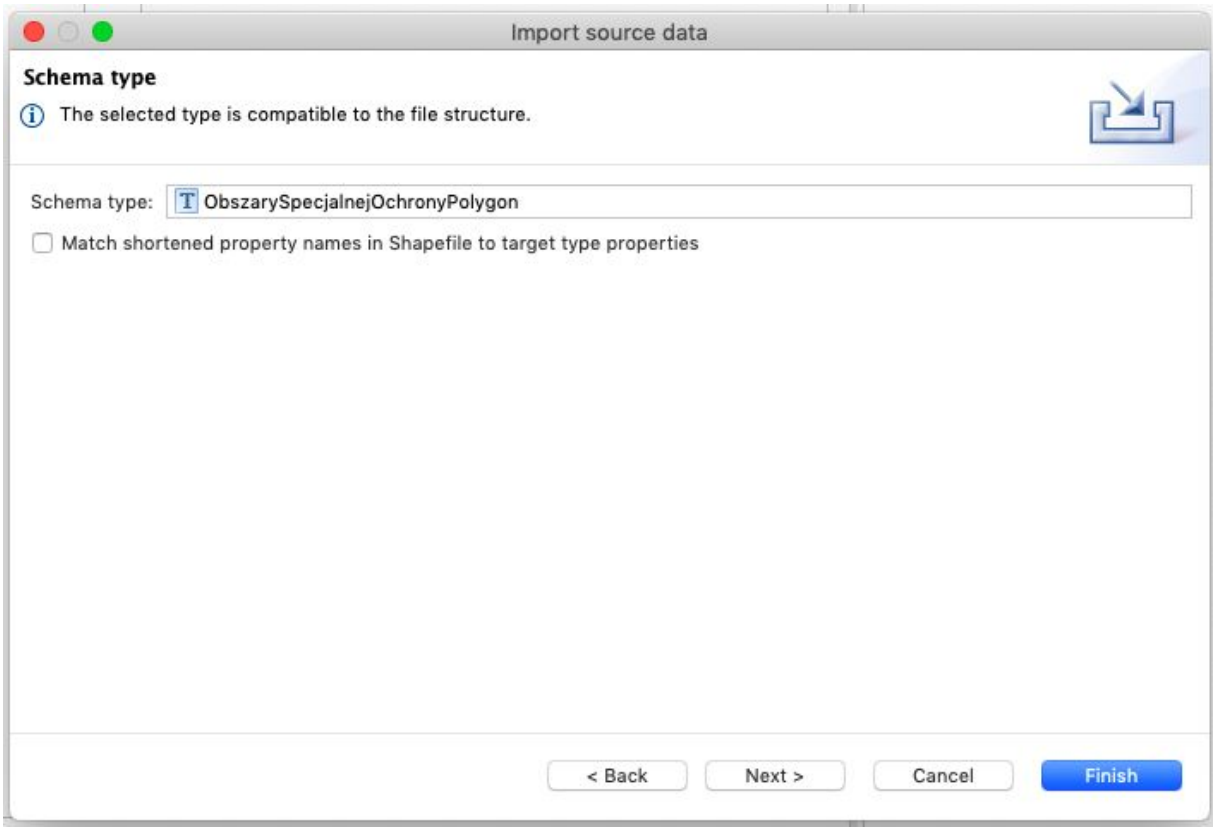



## Import danych źródłowych

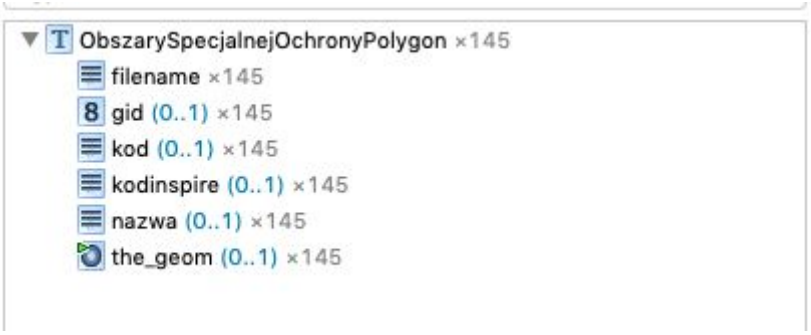
Należy wybrać z menu **File -> Import -> Source data**, a następnie ponownie wybrać plik ObszarySpecjalnejOchronyPolygon.shp.




Po zatwierdzeniu poprzez **Next** należy zatwierdzić relację między schematem a danymi.

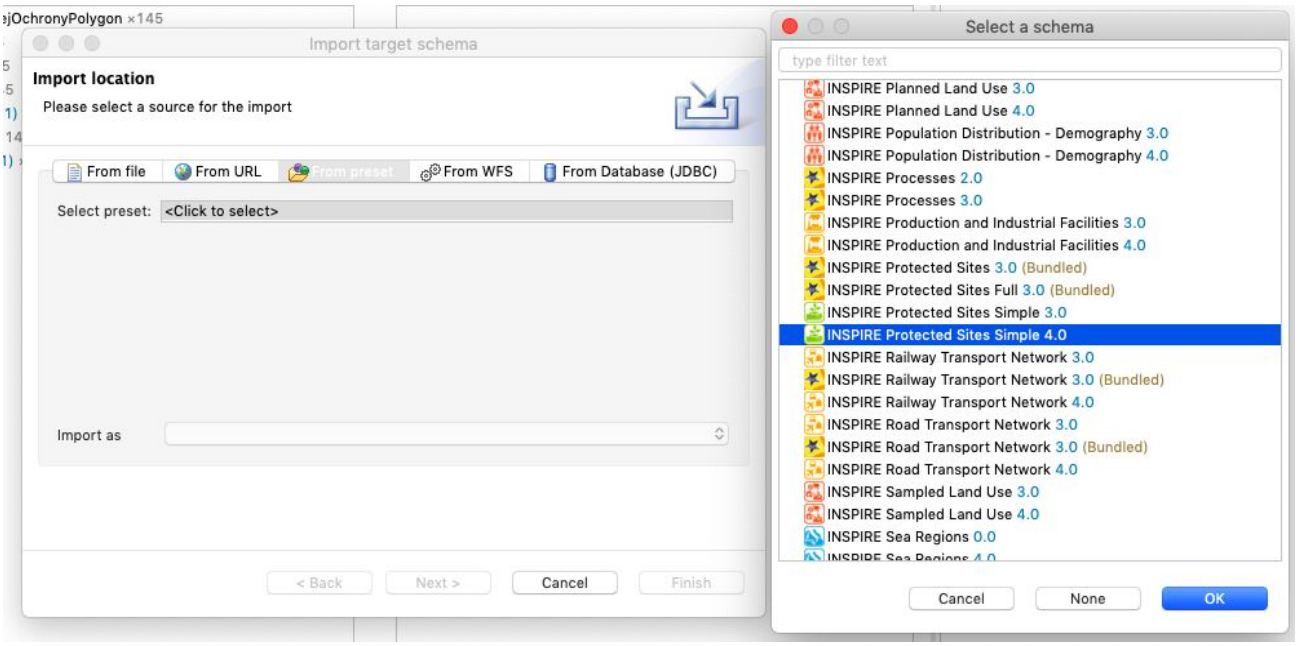



Kodowanie znaków należy wybrać ponownie jako windows-1250 i zatwierdzić przez **Finish**. Panel **Schema Explorer** powinien wyświetlić liczbę obiektów.

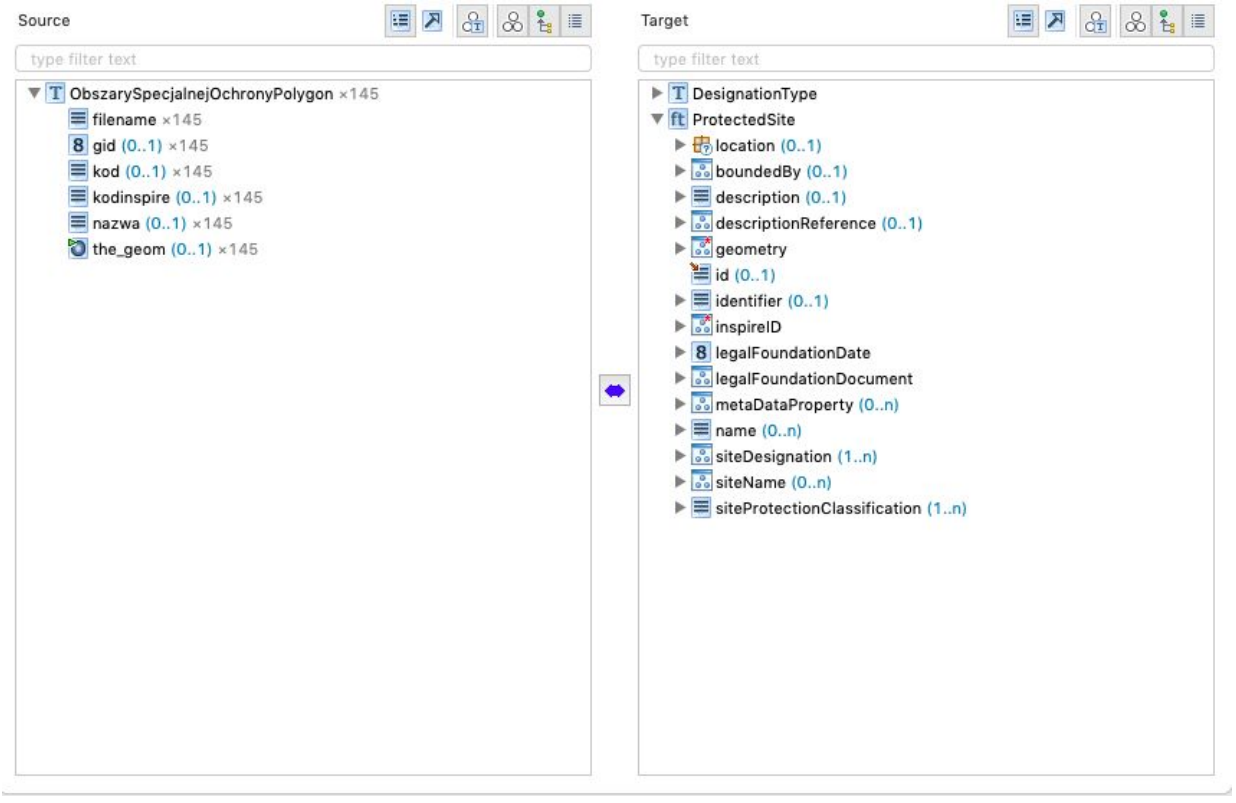



## Import schematu wynikowego

Należy wybrać z menu **Import -> Target schema,** zakładkę **From preset** i z listy dostępnych schematów - **INSPIRE Protected Sites Simple 4.0.**



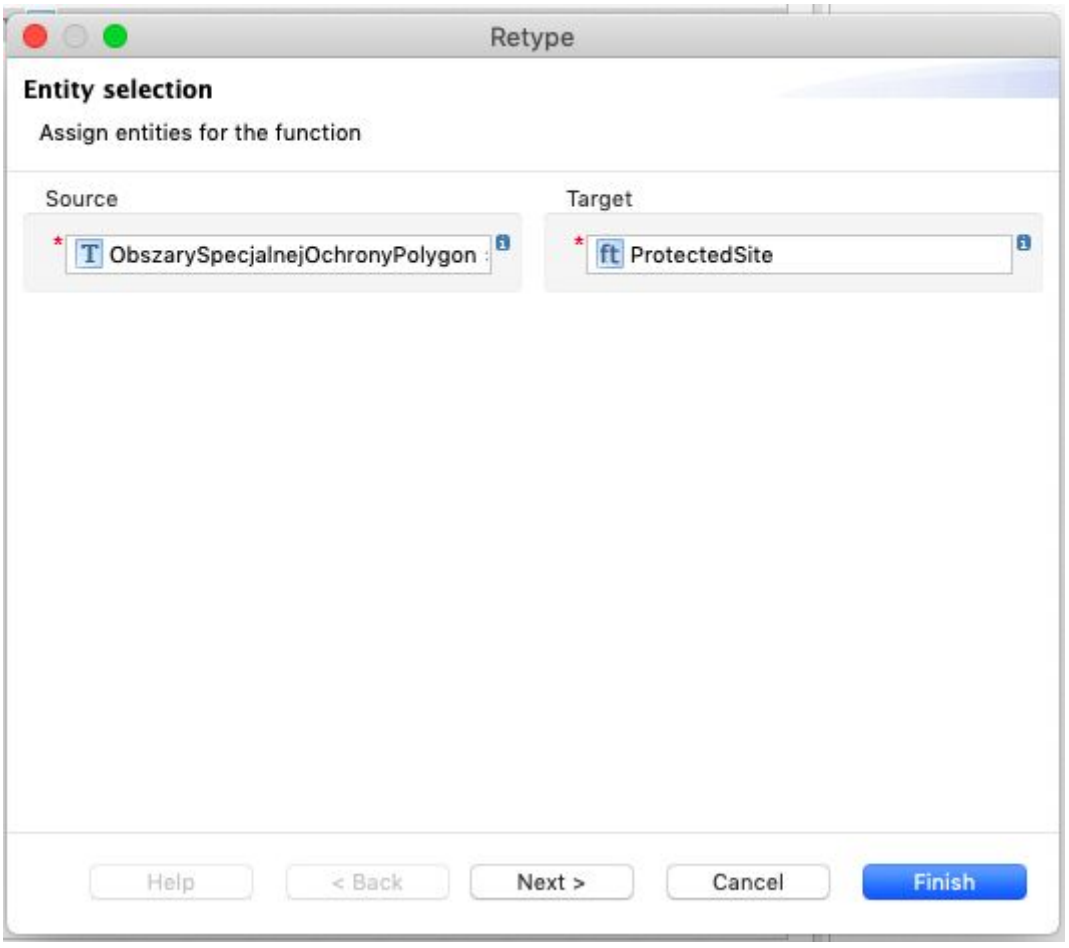

Po zatwierdzeniu schemat docelowy stanie się widoczny w panelu Schema Explorer.




## Utworzenie mapowania

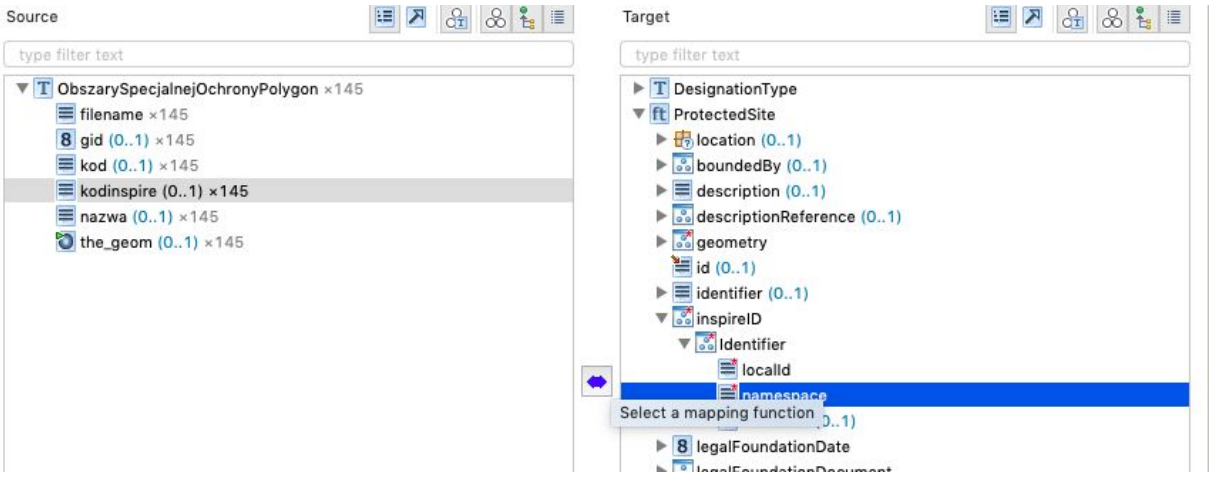
W pierwszej kolejności należy zaznaczyć w schemacie źródłowym

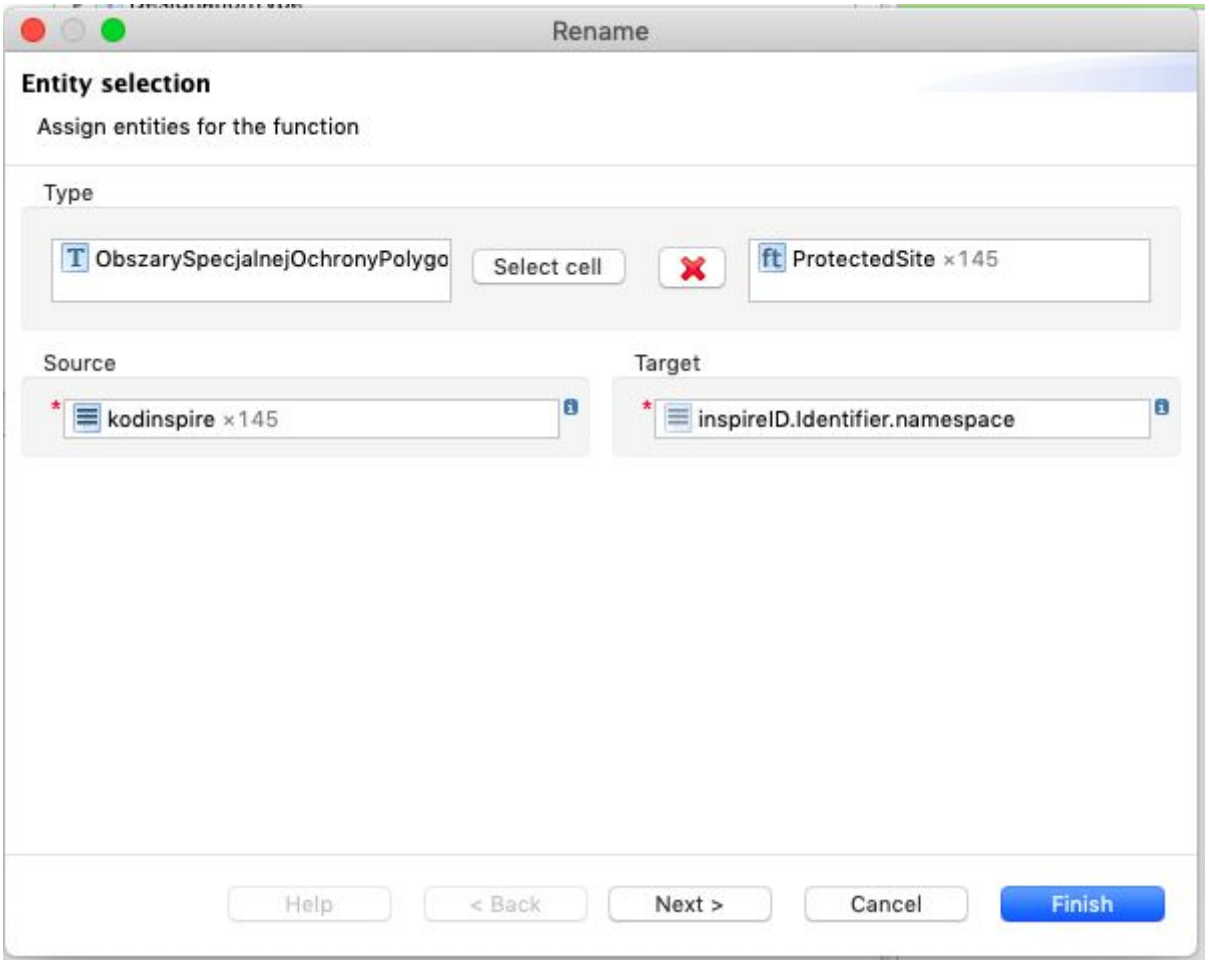
"**ObszarySpecjalnejOchronyPolygon"**, a w wynikowym - "**ProtectedSite"** i kliknąć  , a z menu kontekstowego wybrać funkcję **Retype.**




Mapowanie zatwierdzić poprzez **Finish.**

Następnie można przystąpić do mapowania atrybutów. Należy zaznaczyć w schemacie źródłowym pole **"kodinspire"**, a w schemacie wynikowym "**ProtectedSite > inspireID -> Identifier -> namespace"** , a następnie kliknąć przycisk  i wybrać funkcję **Rename**.



