INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Harmonização de CDG's com Hale

DMSI/GEO

(05-07-2016)

Harmonização

- **O que é a Harmonização?**
- Para ***Hale***
  - "*The alignment is the mapping between source and target schemas. It defines relations between source and target entities (types or properties). Based on the defined relations a transformation is derived.*"
- Segundo ***Dean M. Hintz***
  - "*Core to the harmonization workflow is the transformation process which reshapes source schema and geometry to match the required destination structure*." (Hintz,2012,1)

Boas práticas da Harmonização

- **Antes de começar a trabalhar com Hale ou outra ferramenta de harmonização**
- Ler (e reler) as especificações dos dados!
  - ***Data Specification - Technical Guideline***
  - Modelos UML
- Preparar com antecedência a Matching Table
- Procurar exemplos de CDG's já harmonizados nos ***Thematic Clusters***.
- Partilhar conhecimentos e experiências com a comunidade!

Hale

- **H**UMBOLDT **AL**ignment **E**ditor

- Software utilizado para criar mapping entre schemas diferentes e aplicar a transformação resultante

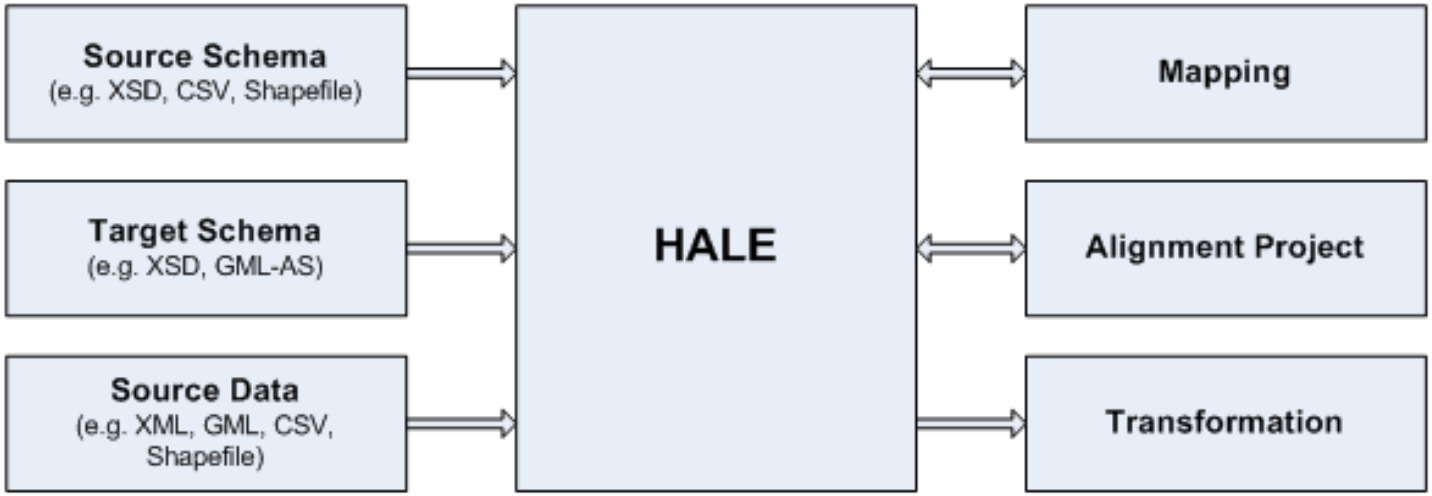
- *Software Open Source*

- Fornece uma interface gráfica rica, textual e especificamente aprovado para especialistas em IG

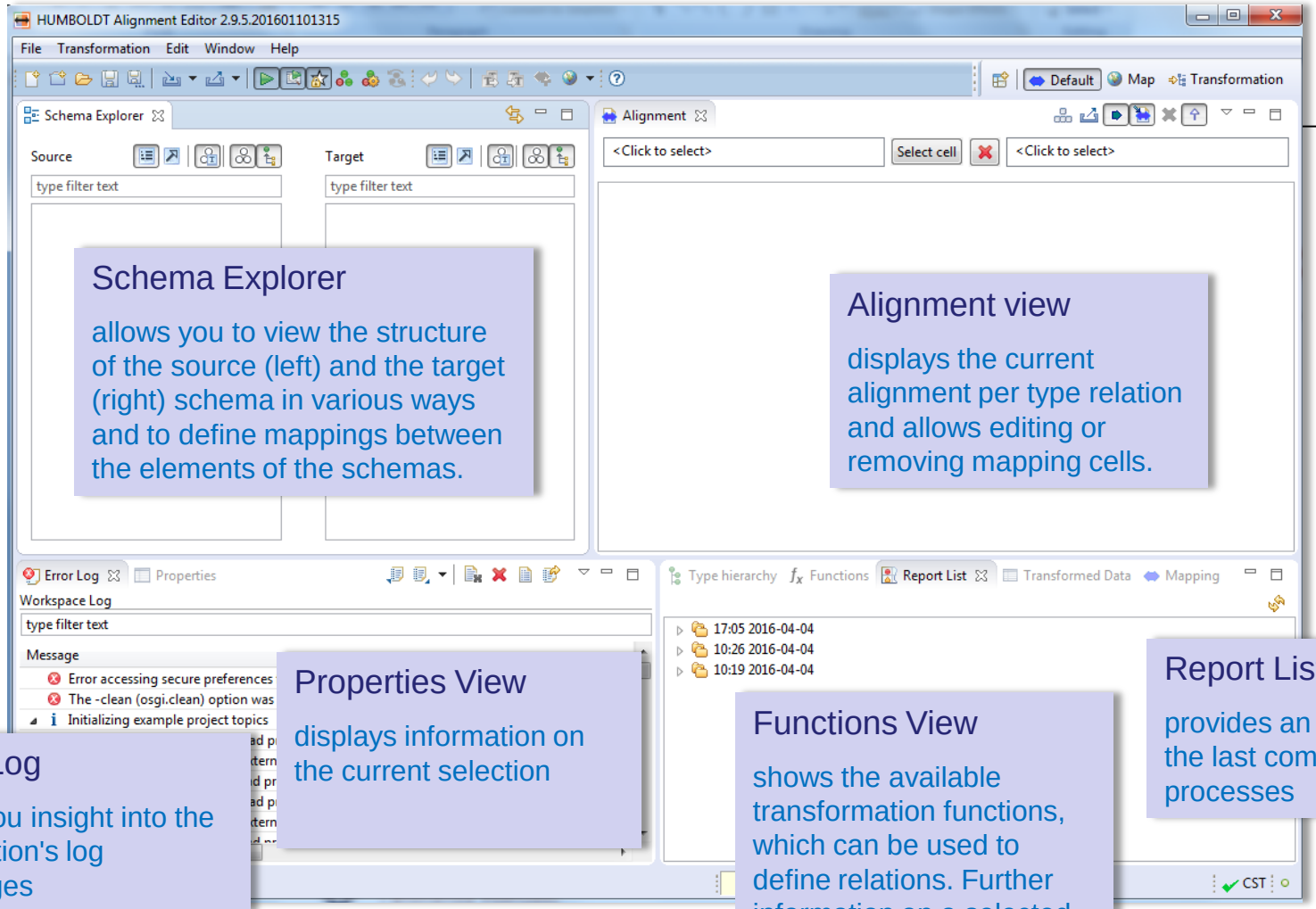
- *Feedback* instantâneo sobre o processo de harmonização de dados