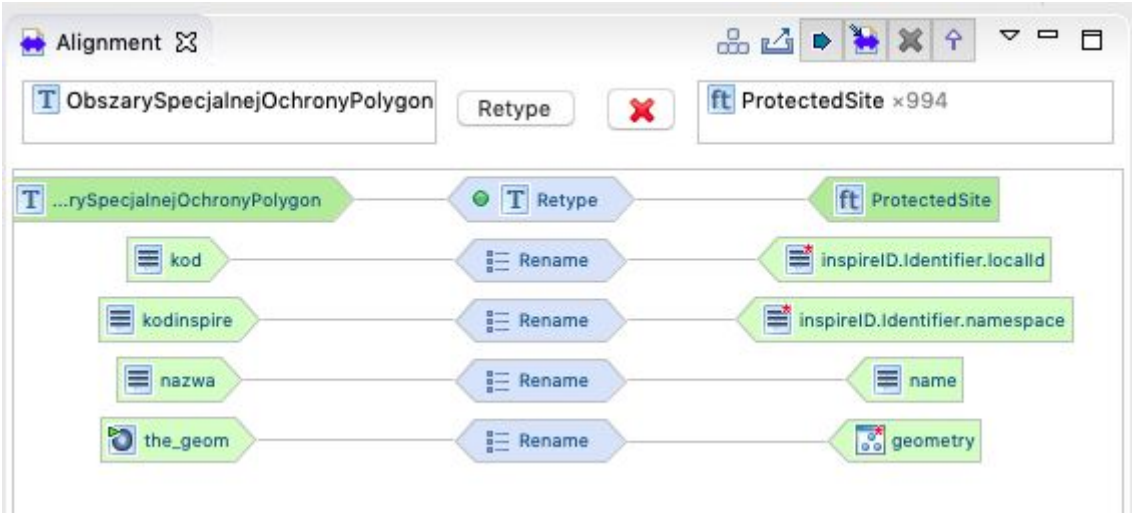




Czynność należy powtórzyć dla następujących mapowań:

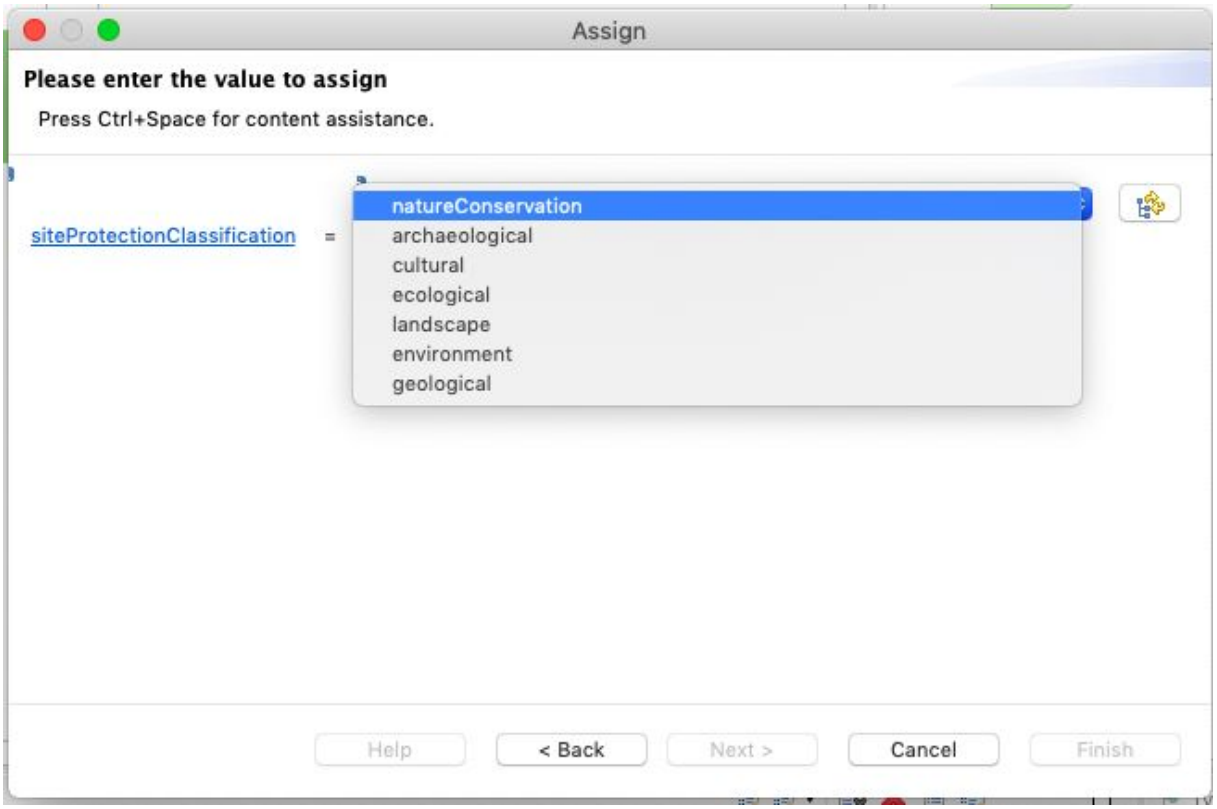
**kod** na **ProtectedSite -> inspireID -> Identifier -> localID**

**nazwa** na **ProtectedSite -> name**
**the_geom** na **ProtectedSite-> geometry**

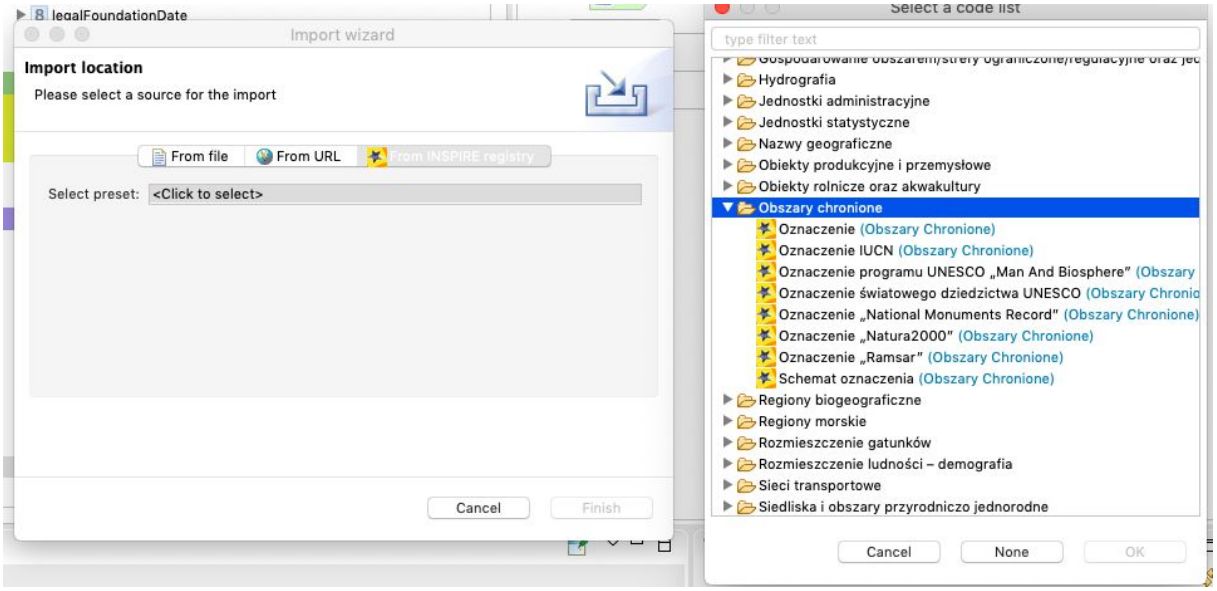
Panel **Alignment** powinien wyglądać następująco:




Następnie należy przystąpić do przypisania wartości stałych. W schemacie źródłowym nie może pozostać zaznaczony żaden atrybut - należy w tym celu kliknąć w białą przestrzeń w panelu **Source** przedstawiającym schemat źródłowy i upewnić się, że ikona  pomiędzy panelami uległa zmianie na  . Następnie w schemacie wynikowym należy zaznaczyć pole **siteProtectionClassification** i kliknąć  , a z listy dostępnych funkcji wybrać **Assign.** Następnie z listy dostępnych wartości wybrać **natureConservation.**

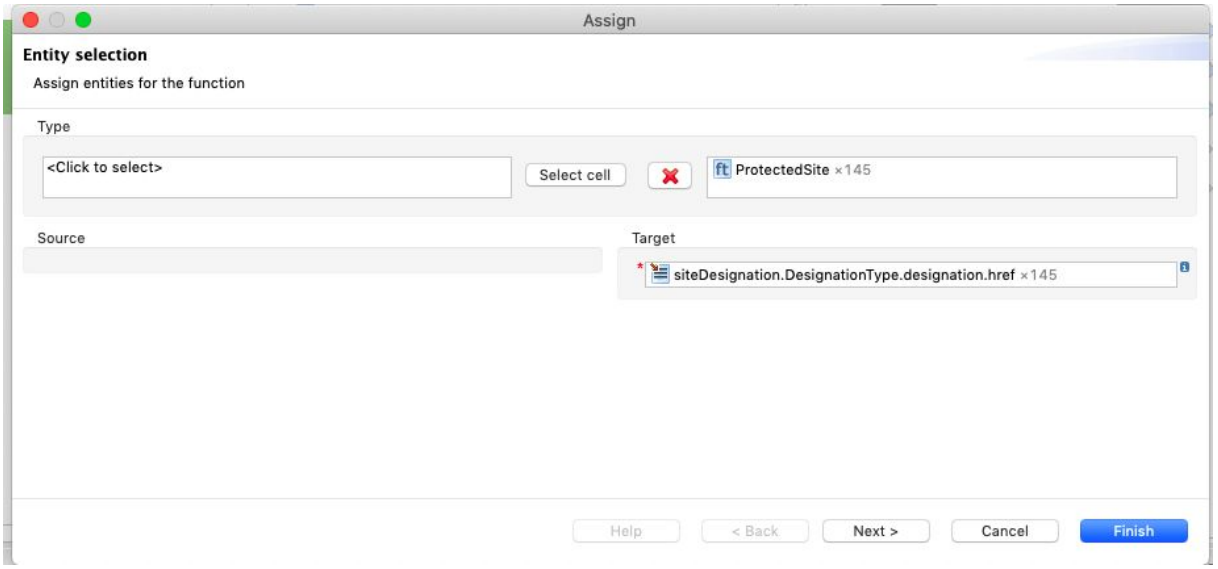



Atrybut **SiteDesignation** należy uzupełnić z wykorzystaniem list kodowych (słowników) INSPIRE. W tym celu należy je importować do HALE: **File -> Import -> Code list…** a następnie wybrać **From INSPIRE Registry** i **Obszary chronione:**

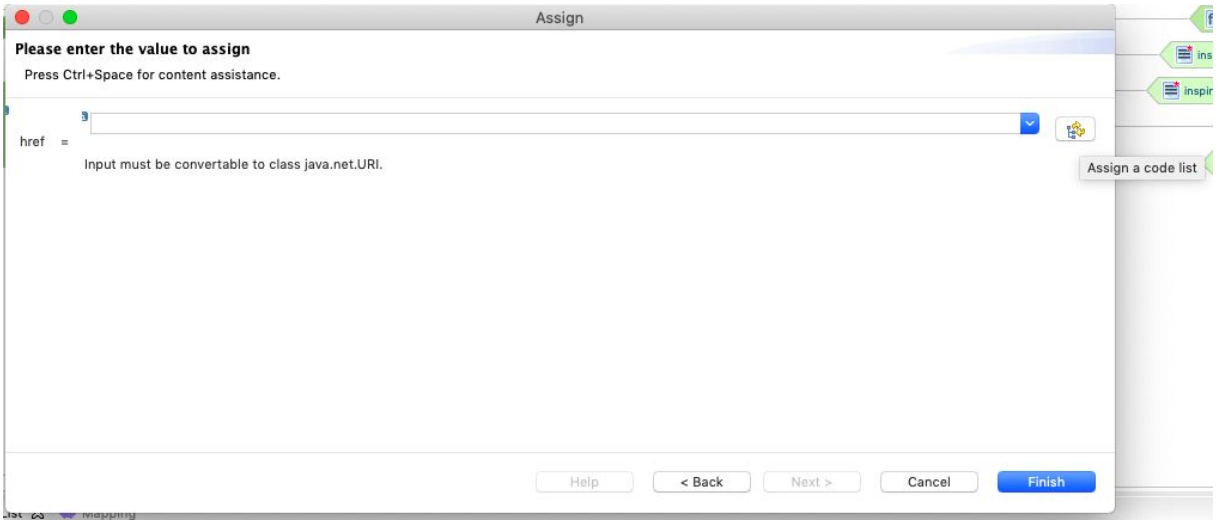



Do wykonania ćwiczenia potrzebne są następujące listy kodowe: **Oznaczenie, Oznaczenie "Natura 2000" i Schemat oznaczenia.** Każdą należy importować osobno.

Następnie należy zmapować atrybut **ProtectedSite -> siteDesignation** -> **DesignationType -> designation -> href** z użyciem funkcji **Assign**:

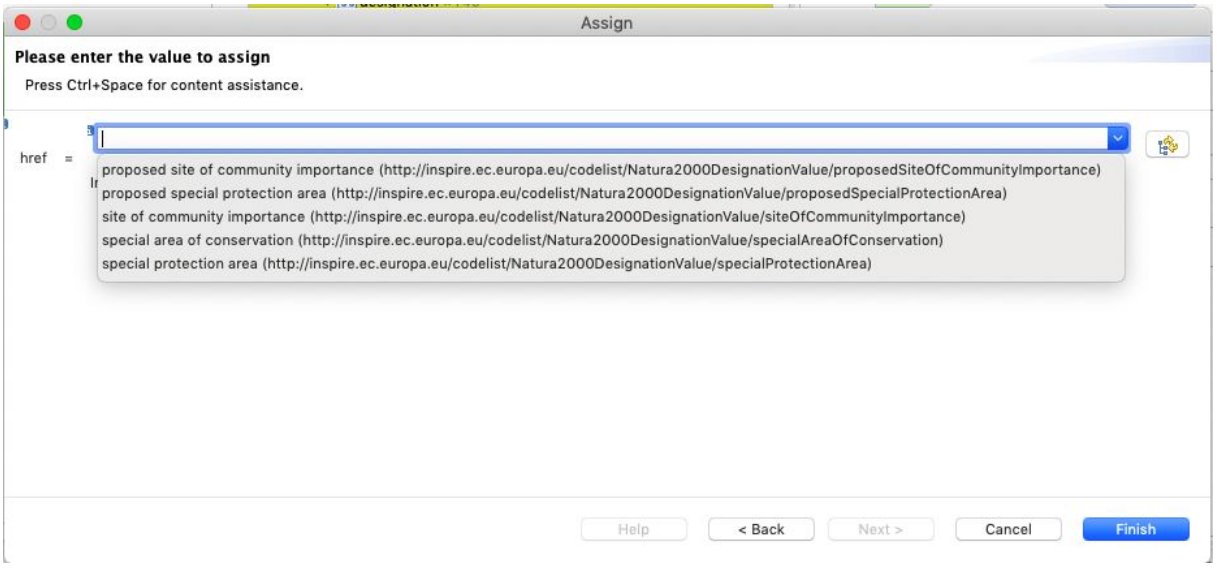



po potwierdzeniu atrybutu docelowego należy kliknąć **Next**. Pojawi się okno do wpisania wartości:

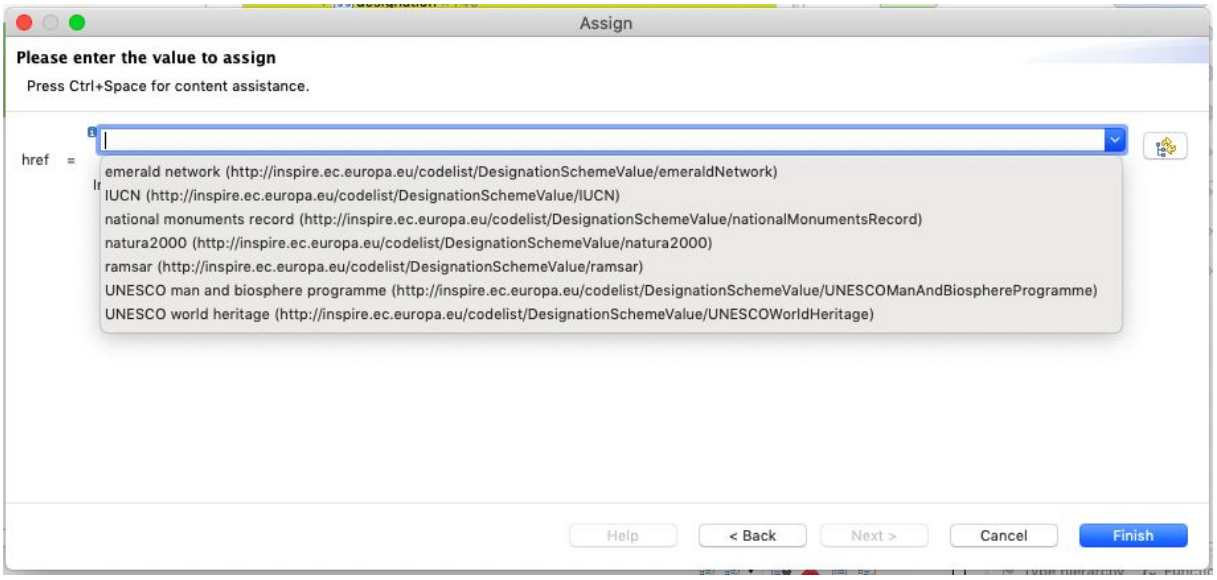



Dostęp do list kodowych uzyskuje się po naciśnięciu przycisku **Assign a code list**  znajdującego się po prawej stronie pola tekstowego.


Należy wybrać listę "**Natura2000DesignationValue**". W oknie do podania wartości należy rozwinąć listę przy użyciu przycisku  i wybrać wartość **"specialProtectionArea".**

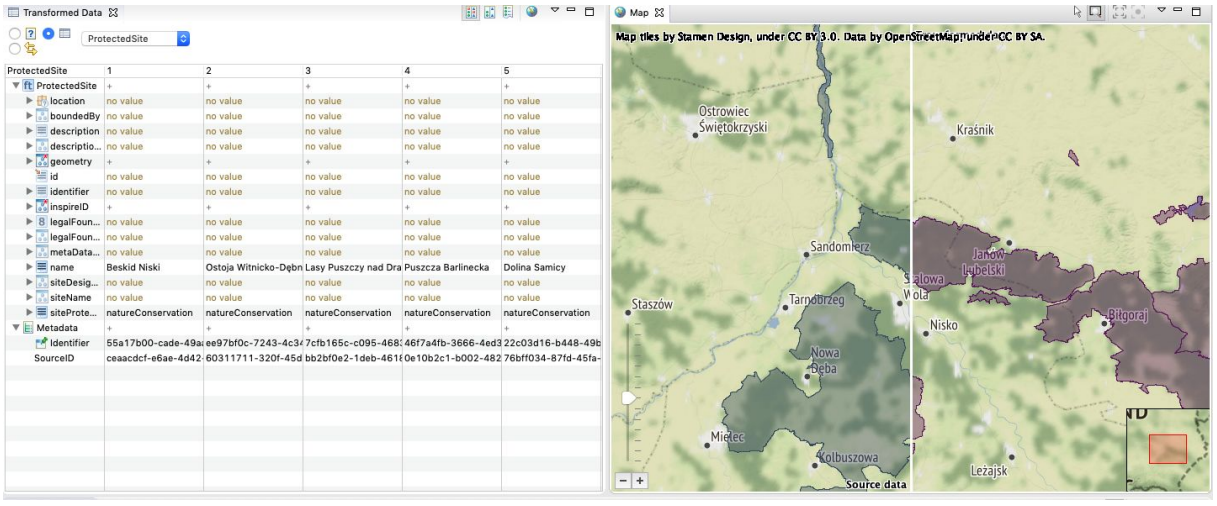


Następnie należy - postępując analogicznie jak w poprzednim przykładzie - zmapować atrybut **ProtectedSite -> siteDesignation -> DesignationType -> designationScheme -> href ,** wybrać listę kodową **DesignationScheme** i wartość **"natura2000".**



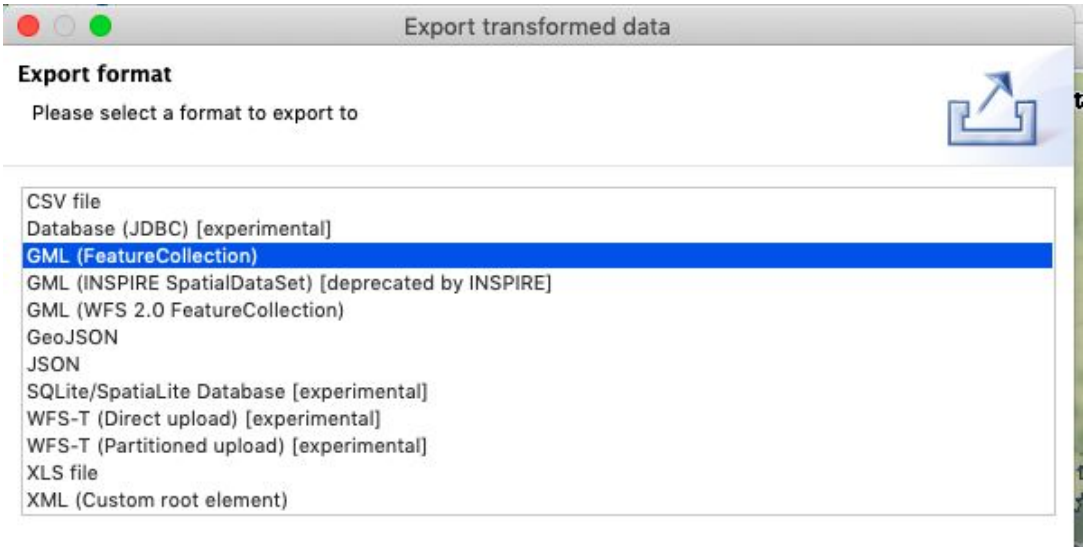

## Ocena wyników

Kliknięcie ikony  w prawym górnym rogu ekranu pozwala zwizualizować wynik transformacji na podkładzie mapowym. Narzędzie służy inspekcji wizualnej wyników transformacji.

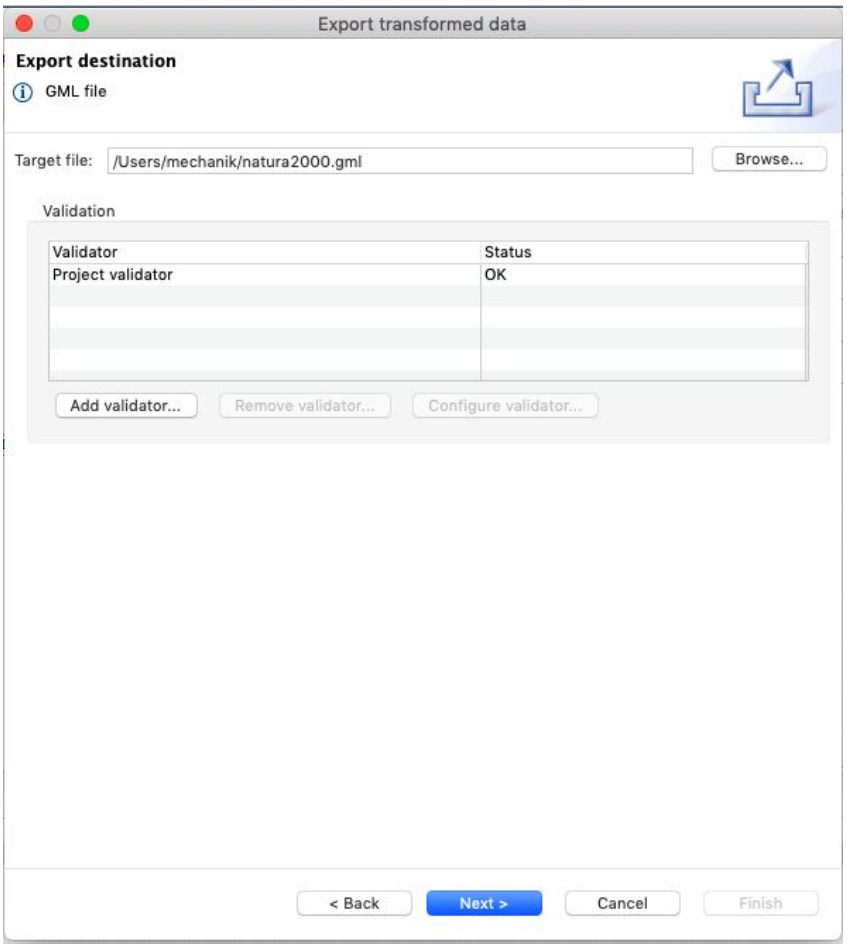


## Eksport danych

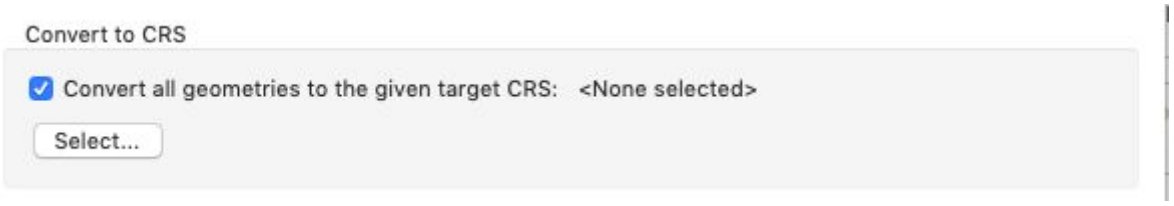
Należy wybrać z menu **File -> Export -> Transformed data,** a jako format **GML (FeatureCollection)**.Wskazany format zapewnia zgodność z dyrektywą Inspire i zastępuje stosowany do 2016 r. GML (INSPIRE SpatialDataSet).



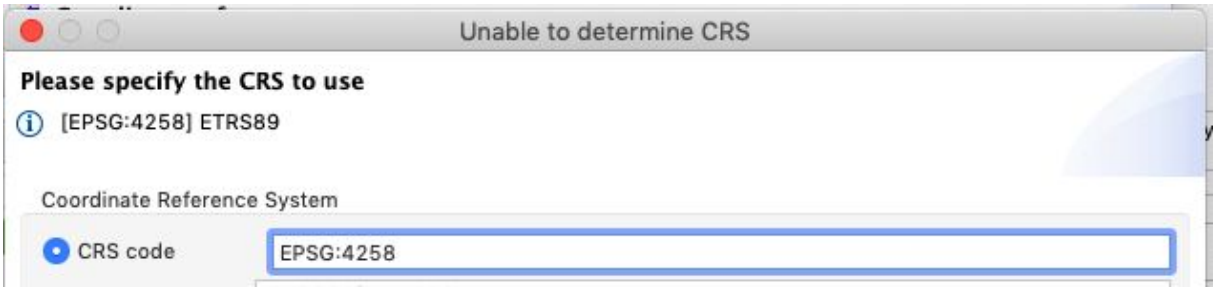

w kolejnym kroku należy wybrać lokalizację eksportowanego pliku:

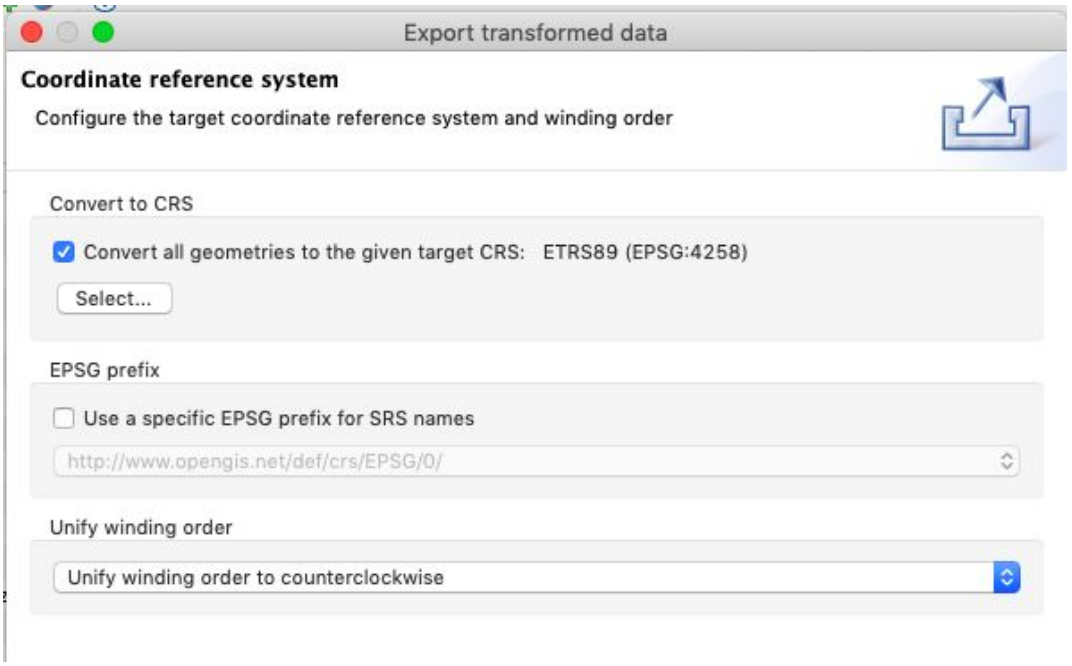



a następnie wybrać układ współrzędnych: zaznaczyć pole "Convert all geometries to the given target CRS". Zaznaczenie tej opcji spowoduje transformację źródłowego układu współrzędnych do układu współrzędnych wybranego jako wyjściowy.




i po kliknięciu przycisku **Select** podać kod: EPSG:4258

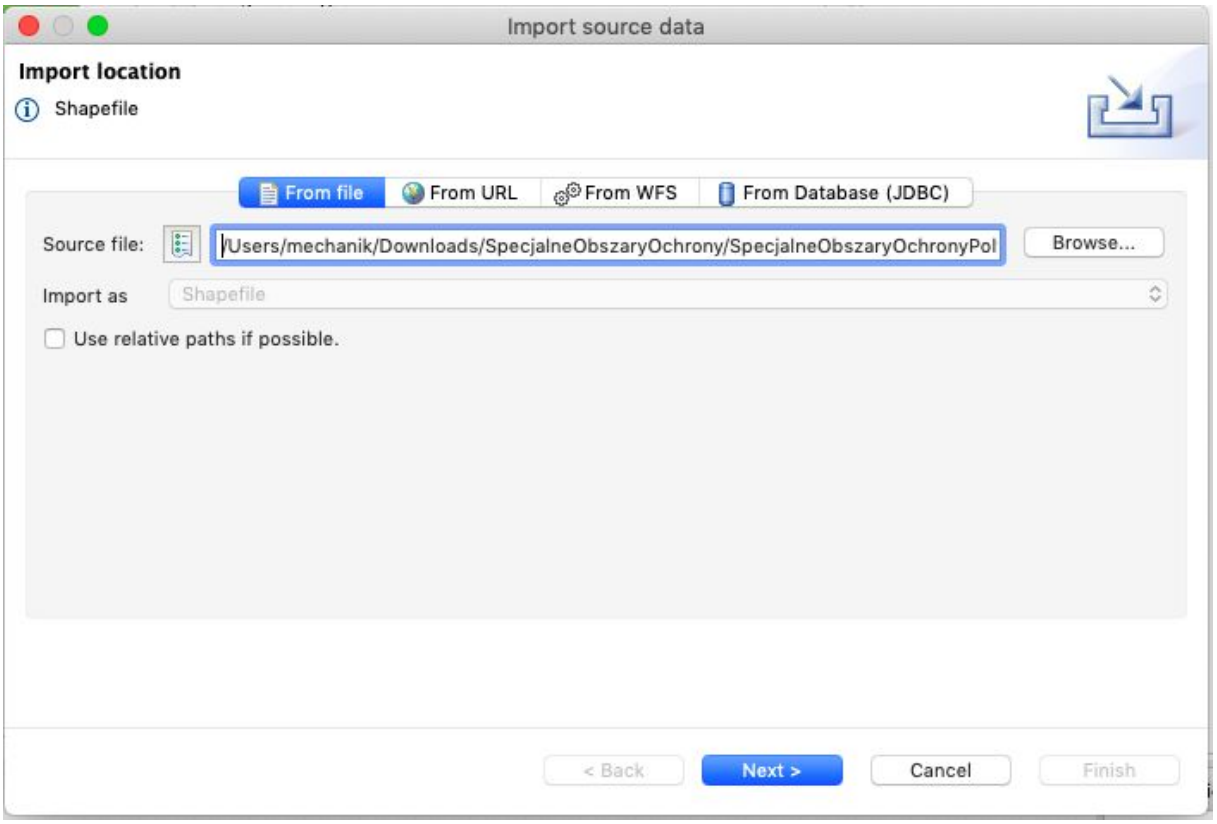




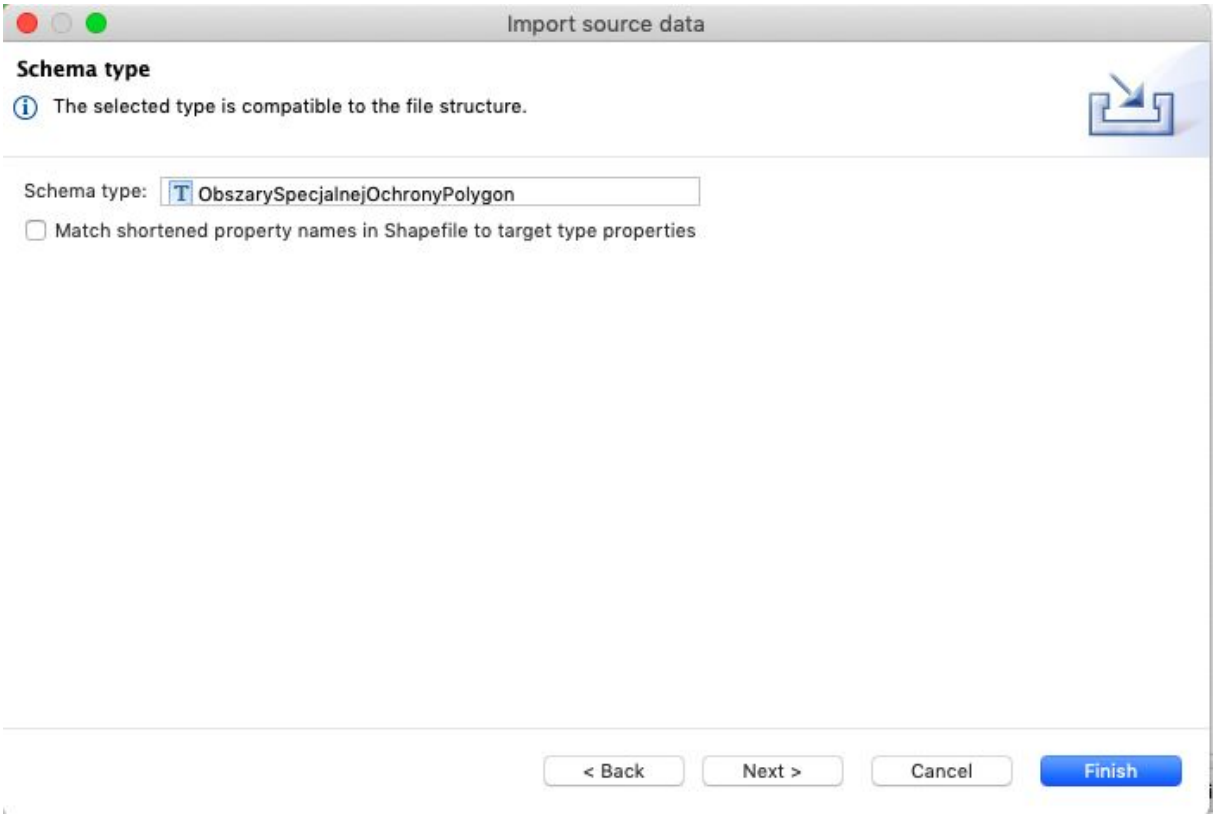

Po udanym eksporcie projekt należy zapisać: **File -> Save Alignment Project** w formacie **"hale project"** jako plik "**cwiczenie2.hale"**, a wynikowy plik GML zweryfikować w QGIS.