- Download versão 2.9.4 (2015-11-01) versão 32 e 64 bit para windows, Mac OS, Linux

Hale


Source Schema
(e.g. XSD, CSV, Shapefile)
Target Schema
(e.g. XSD, GML-AS)
Source Data
(e.g. XML, GML, CSV, Shapefile)
HALE
Mapping
Alignment Project
Transformation

Hale Interface



Schema Explorer

allows you to view the structure of the source (left) and the target (right) schema in various ways and to define mappings between the elements of the schemas.

Alignment view

displays the current alignment per type relation and allows editing or removing mapping cells.

Error Log

gives you insight into the application's log messages

Properties View

displays information on the current selection

Functions View

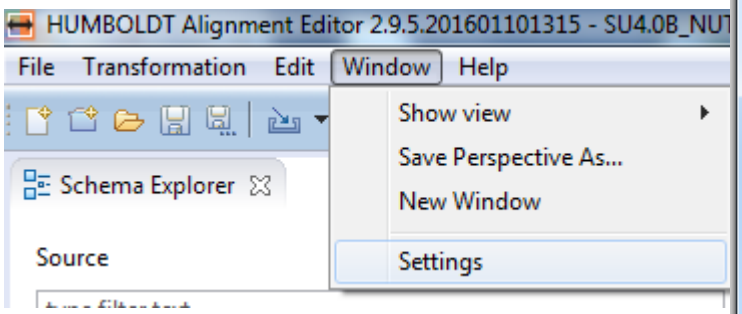
shows the available transformation functions, which can be used to define relations. Further information on a selected function will be displayed in the *Properties* view.

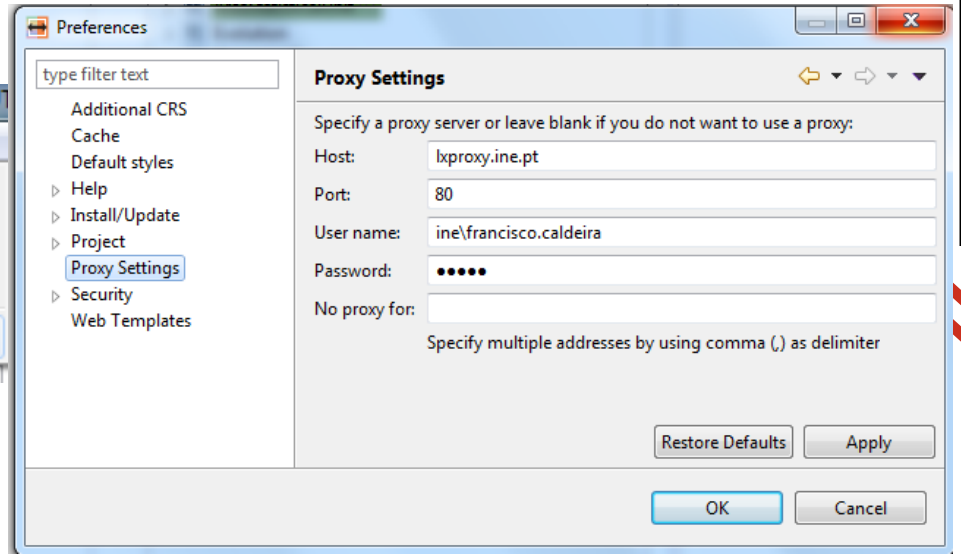
Report List

provides an overview of the last completed processes

Proxy Server

- A versão 2.9.4 de ***Hale*** parece não funcionar bem com proxy
- Fundamental para acesso a recursos na Web (***Codelists***, …)
- Como definir *proxy*?



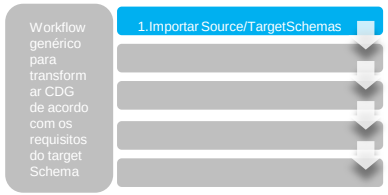


Hale - Workflow

Workflow genérico para transformar CDG de acordo com os requisitos do target Schema

1. Importar Source/TargetSchemas
2. Importar dados
3. Definir mapping rules
4. Exportar dados transformados
5. Validar dados transformados

Importar Source Schema

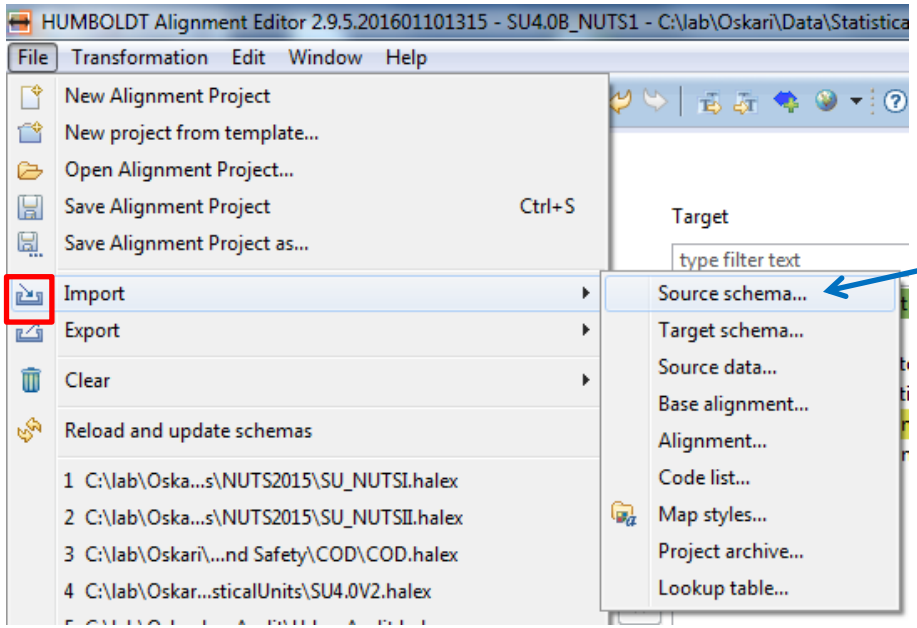


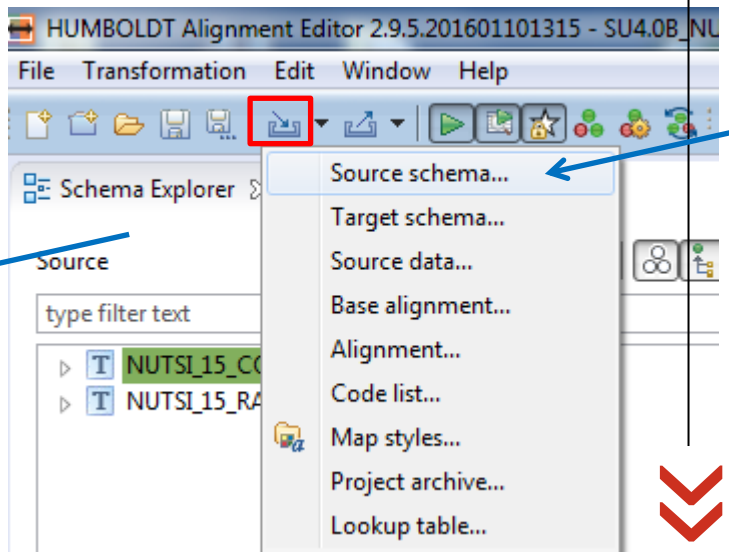
- ***Source Schema***
  - Define a estrutura dos dados que desejamos transformar
  - Várias fontes para dados, incluindo *online* (URL, WFS) , base de dados (PostgreSQL/PostGIS, SpatiaLite)

Shapefile (*.shp)
XML schema (*.xsd, *.xml)
CSV file (*.csv)
MS OOXML Format Spreadsheet (XLSX) (*.xlsx)
SpatiaLite Database (*.sqlite)
HALE Schema Definition (*.hsd, *.haleschema)
GZiped HALE Schema Definition (*.hsd.gz, *.haleschema.gz)
Excel Spreadsheet (XLS) (*.xls)

Importar Source Schema

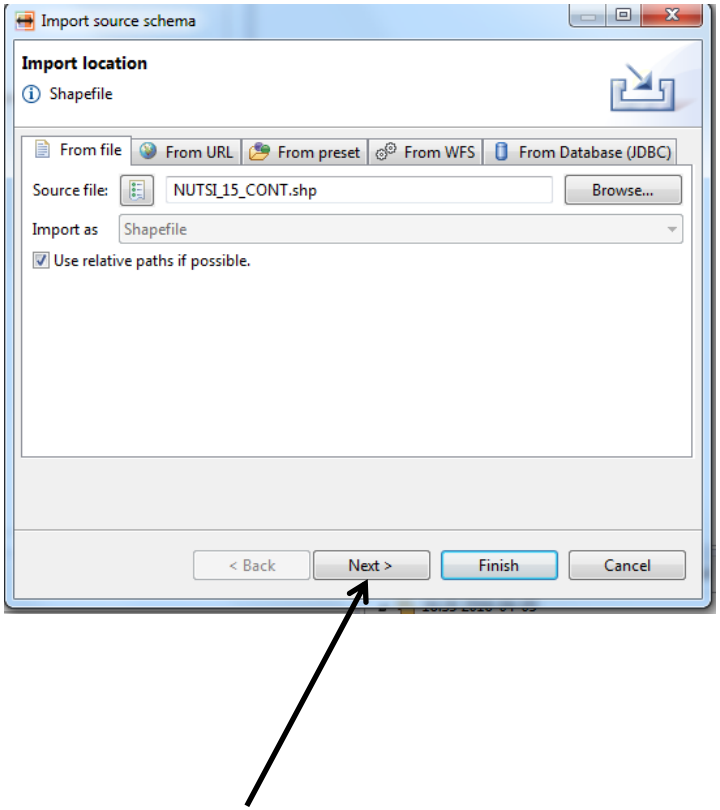
- ***Source Schema***
  - Várias formas de invocar a funcionalidade