## Ćwiczenie 2: Mapowanie danych z kilku źródeł jednocześnie

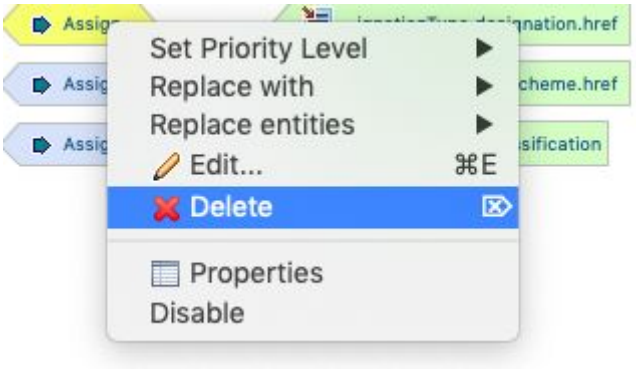
Korzystając z projektu utworzonego w Ćwiczeniu 1, należy dodać kolejny zbiór danych źródłowych - **Import -> Source data.** Należy załadować plik **SpecjalneObszaryOchronyPolygon.shp** z katalogu HALE/masowa_konwersja w danych do ćwiczeń.



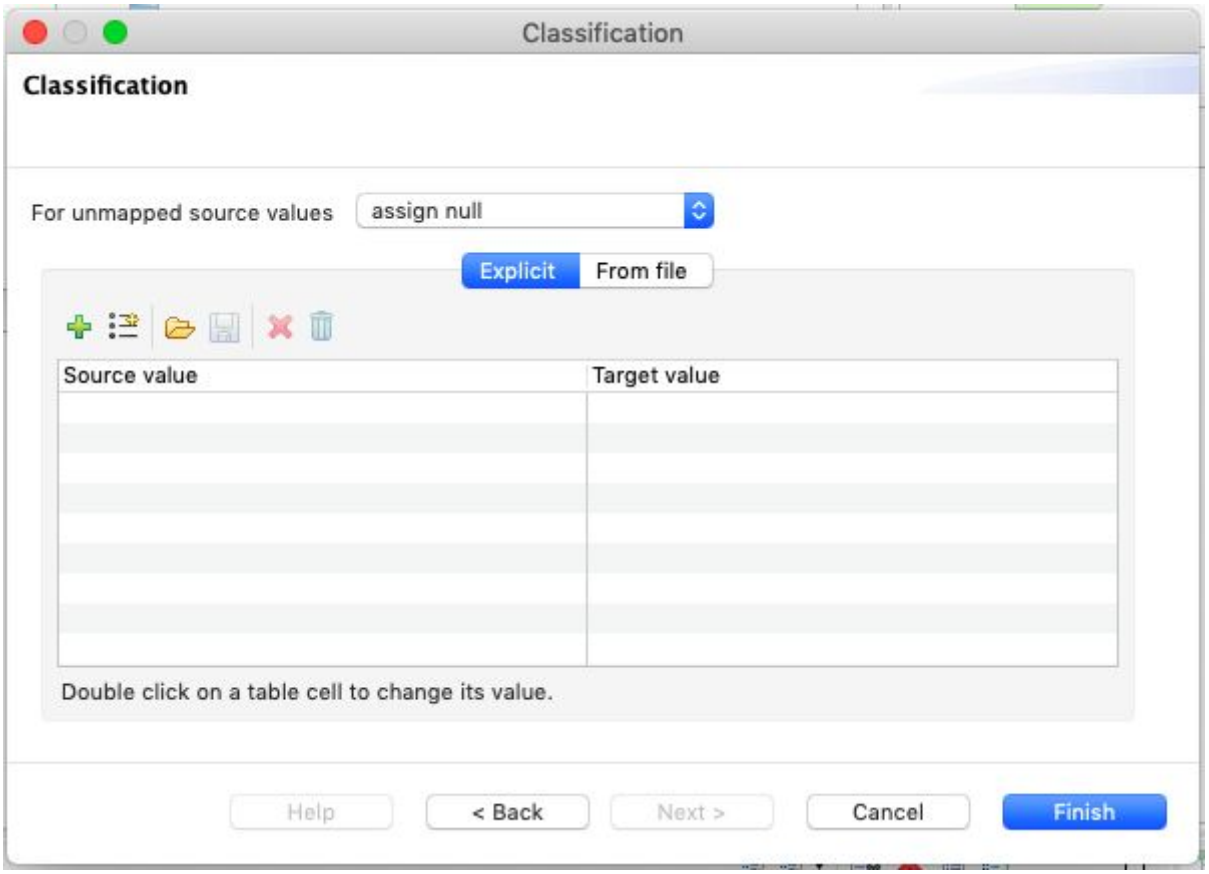

Jako schemat docelowy należy wybrać **ObszarySpecjalnejOchronyPolygon.**



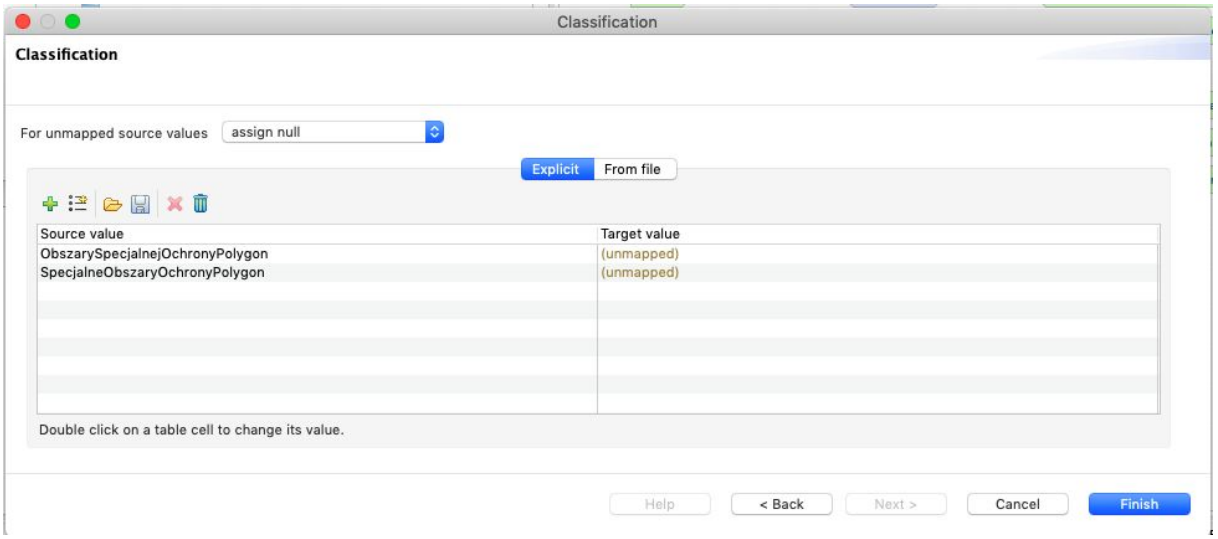
Następnie należy usunąć mapowanie atrybutu **ProtectedSite -> siteDesignation -> DesignationType -> designation -> href**, poprzez kliknięcie prawym klawiszem na **Assign** i wybór **Delete:**




Nowe mapowanie będzie uwzględniało rozróżnienie na OSO i SOO. W tym celu należy zaznaczyć po stronie źródłowej atrybut **filename**, a po stronie wynikowej **ProtectedSite -> siteDesignation -> DesignationType -> designation -> href ,** kliknąć  i wybrać funkcję **Classification.**

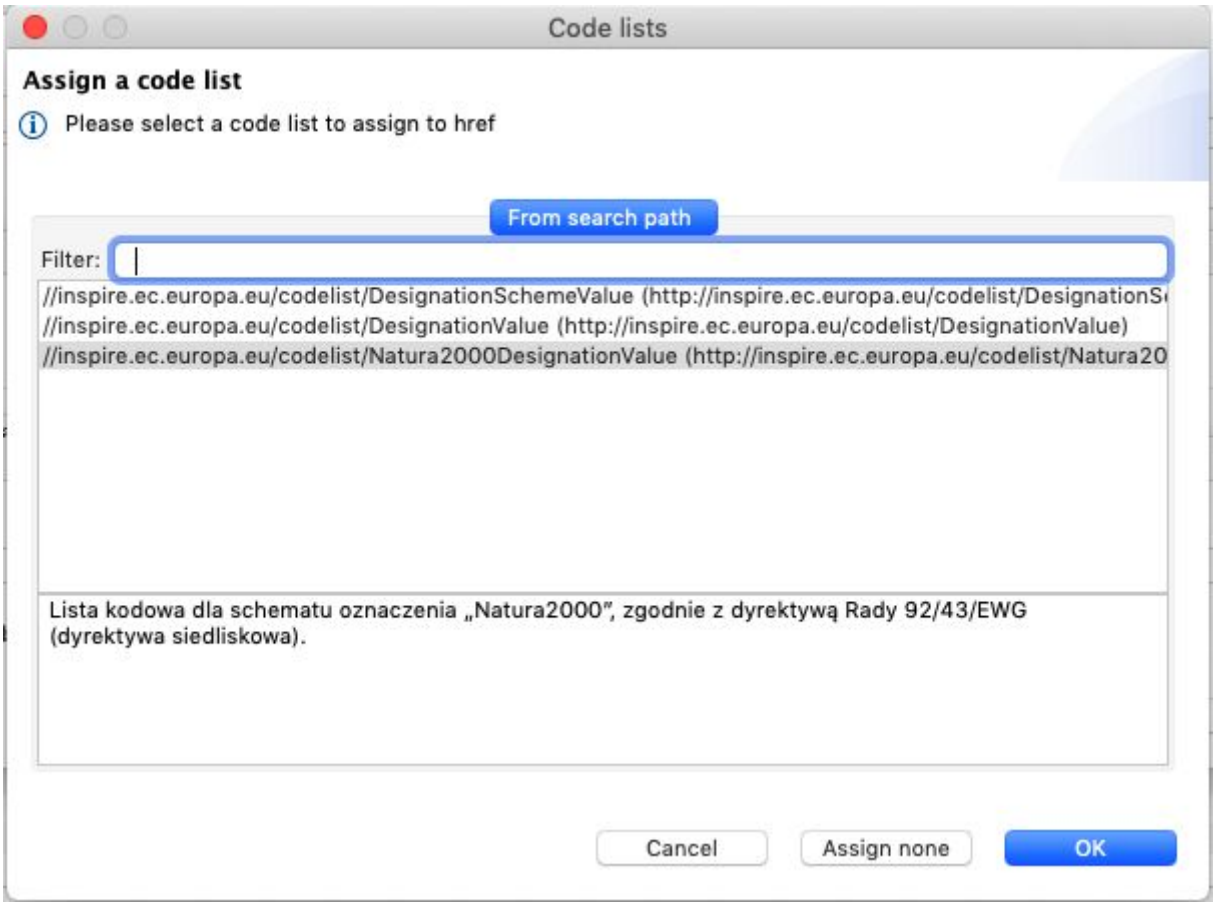



Za pomocą przycisku  należy zaimportować istniejące wartości w danych źródłowych:

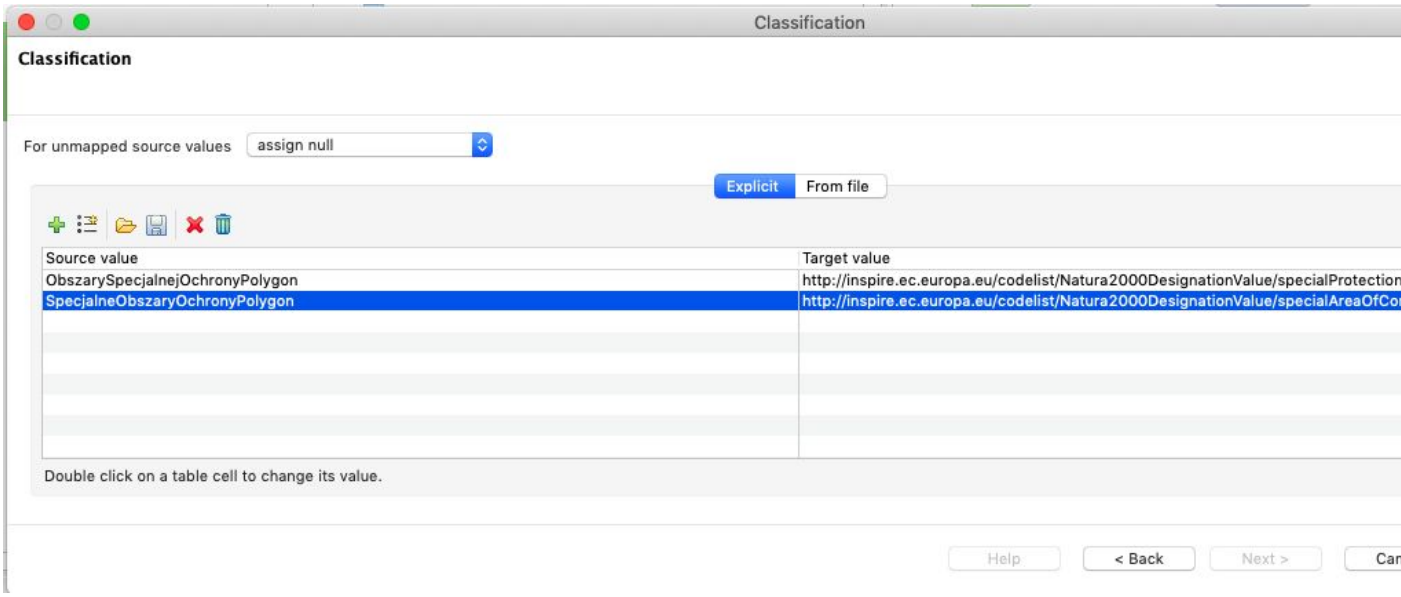



a następnie poprzez podwójne kliknięcie w **(unmapped)** w kolumnie **Target value**, wybrać wartości z listy kodowej Natura2000.

Jeśli pomiędzy wykonaniem Ćwiczenia 2 i Ćwiczenia 3 nie wybierano innej listy kodowej niż Natura 2000, nie ma potrzeby ponownego wskazywania tej listy za pomocą przycisku  i poniższego okna. W przeciwnym razie należy ponownie wybrać listę kodową Natura 2000.