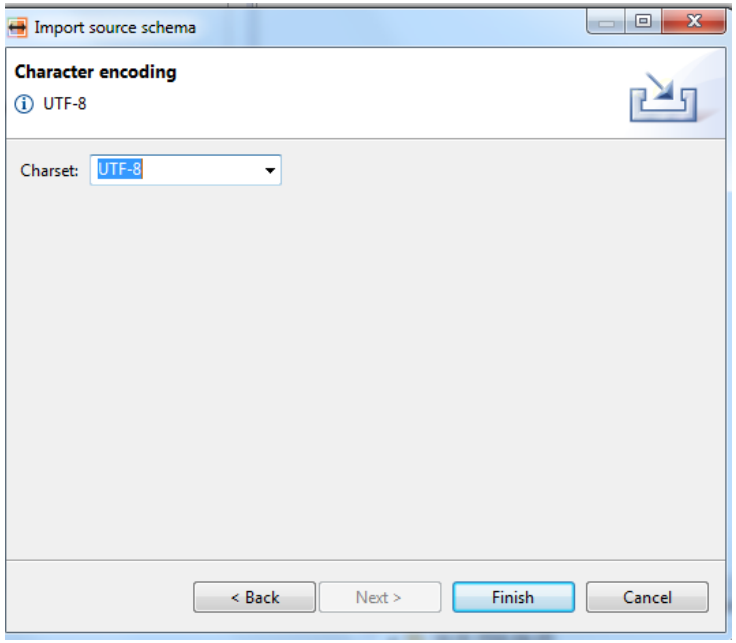


Importar Source Schema

- **importar *Source Schema***



- **Cuidado com os caracteres *PT***



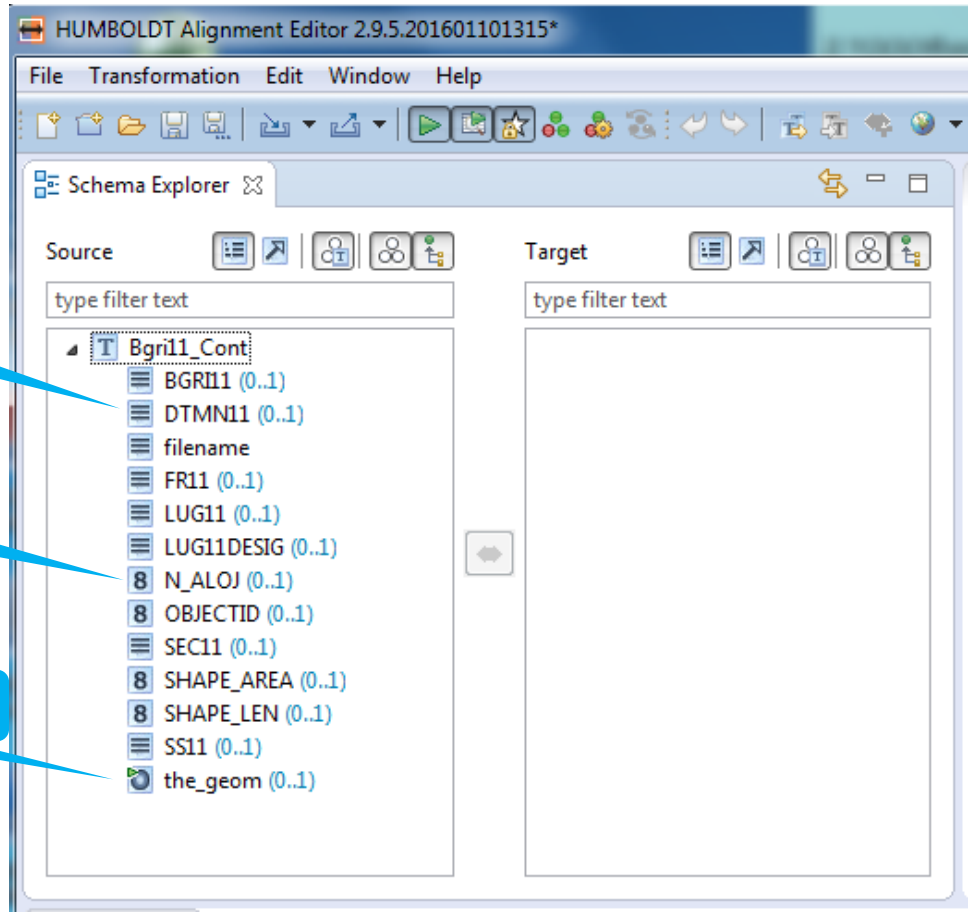
Importar Source Schema

- **importar *Source Schema* - Simbologia**

String

Numérico

Geometria

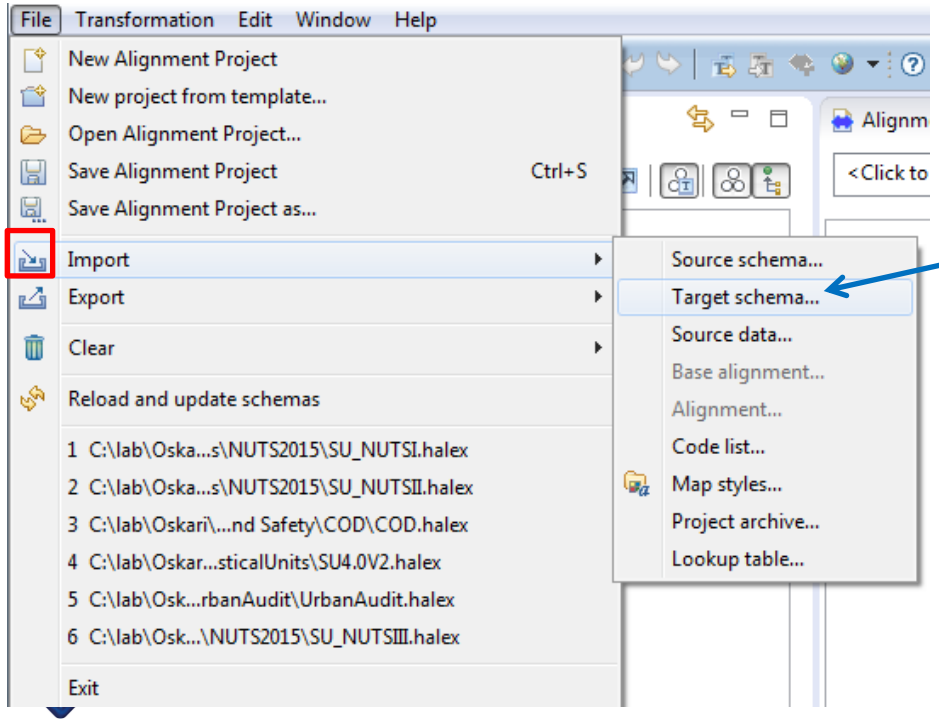


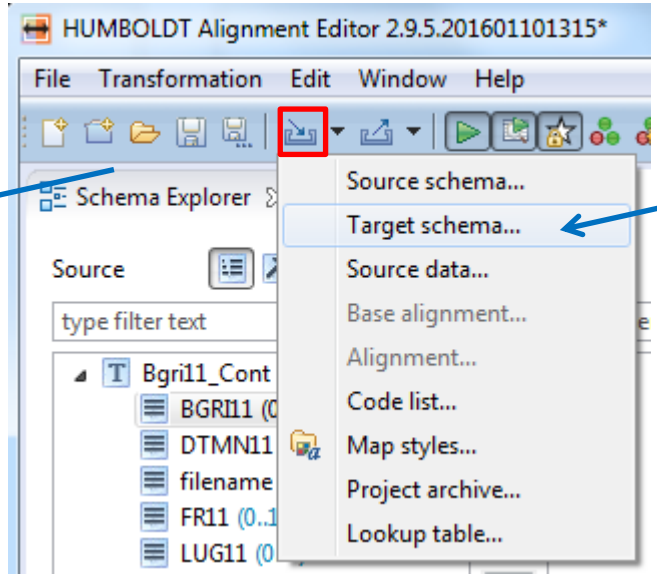
Properties

- String property
- Numeric property
- Geometry property
- Other (complex) property