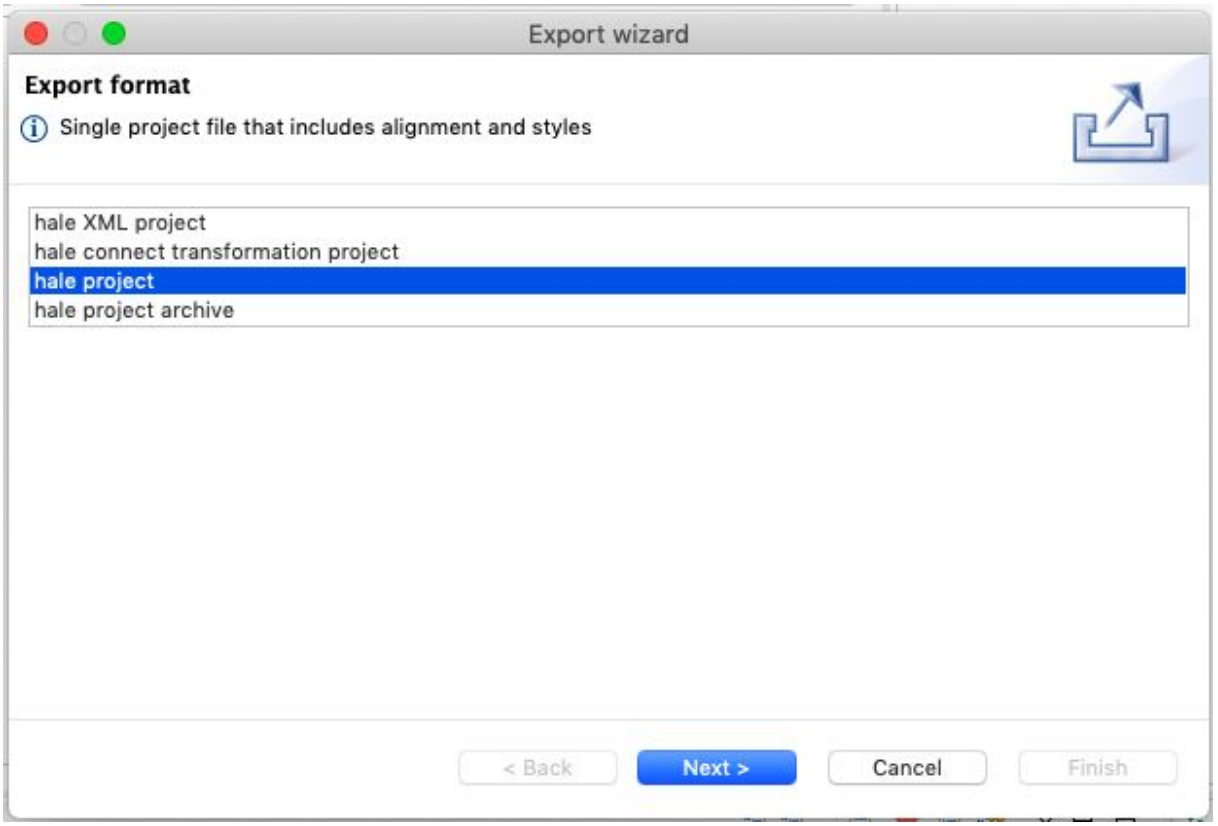



mapując **ObszarySpecjalnejOchronyPolygon** na **specialProtectionArea**, a **SpecjalneObszaryOchrony** na **specialAreaOfConservation**:

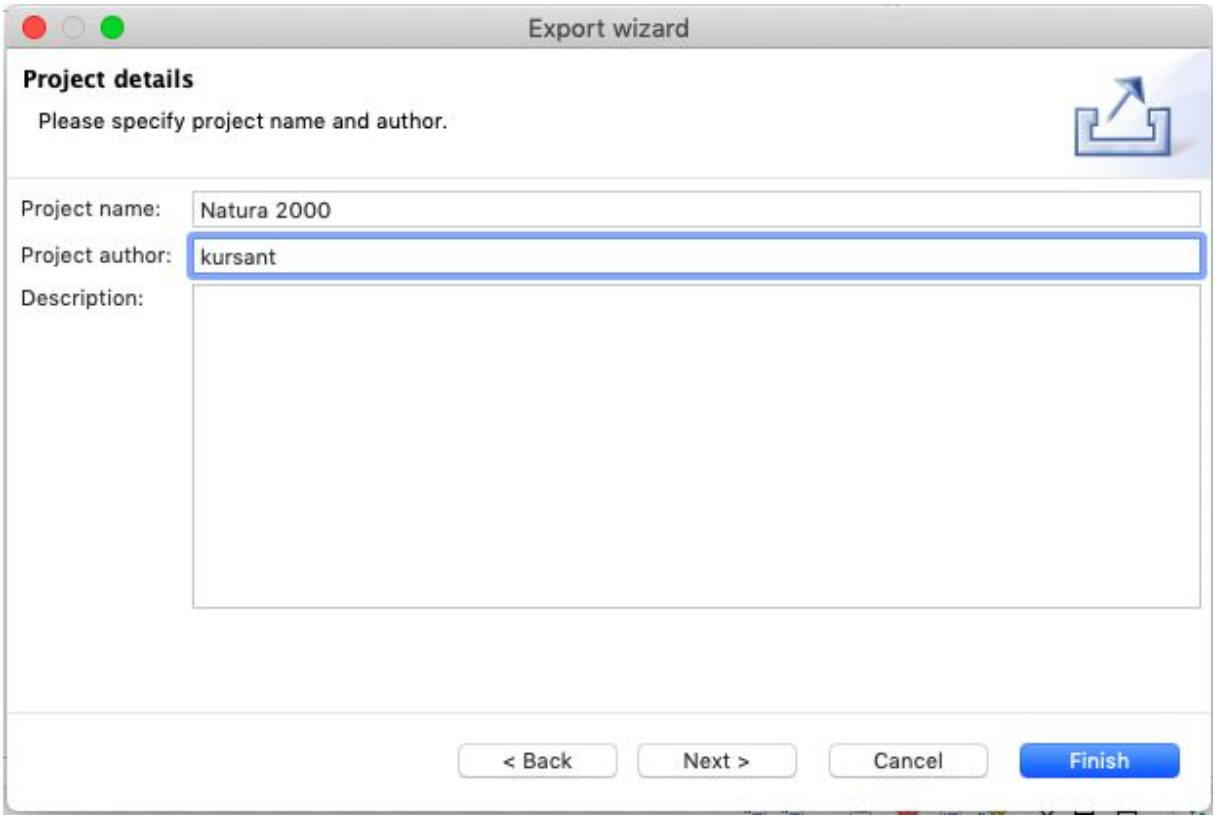


Po zakończeniu transformacji należy ponownie wyeksportować plik GML oraz zapisać projekt:

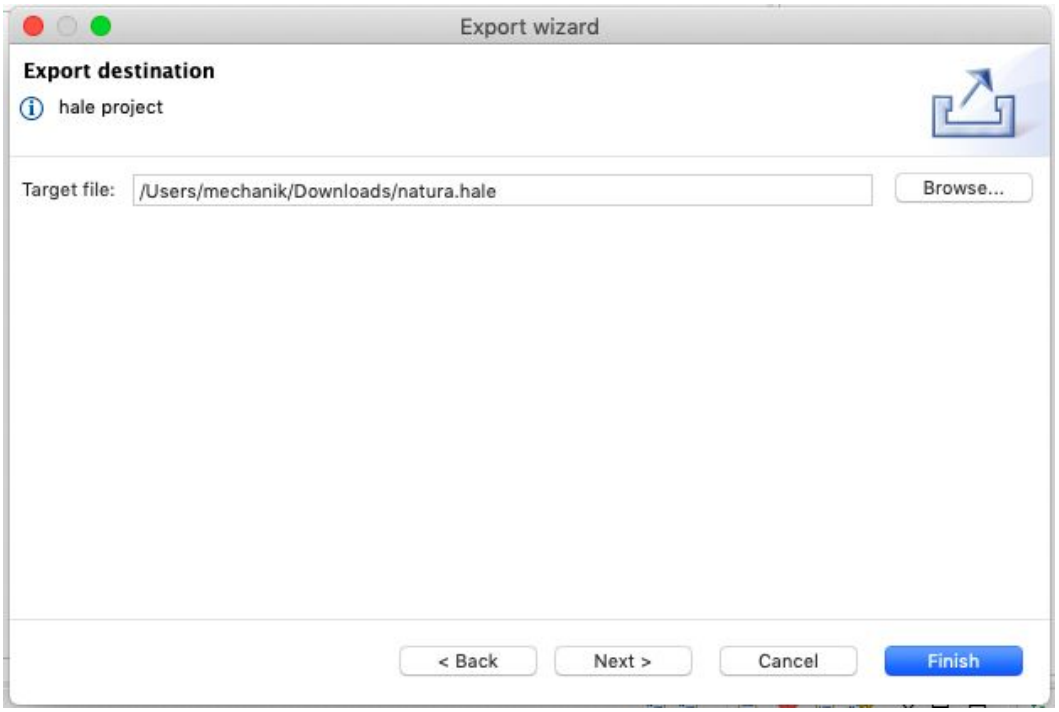
**File -> Save Alignment Project as...**




i nadać mu nazwę "Natura 2000".




jako plik wynikowy należy wybrać "natura.hale".



## Ćwiczenie 3: Wykorzystanie struktury i danych zapisanych w bazie danych

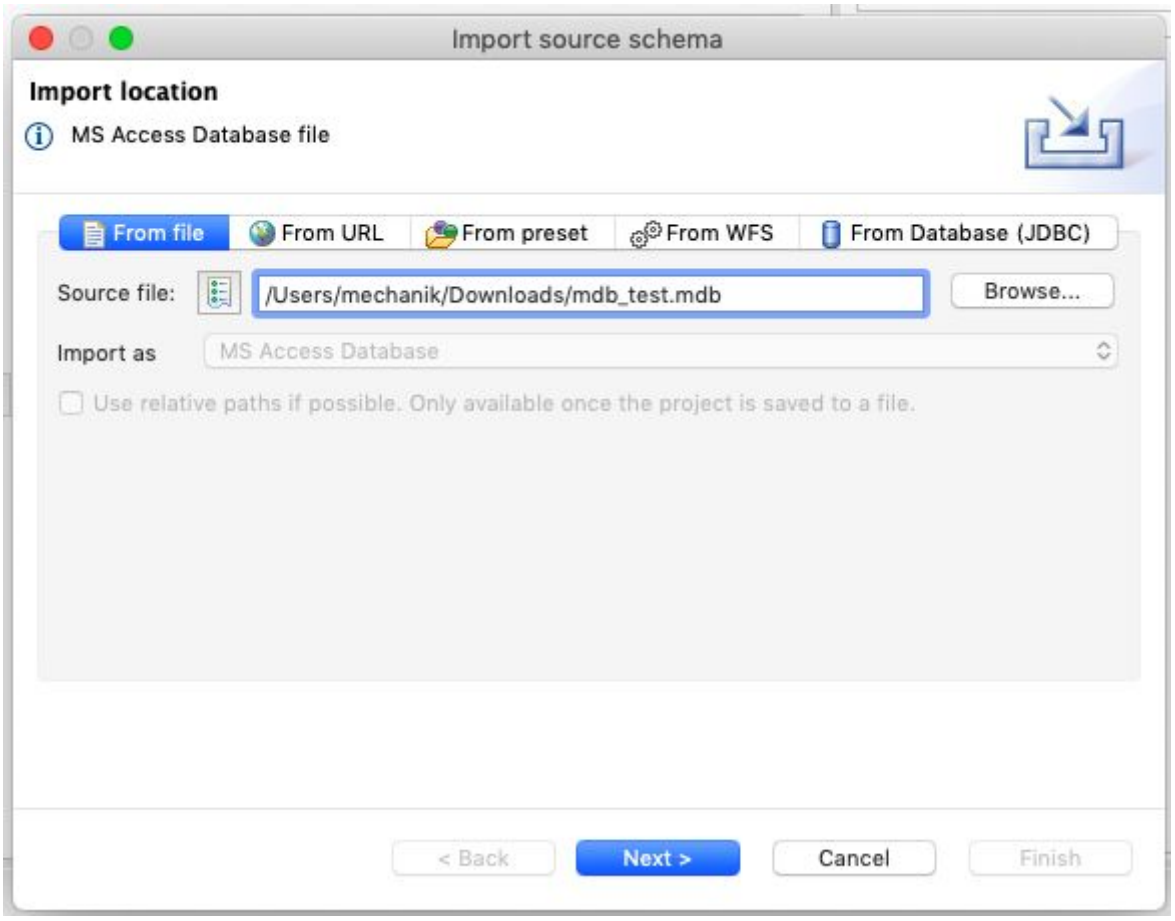
Celem ćwiczenia jest pozyskanie wiedzy praktycznej w zakresie wykorzystania struktur i danych zapisanych w bazach danych w celu ich transformacji. Proces składa się z kroków: Ćwiczenie ma na celu przejście procesu:

1. Importu schematu danych wejściowych z bazy danych
2. Importu danych źródłowych z bazy danych
3. Import schematu wyjściowego
4. Ustawienie reguł transformacji
5. Wykonanie transformacji
6. Eksport wyników przetworzenia

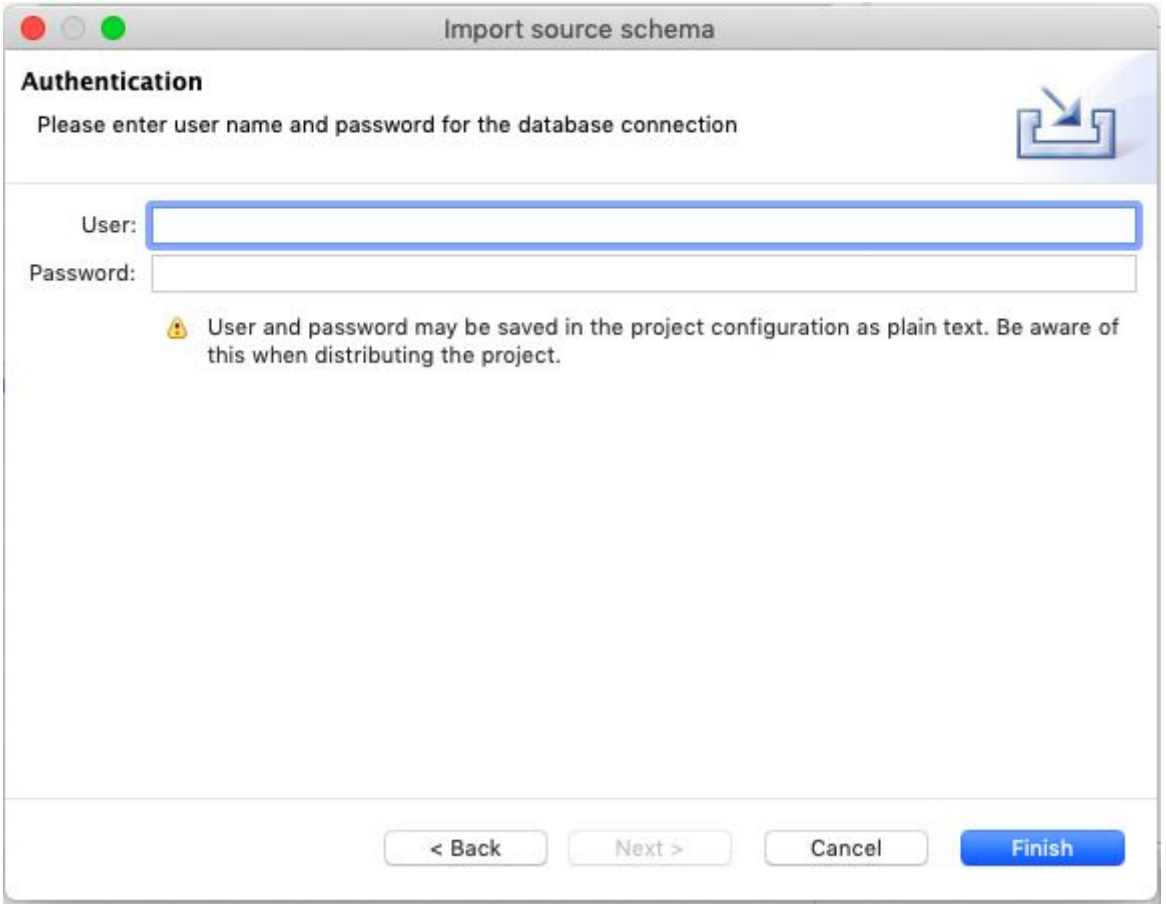
Pracę z bazą danych należy zacząć od utworzenia nowego projektu: **File -> New alignment project.** Widok programu zostanie wyczyszczony z istniejących schematów i mapowań.

## Geobaza ESRI MDB

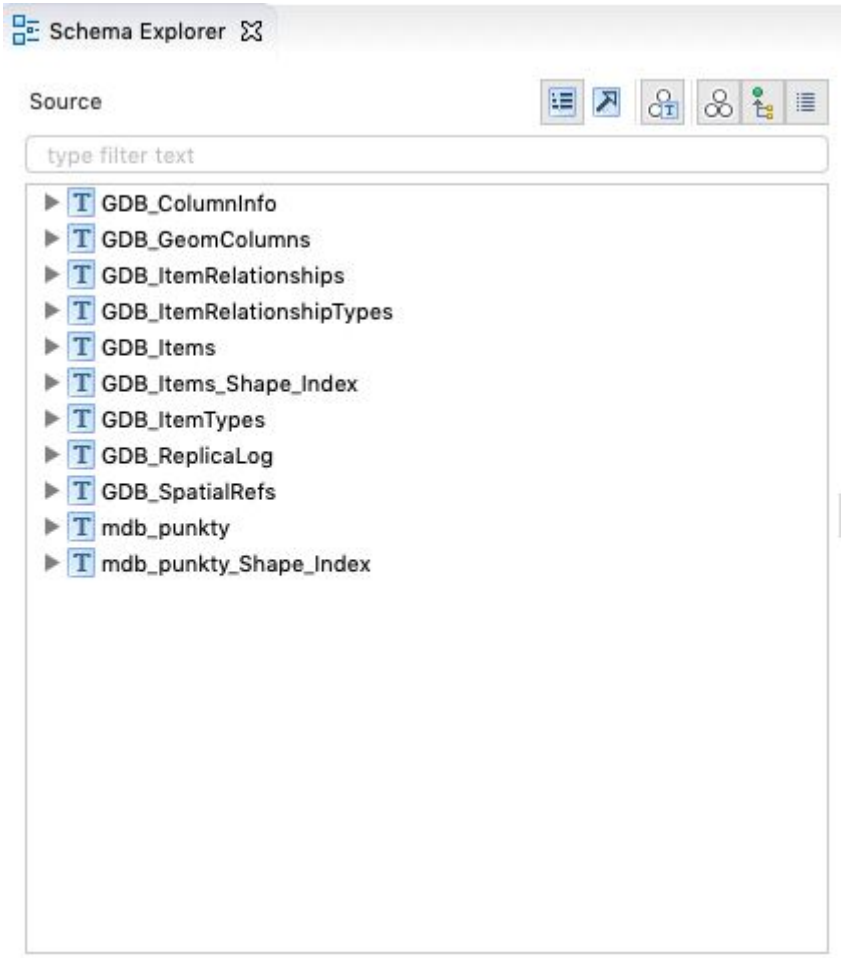
Należy wczytać bazę MDB jako schemat: **File -> Import -> Source schema,** wybrać plik "**mdb_test.mdb**". W polu Import as powinien pokazać się wpis **MS Access database.**



W następnym kroku nazwę użytkownika i hasło należy pozostawić puste. Kliknąć **Finish**.

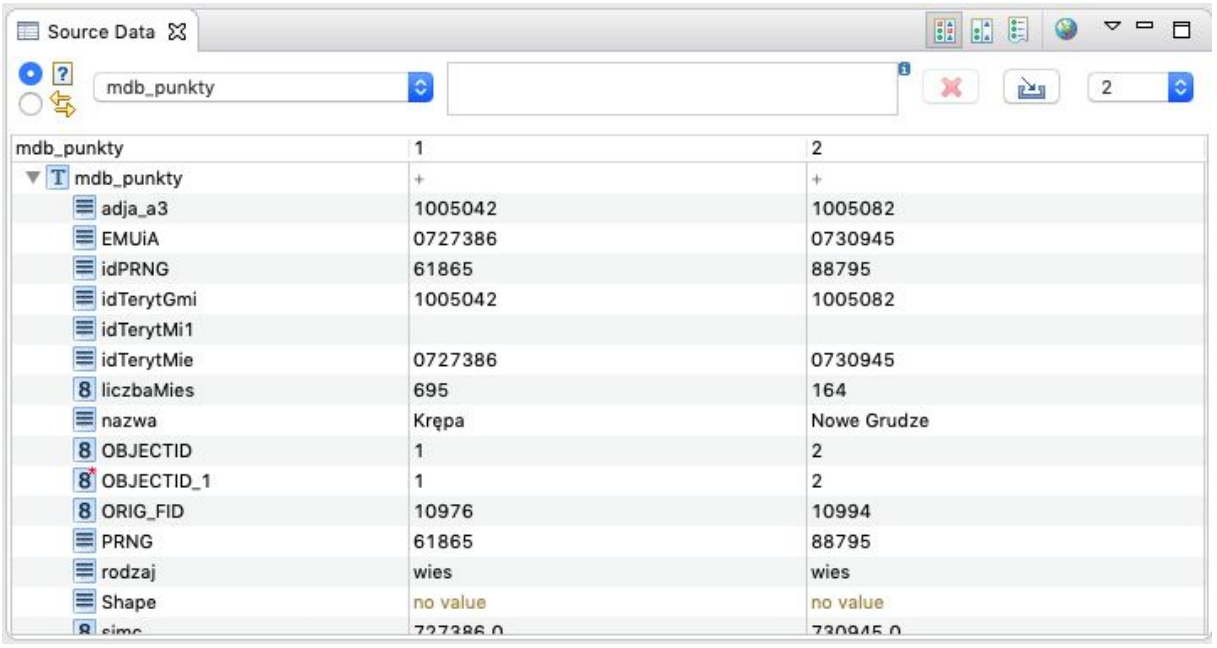



W panelu Schema Explorer pokaże się lista tabel z geobazy jako klas obiektów źródłowych.




Następnie należy ponownie wczytać bazę, tym razem jako źródło danych: **Import -> Source data.**

Dane powinny być widoczne w widoku **Data**.




| mdb_punkty | 1 | 2 |
|---|---|---|
| mdb_punkty | + | + |
| adja_a3 | 1005042 | 1005082 |
| EMUiA | 0727386 | 0730945 |
| idPRNG | 61865 | 88795 |
| idTerytGmi | 1005042 | 1005082 |
| idTerytMi1 | | |
| idTerytMie | 0727386 | 0730945 |
| liczbaMies | 695 | 164 |
| nazwa | Krępa | Nowe Grudze |
| OBJECTID | 1 | 2 |
| OBJECTID_1 | 1 | 2 |
| ORIG_FID | 10976 | 10994 |
| PRNG | 61865 | 88795 |
| rodzaj | wies | wies |
| Shape | no value | no value |
| simc | 727386.0 | 730945.0 |

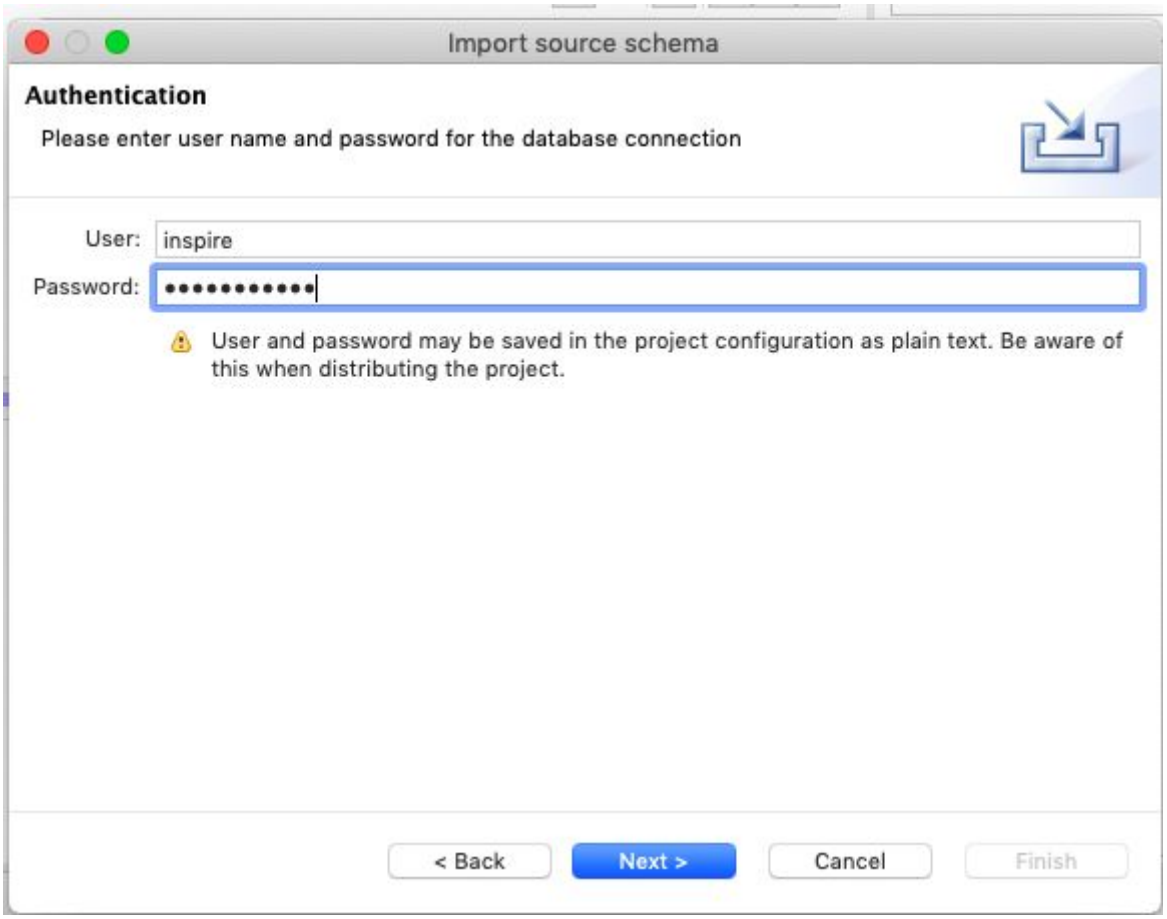
## Baza PostGIS

Należy utworzyć nowy projekt poprzez **File -> New alignment project.**
Następnie należy uruchomić narzędzie **File -> Import -> Source schema,** wybrać opcję **From database (JDBC)**.

Parametry połączenia są następujące:

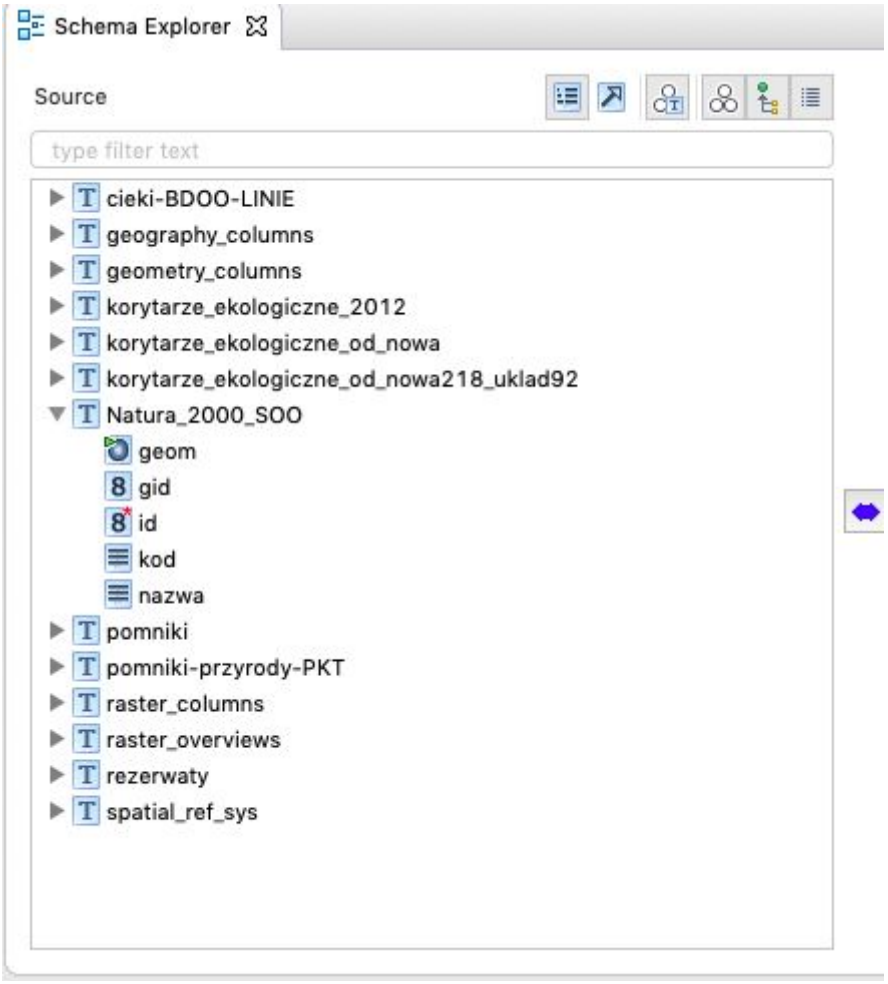
Driver: PostgreSQL / PostGIS
Host:Port: dragon.gis-support.pl:5432
Database: inspire
Import as: Database schema

Po kliknięciu **Next** należy się autoryzować jako użytkownik "inspire", hasło zostanie podane na szkoleniu.



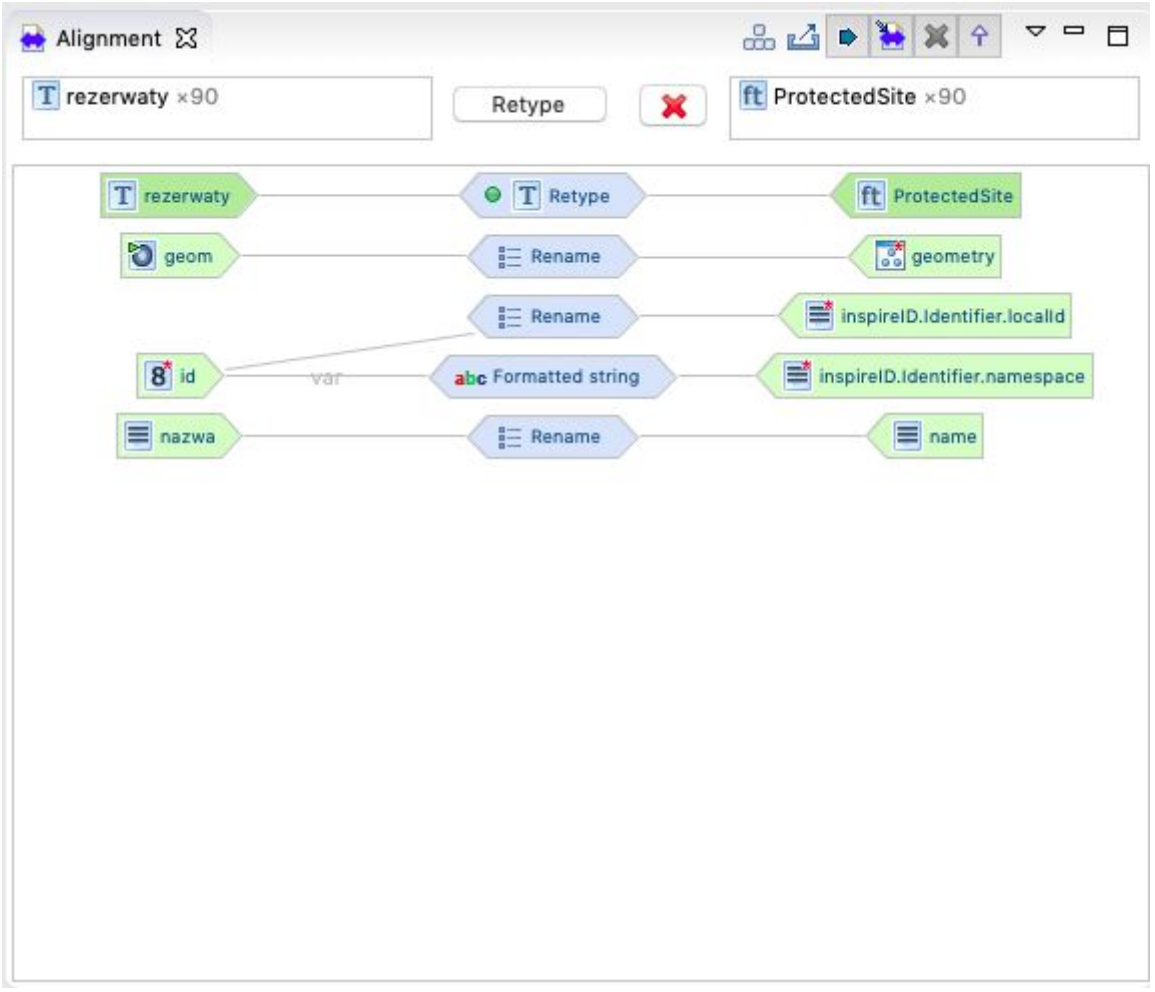

W następnym kroku należy wybrać schemat "**public**".

Po imporcie schematu powinna pokazać się lista dostępnych tabel.




Następnie należy ponowić operację połączenia z bazą, wybierając tym razem import danych: **Import -> Source data -> From database.**
W następnej kolejności należy przeprowadzić mapowanie tabeli **rezerwaty** na klasę **ProtectedSite** z schematu **INSPIRE Protected Sites Simple 4.0**, wykorzystując następujące mapowania:

| **rezerwaty** | **Funkcja** | **ProtectedSite** |
|---|---|---|
| data_utw | Rename | legalFoundationDate |
| geom | Rename | geometry |
| id | Rename | inspireID.Identifier.localId |
| id | Formatted string - PL.RP.{id} | inspireID.Identifier.namespace |
| nazwa | Rename | name |
| | Assign - natureConservation | siteProtectionClassification |



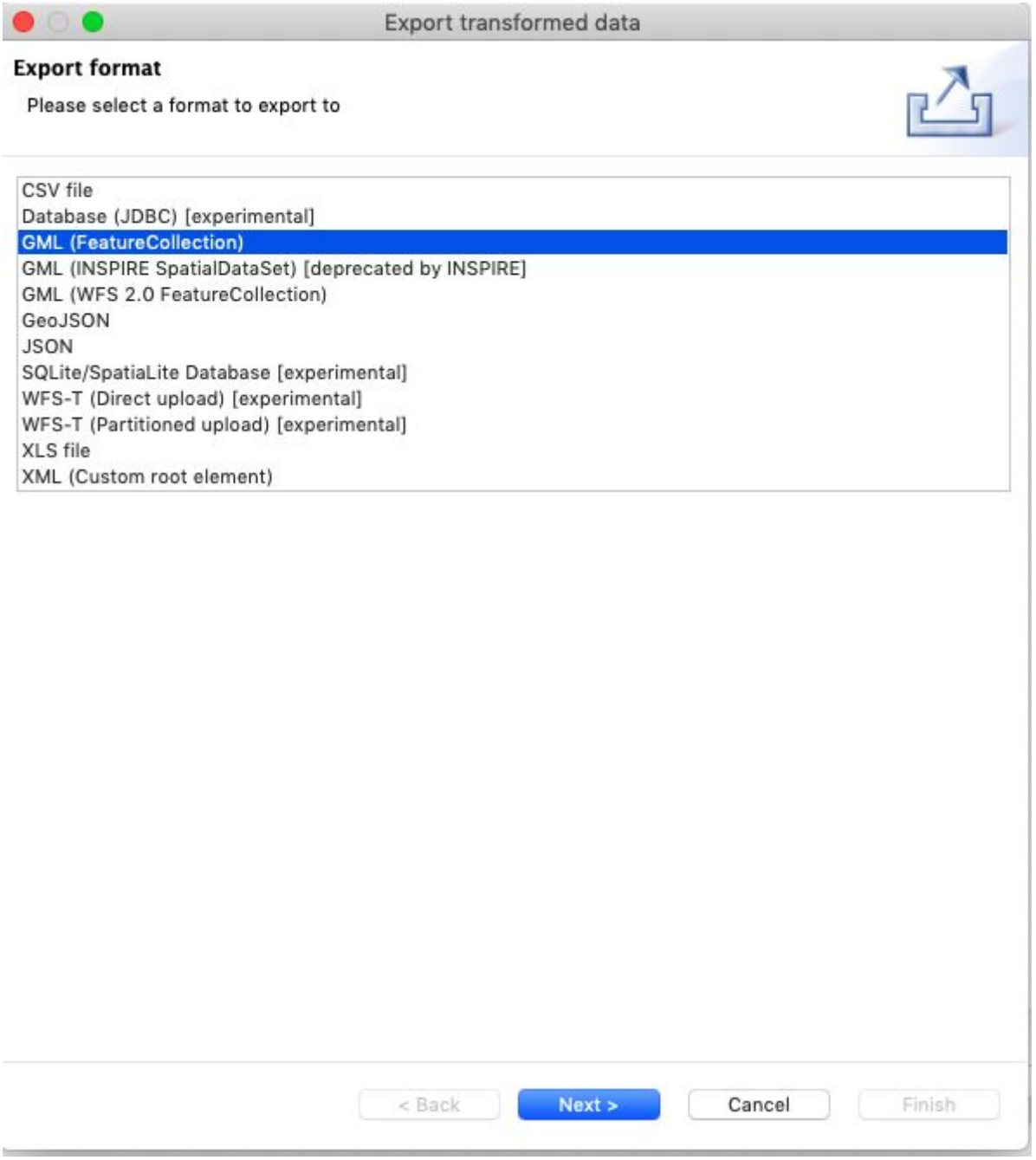

Następnie należy dokonać wizualnej inspekcji wyniku w widoku mapy  i zapisać projekt: **File -> Save alignment project as…** jako plik "**postgis.halez**".