Importar target Schema

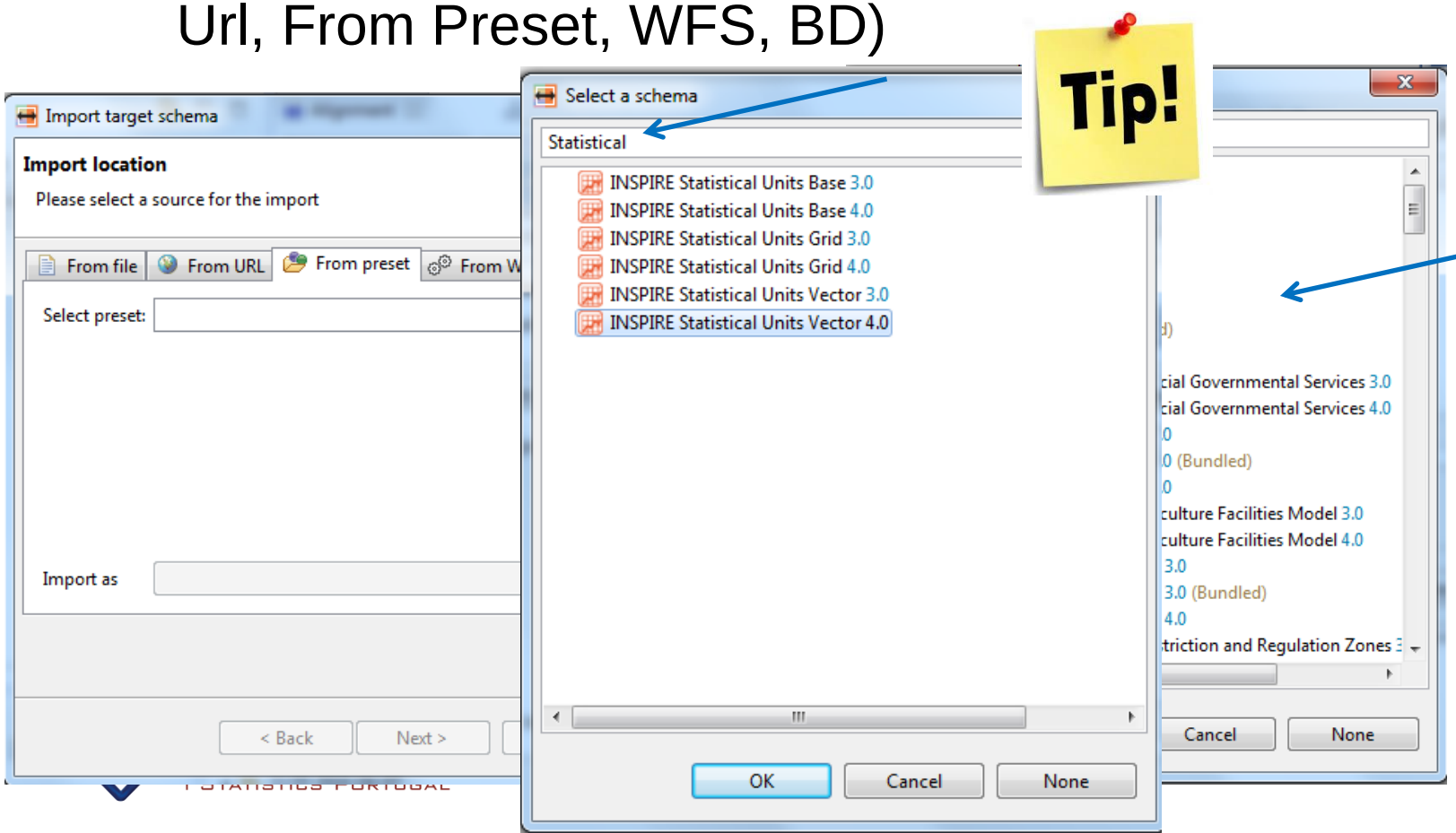
- ***Target Schema***
  - Define a estrutura para a qual desejamos transformar os dados
  - Schemas definidos nas ***Data Specifications Inspire***