## Ćwiczenie 3: Generowanie plików XML i GML

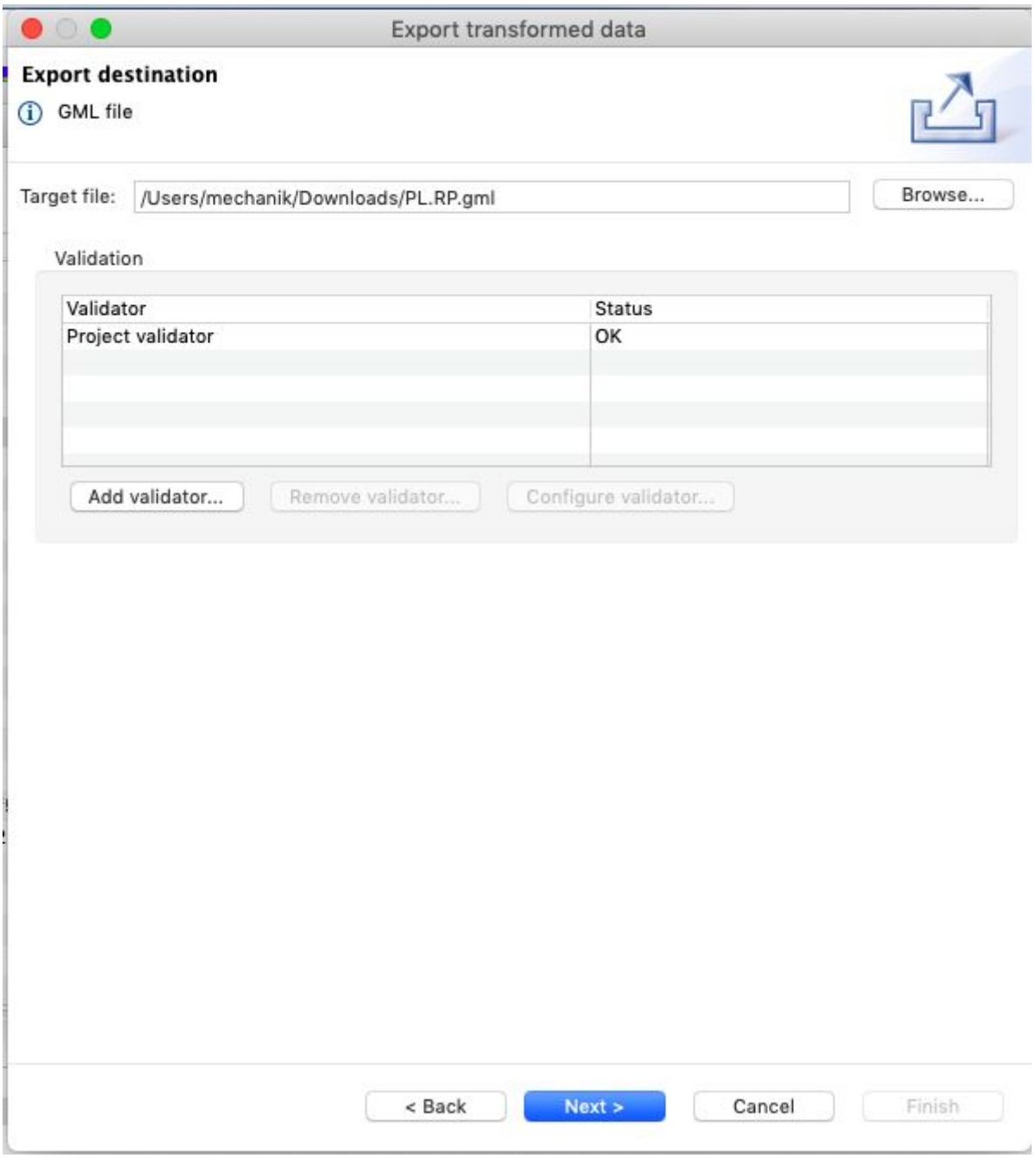
Wykorzystując projekt "**postgis.halez**" z Ćwiczenia 2, należy wygenerować eksport do pliku GML.

Narzędzie eksportu aktywuje się poprzez **File -> Export -> Transformed data**.

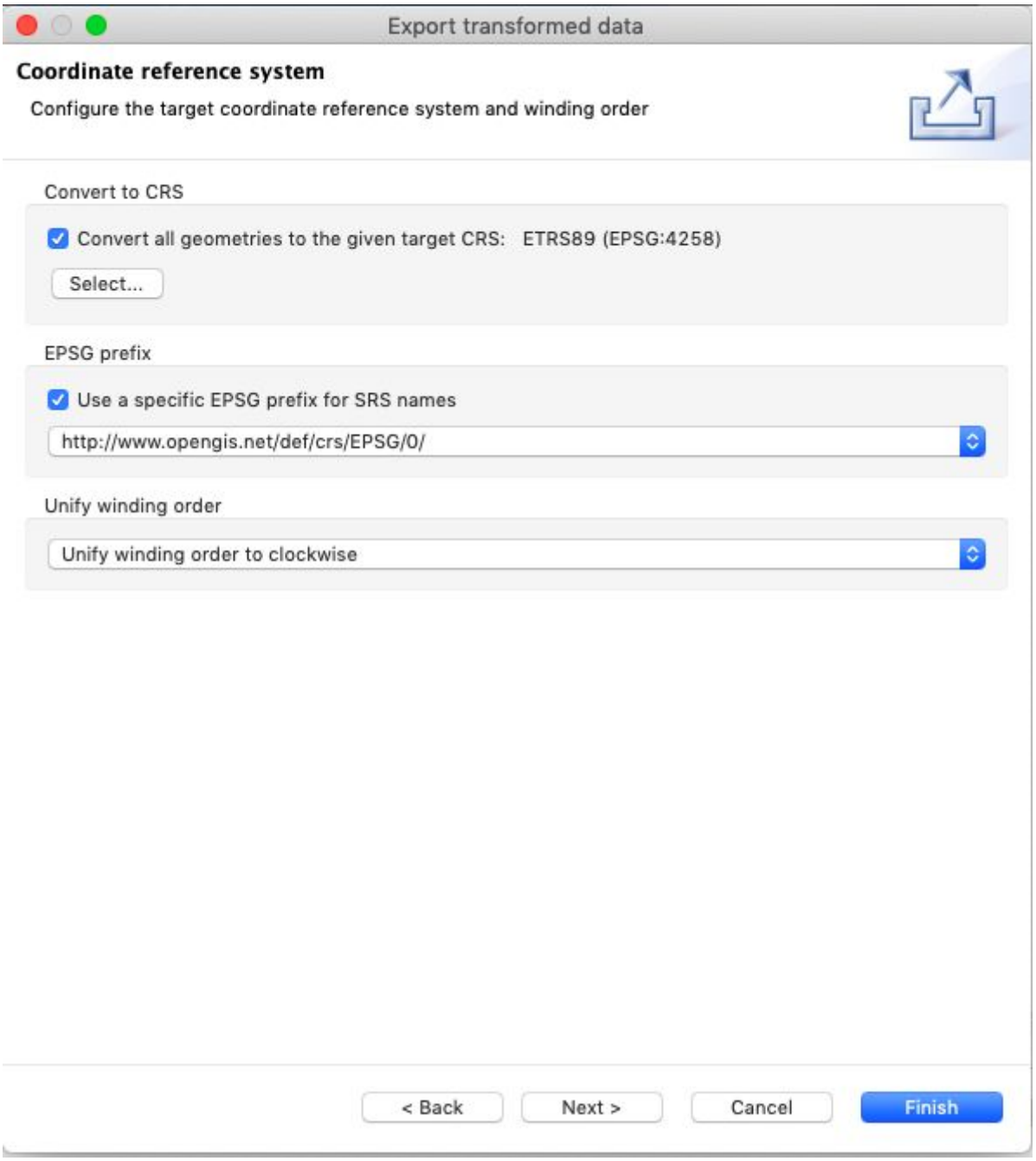
Należy wybrać GML (FeatureCollection).




Jako nazwę pliku wynikowego należy wpisać "**PL.RP.gml**"

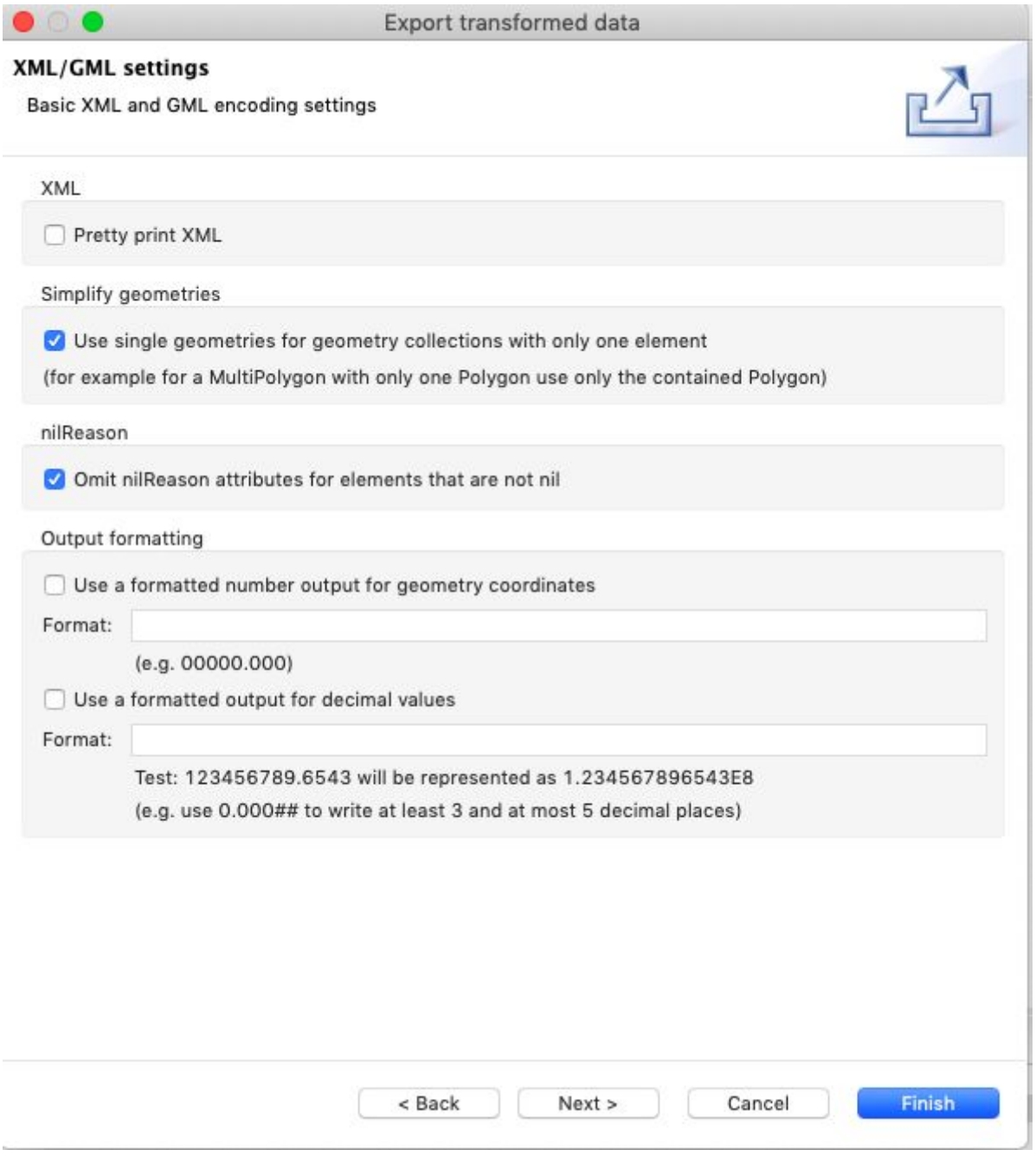



Następnie należy ustawić parametry eksportu: układ współrzędnych - **EPSG:4258**, prefiks układu współrzędnych - **http://opengis.net/def/crs/EPSG/0**, kolejność wierzchołków - Unify winding order to clockwise. Układ współrzędnych 4258 jest układem stosowanym do prezentacji danych przestrzennych w ujęciu Unii Europejskiej. Adres prefiksowy kieruje do bazy danych definiującej układy współrzędnych w formacie maszynowym, co pozwala na automatyczne zaimportowanie właściwych zmiennych układu odniesienia. Wymuszenie kolejności wierzchołków zapobiega błędom wynikającym z różnic w kolejności zapisywania wierzchołków w plikach SHP (zgodnie z ruchem wskazówek zegara) oraz GML (przeciwnie do ruchu wskazówek zegara).

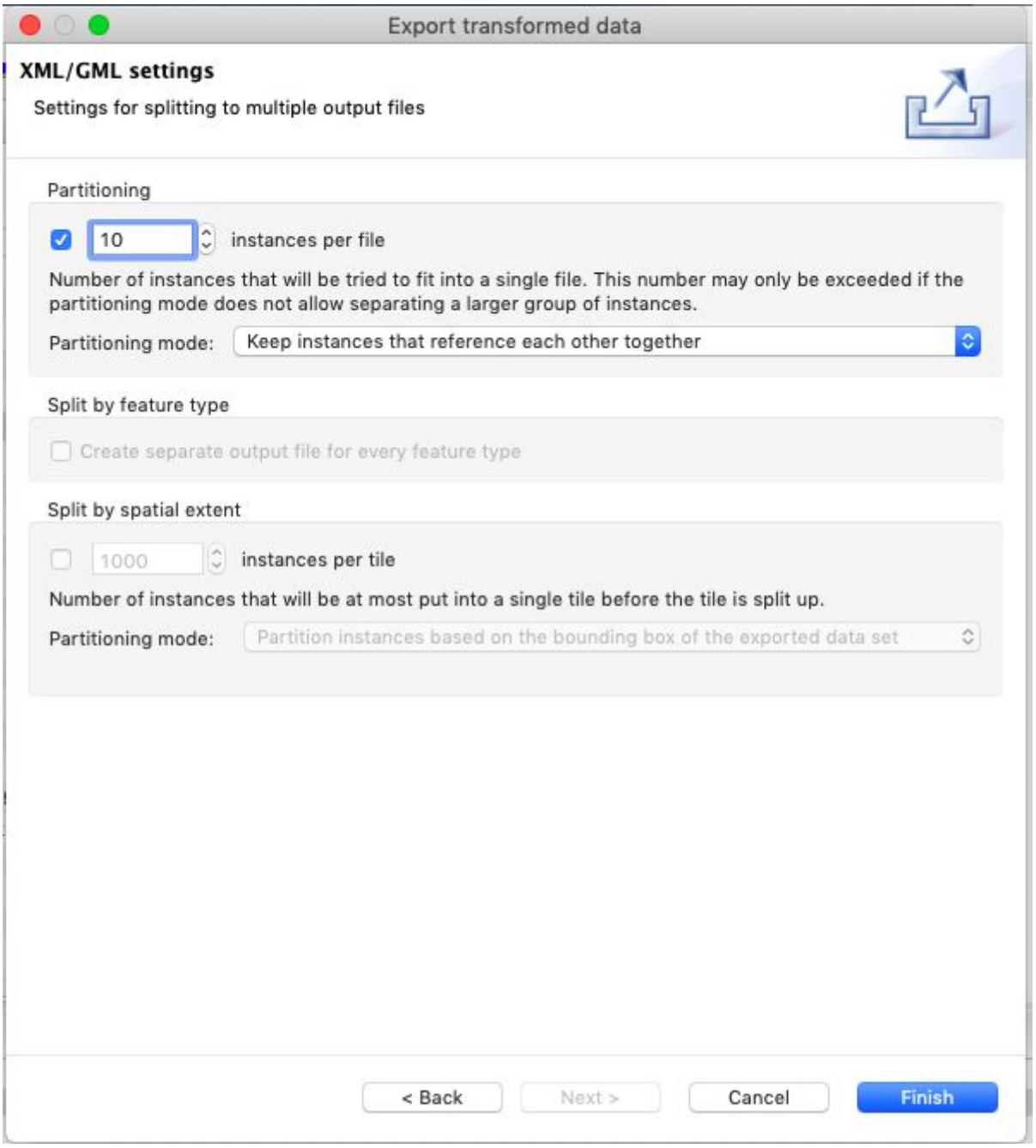



W kolejnym etapie należy zaznaczyć opcję:

**Use single geometries for geometry collections with only one element**
**Omit nilReason attributes for elements that are not nil.** Wybranie tej opcji pozwala na właściwe zakodowanie opisu atrybutów wskazujących na powód braku wartości w analizowanym zbiorze (np.: inapplicable, missing, template, unknown, withheld, other). Oznacza to, że jeśli w zbiorze nie pojawią się wymagane wartości, plik wynikowy będzie posiadał ich opis, w przeciwnym przypadku (dane wpisane prawidłowo) opis nie będzie potrzebny.




W ostatnim etapie należy wybrać partycjonowanie po 10 obiektów na plik.




Na koniec należy dokonać wizualnej inspekcji wygenerowanych plików GML (podobnie jak we wcześniejszych ćwiczeniach. Dodatkowo można dokonać walidacji otwierając utworzony plik wynikowy w oprogramowaniu GIS (np. QGIS).

## Ćwiczenie 3: Przygotowanie sparametryzowanych szablonów

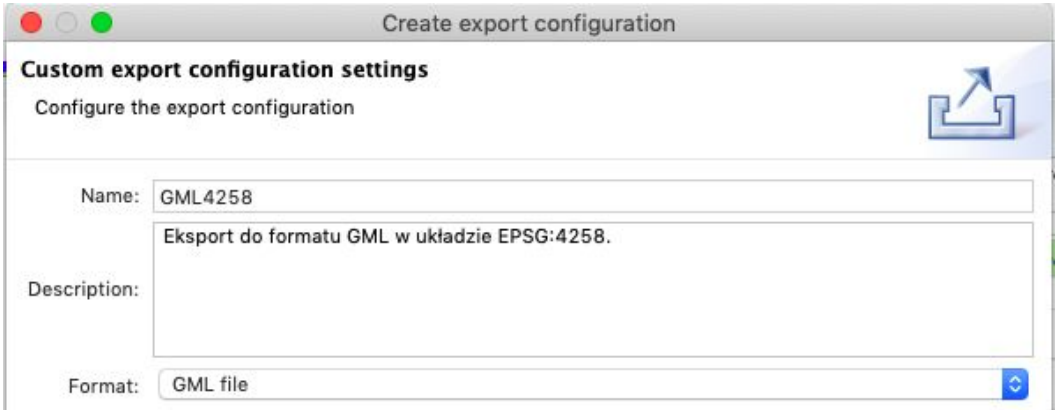
Celem ćwiczenia jest praktyczne tworzenie Własnych schematów eksportu, pozwalających automatyzować proces poprzez przypisanie mu serii parametrów, które użytkownik spodziewa się wykorzystywać w przyszłości.

Do ćwiczenia należy wykorzystać projekt utworzony w Ćwiczeniu 2.