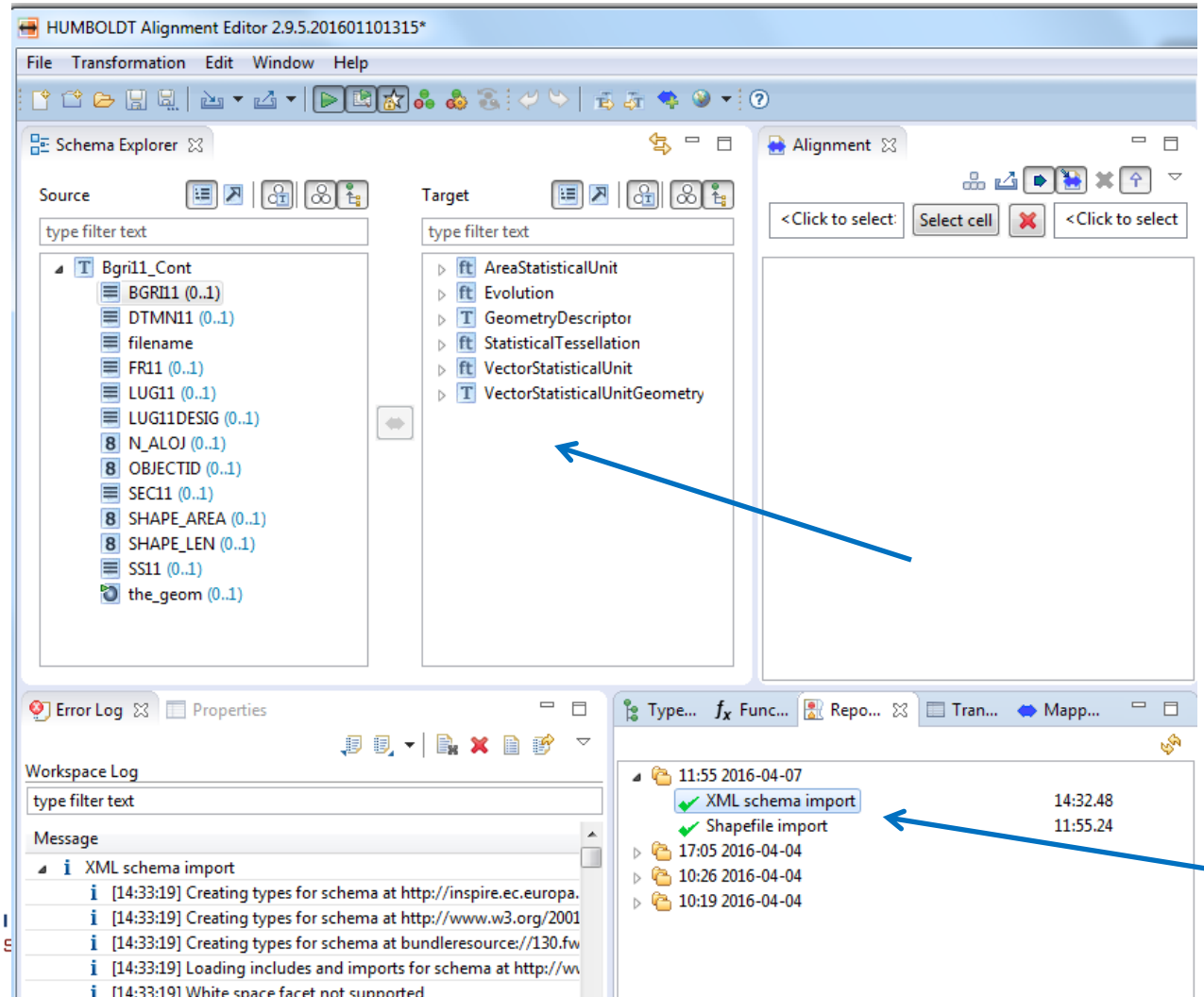
Importar Target Schema

- **Target Schema**
  - Várias fontes,semelhante, ao ***Source Schema*** (File, Url, From Preset, WFS, BD)



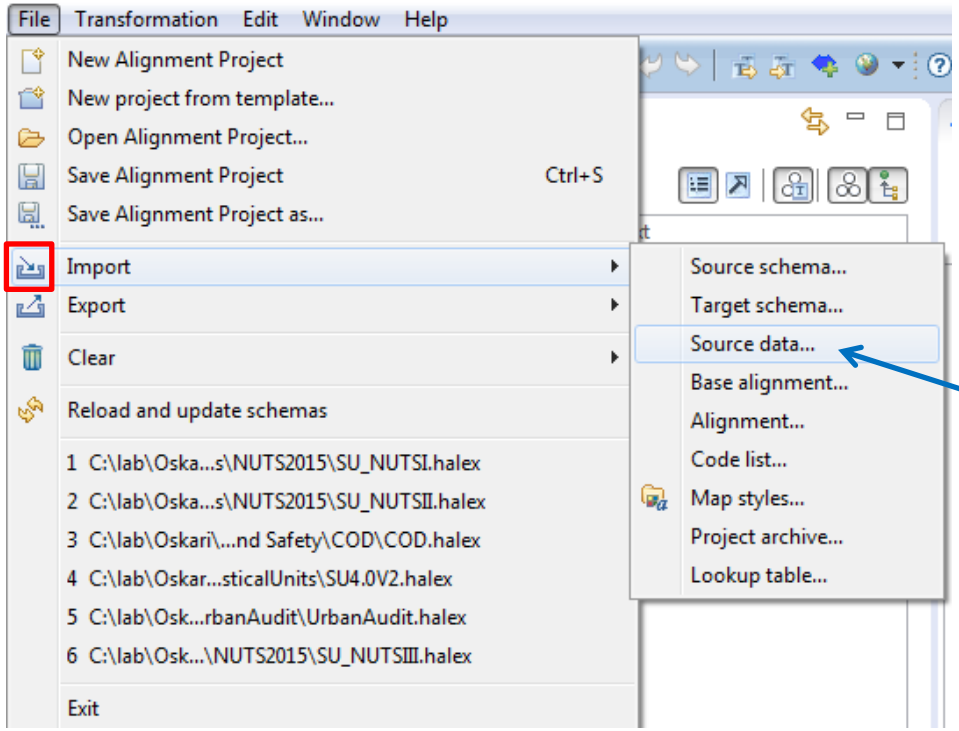
Importar target Schema

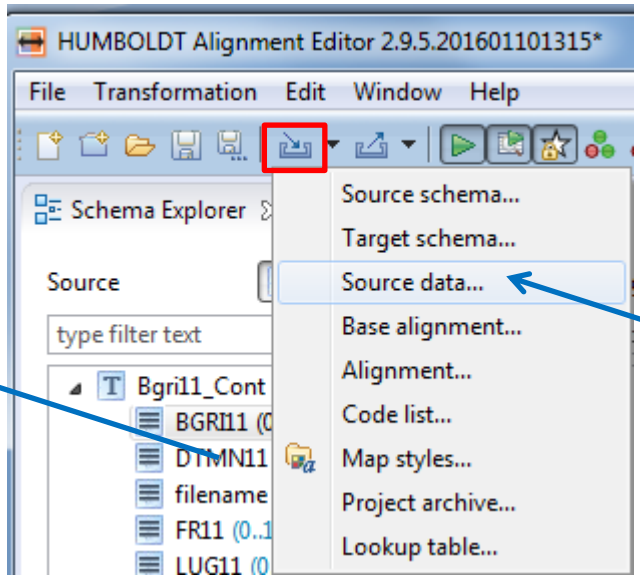
- ***Target Schema***



Importar Source data

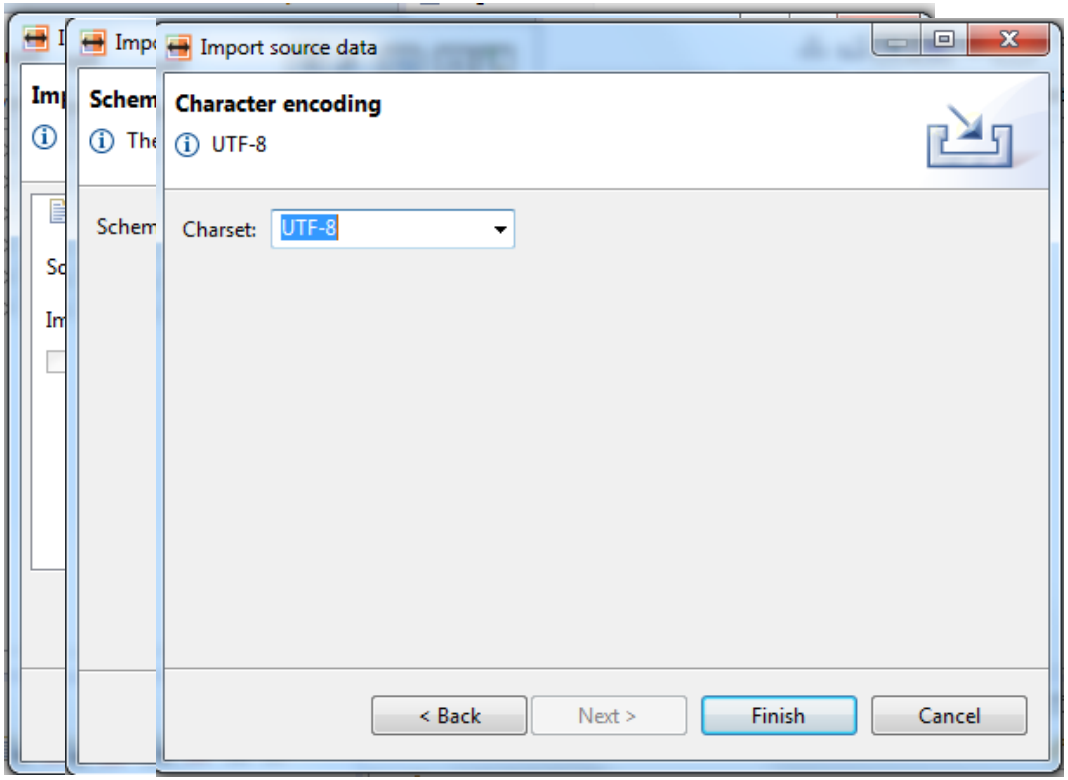
- **Importar *Source Data***
  - Importar o ***source Schema*** não importa as *features* (polígonos, linhas ou pontos).





Importar Source data

- **Importar *Source Data***

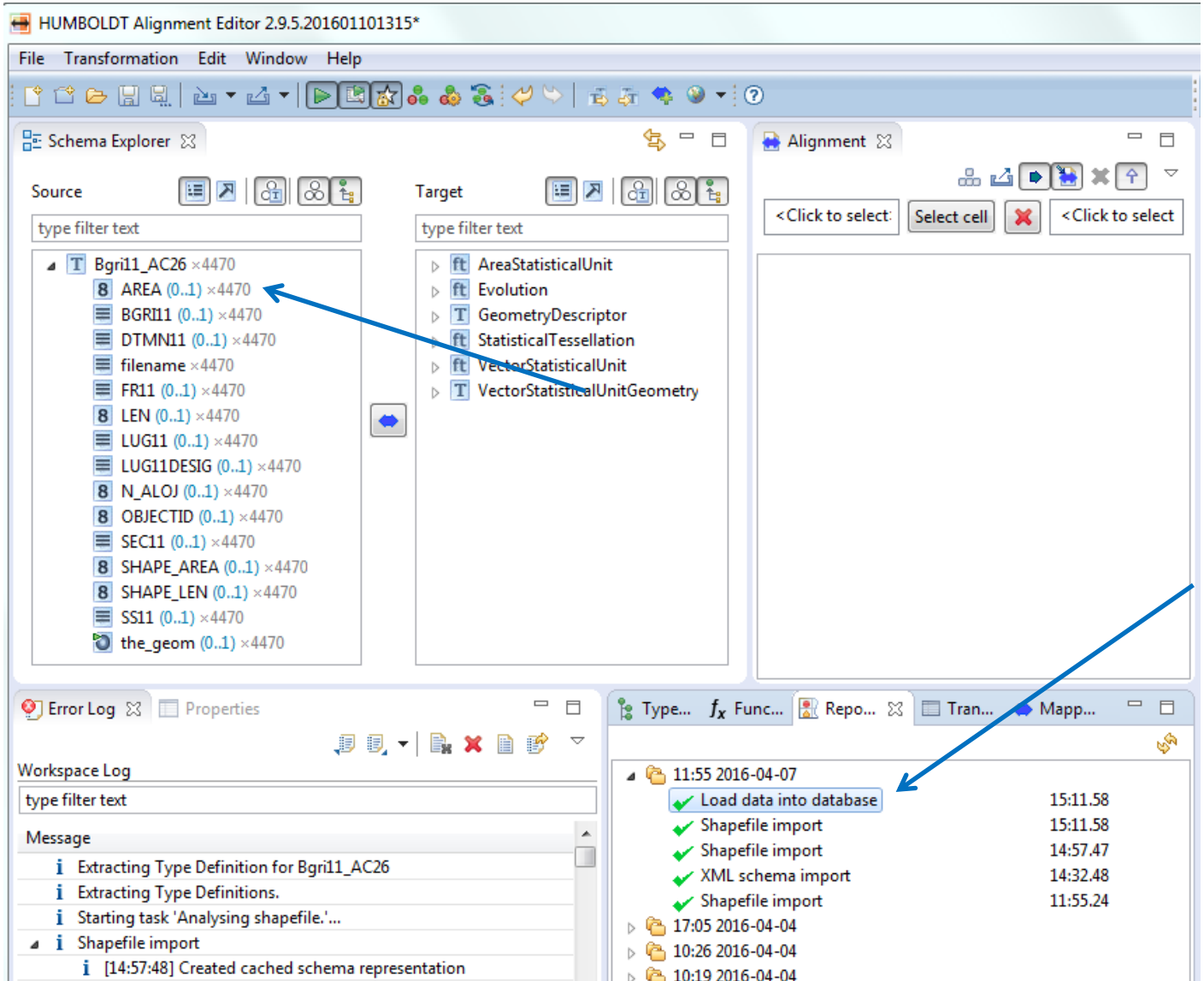


- Cuidado com os caracteres **PT**

Tip!