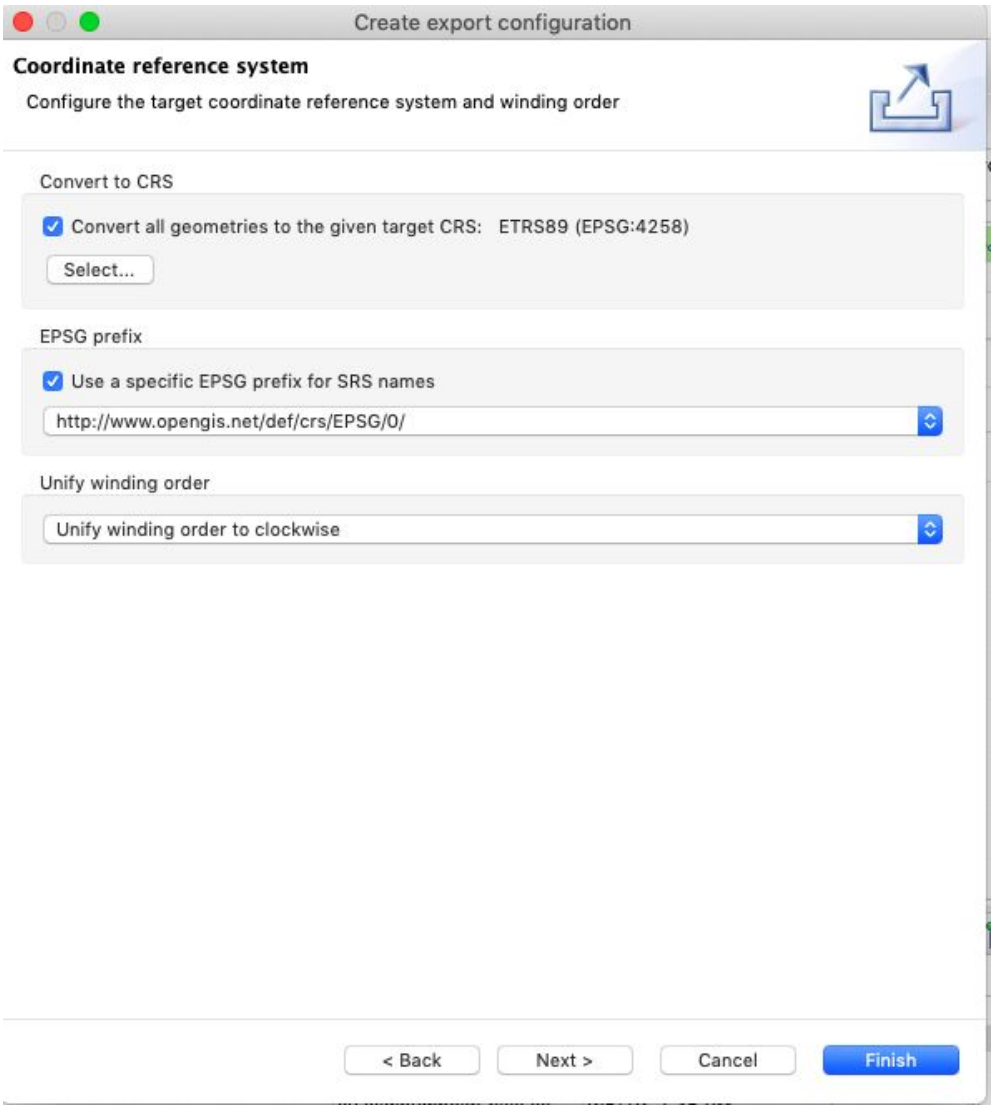
Z menu należy wybrać **File -> Export -> Create custom data export.**

Jako format należy wybrać **GML (FeatureCollection)**.

Jako nazwę należy wybrać "**GML4258**".




Pozostałe parametry eksportu należy ustawić identycznie jak w Ćwiczeniu 2, za wyjątkiem partycjonowania danych, które należy wyłączyć:




Następnie należy zapisać projekt jako "**cwiczenie3.halez**" poprzez **File -> Save alignment project as**.

## Cwiczenie 4: Wykorzystanie szablonów w Hale CLI

Celem ćwiczenia jest pokazanie wykorzystania szablonów Hale CLI służących sterowaniu przy przetwarzaniu wsadowym plików (z poziomu konsoli systemowej) w celu przyspieszenia i automatyzacji procesu transformacji danych.

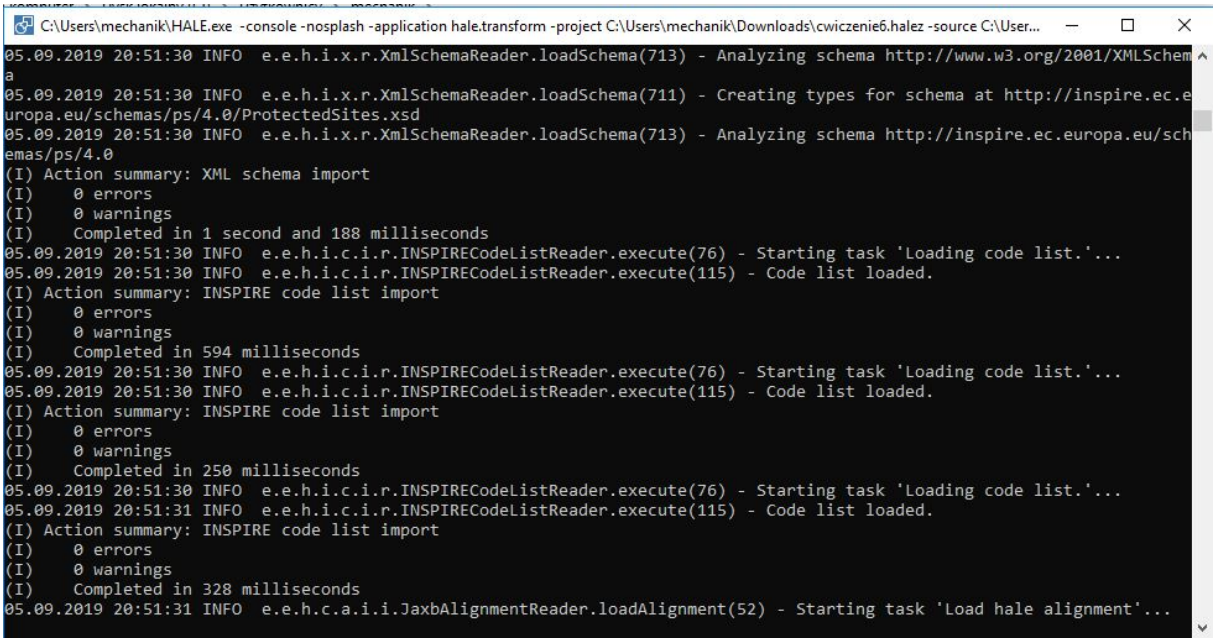
W ćwiczeniu należy wykorzystać projekt z szablonem eksportu, przygotowanym w Ćwiczeniu 3. Plik "**cwiczenie3.halez**" i zestaw plików Shapefile **"ObszarySpecjalnejOchronyPolygon"** należy umieścić w tym samym katalogu, co plik HALE.exe (w przypadku instalacji HALE zgodnie z Ćwiczeniem 1, plik ten znajdzie się w katalogu C:\Users\Nazwa użytkownika\wetransform).

Należy uruchomić wiersz polecenia systemu Windows, wpisując w menu Start komendę "cmd".

Następnie metodą "przeciągnij i upuść" należy przeciągnąć do wiersza polecenia plik wykonywalny HALE.exe, i uzupełnić komendę o:

-console -nosplash -application hale.transform -project cwiczenie3.halez -source ObszarySpecjalnejOchronyPolygon.shp -target oso.gml -preset GML4258

i zatwierdzić klawiszem Enter. Konsola HALE uruchomi się w nowym oknie, pierwsze uruchomienie może potrwać około minuty.




Należy poczekać do wykonania transformacji i zamknięcia okna konsoli HALE, po czym zweryfikować zawartość utworzonego pliku **oso.gml** w QGIS.

## Ćwiczenie 5: Masowe przetwarzanie danych

Do wykonania ćwiczenia należy posłużyć się tym samym projektem "**cwiczenie3.halez**" co w Ćwiczeniu 4.

Plik projektu należy przenieść do katalogu **masowa_konwersja.**

W katalogu tym należy utworzyć plik "harmonizacja.bat" według szablonu (całość powinna zostać zapisana w jednej linii):

```
for %%i in (*.shp) do
```

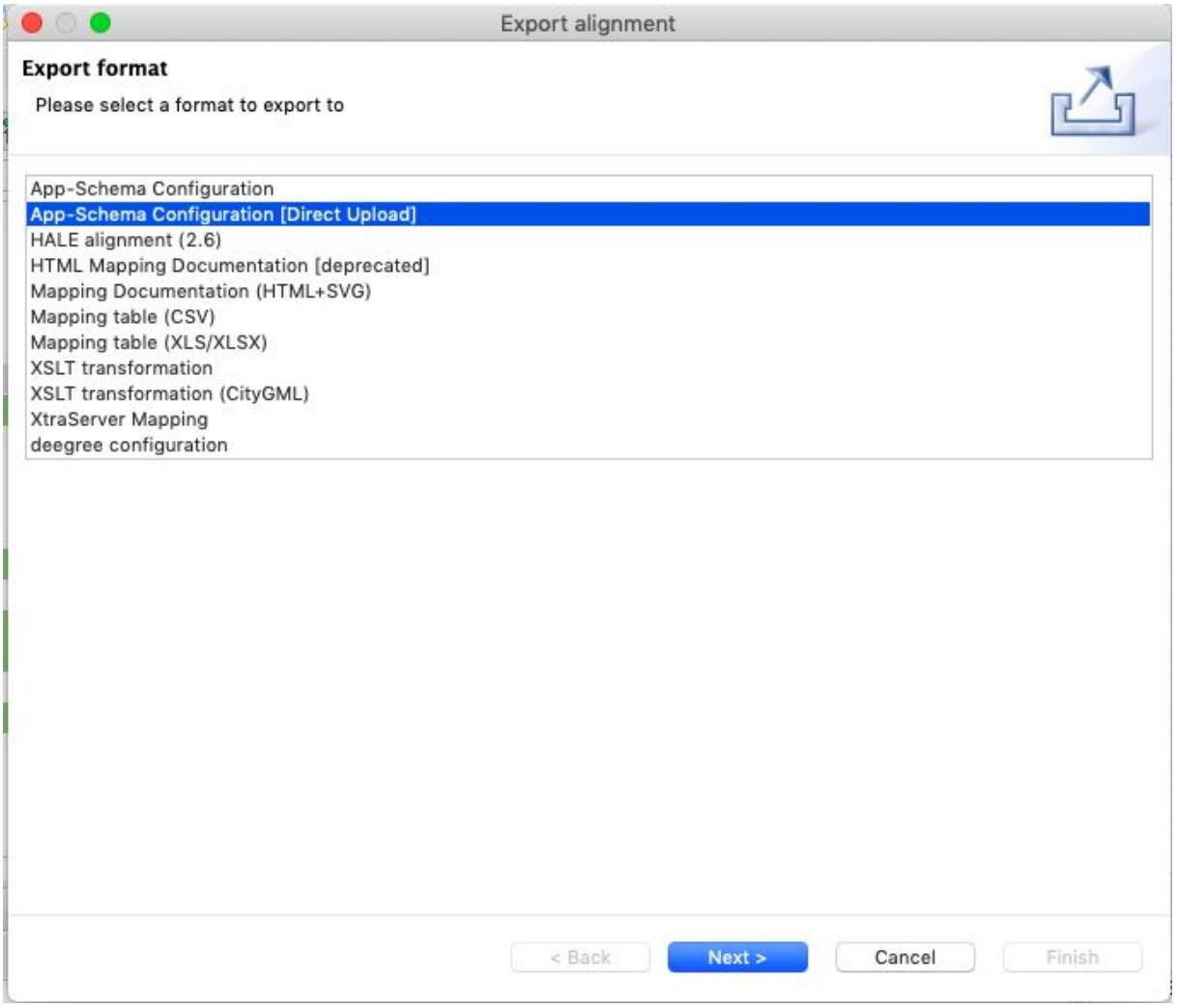
```
C:\Users\inspire\wetransform\HALE.exe -nosplash
-console -application hale.transform -project
cwiczenie3.halez -source %%i -target %%i.gml
-preset GML4258
```

zamieniając fragment zaznaczony **tłustym drukiem** na właściwą dla własnego systemu ścieżkę do HALE.exe. Plik należy zapisać (z wykorzystaniem formatu "Wszystkie pliki", inaczej otrzyma on rozszerzenie .txt i nie będzie wykonywalny) i uruchomić podwójnym kliknięciem.

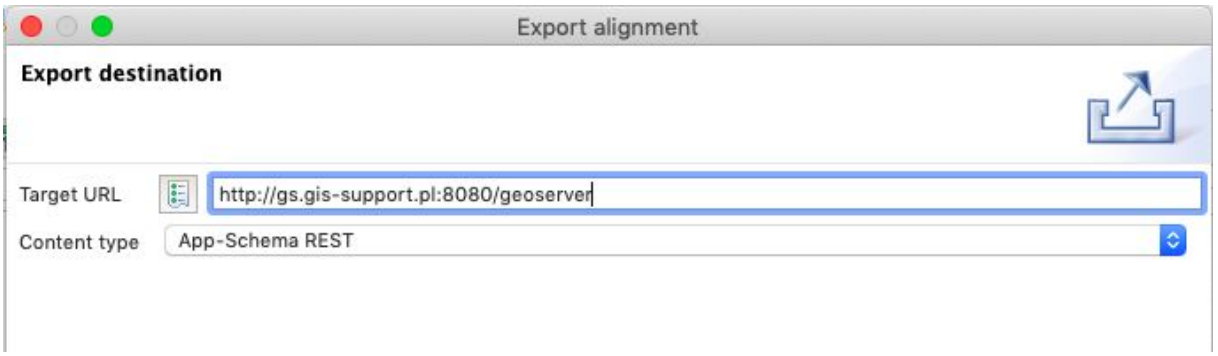
## Ćwiczenie 6: Publikacja danych

Do wykonania ćwiczenia należy posłużyć się projektem wykorzystującym bazę PostGIS z Ćwiczenia 3 - **postgis.halez**.

Aby dokonać eksportu danych do GeoServer z rozszerzeniem AppSchema, należy zastosować narzędzie **Export -> Alignment -> AppSchema [direct upload].**

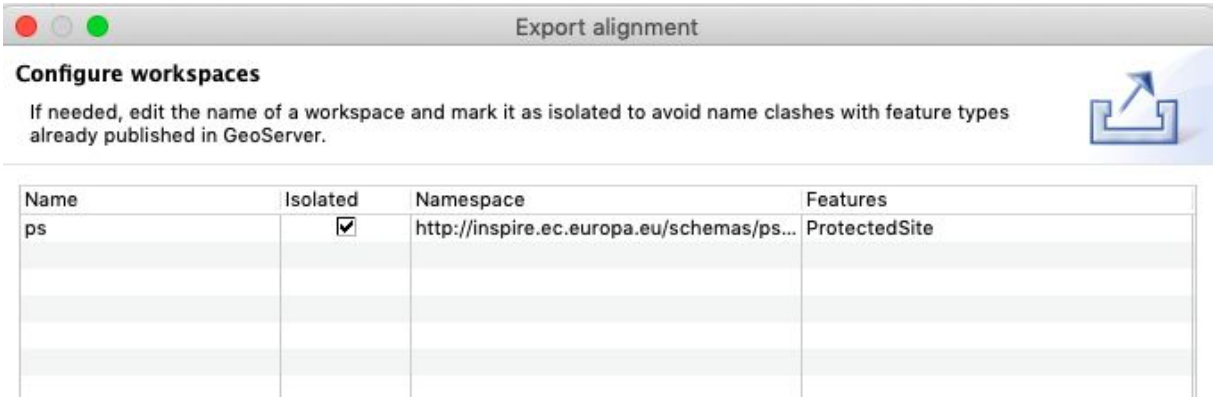



W następnym kroku należy podać bazowy URL do GeoServer (bez podawania ścieżki do REST API rozszerzenia AppSchema): http://gs.gis-support.pl:8080/geoserver

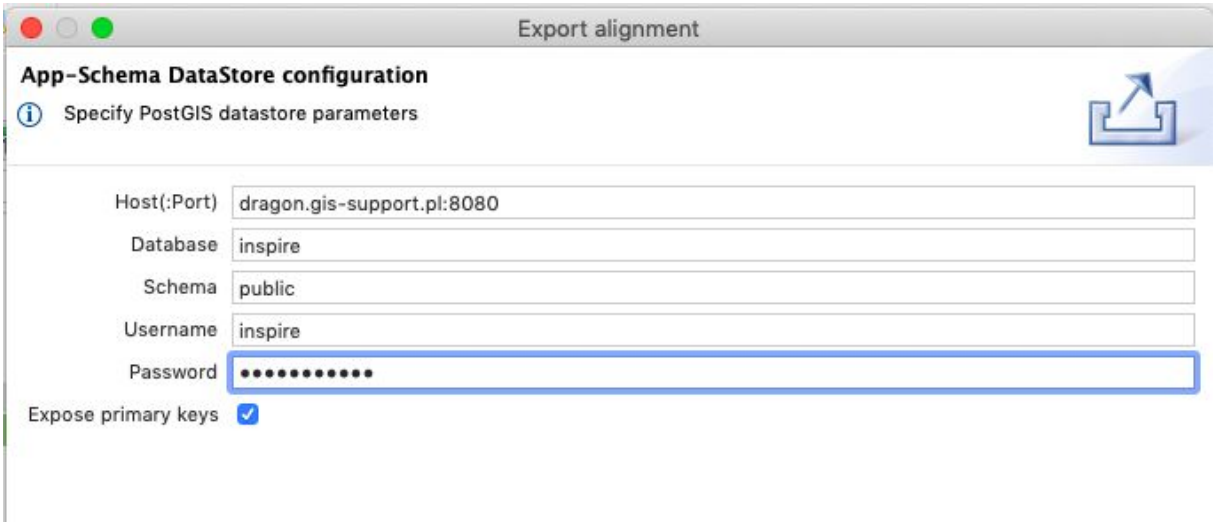



Pole "**Include target schema in archive**" powinno zostać zaznaczone.


W kolejnym kroku należy zaznaczyć opcję "isolated".




a następnie podać parametry połączenia z bazą danych PostGIS, w której znajdują się dane źródłowe.




Ostatnim etapem jest podanie danych logowania do GeoServera i zatwierdzenie wysyłki.