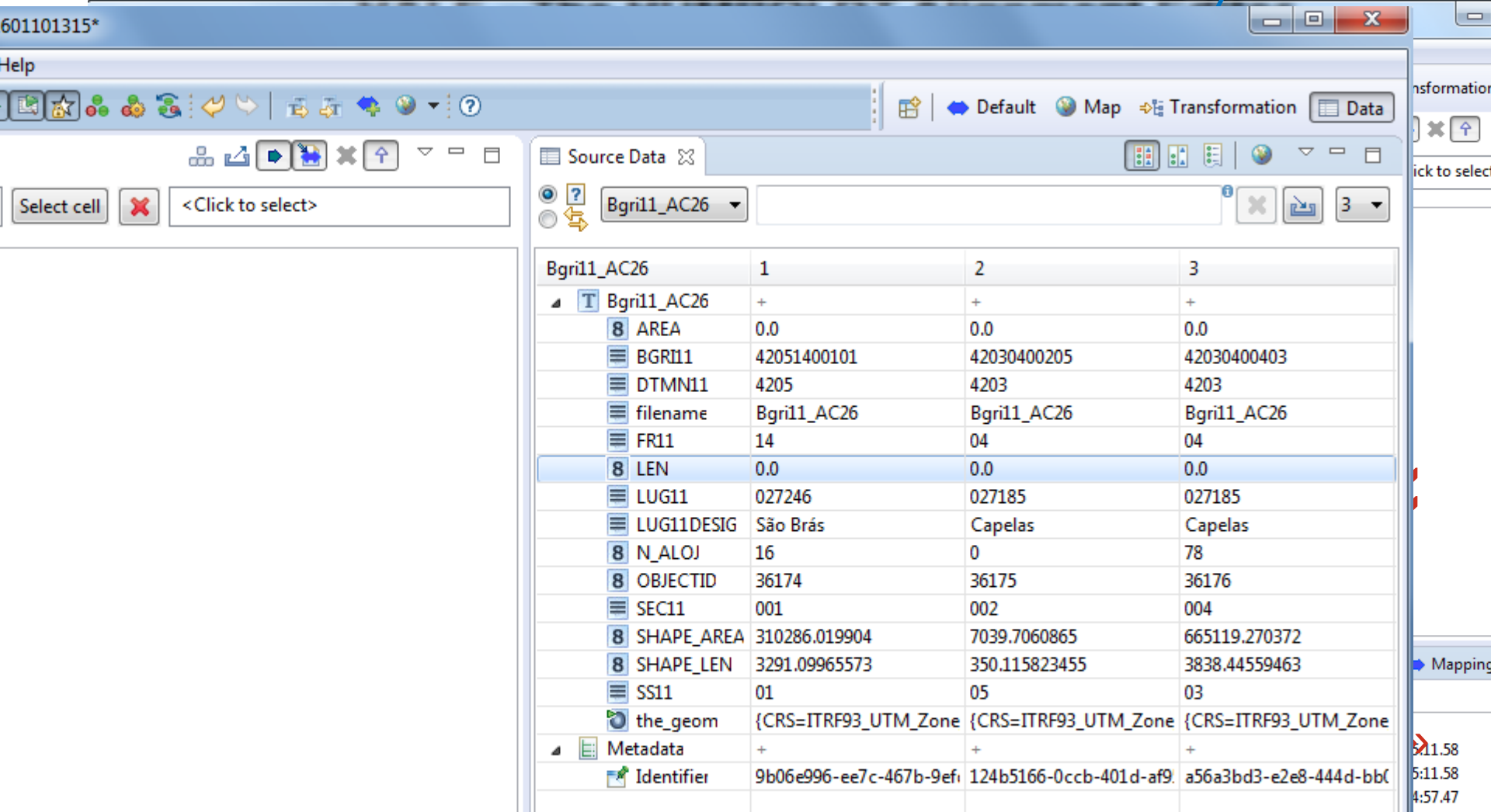
Importar Source data

Importar *Source Data*



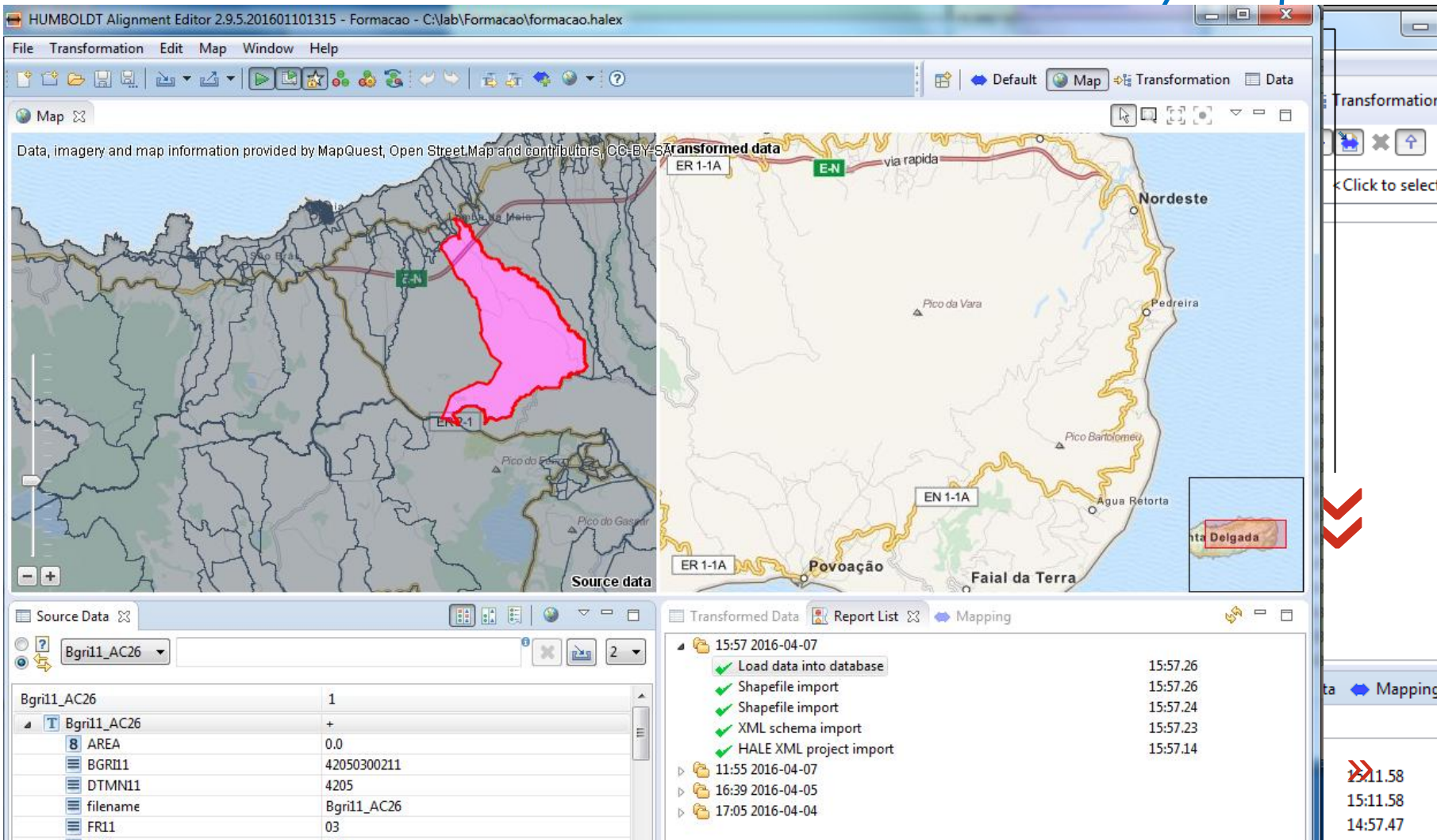
Importar Source data

- **Data perspestive**



Importar Source data

- ***Map perspestive***

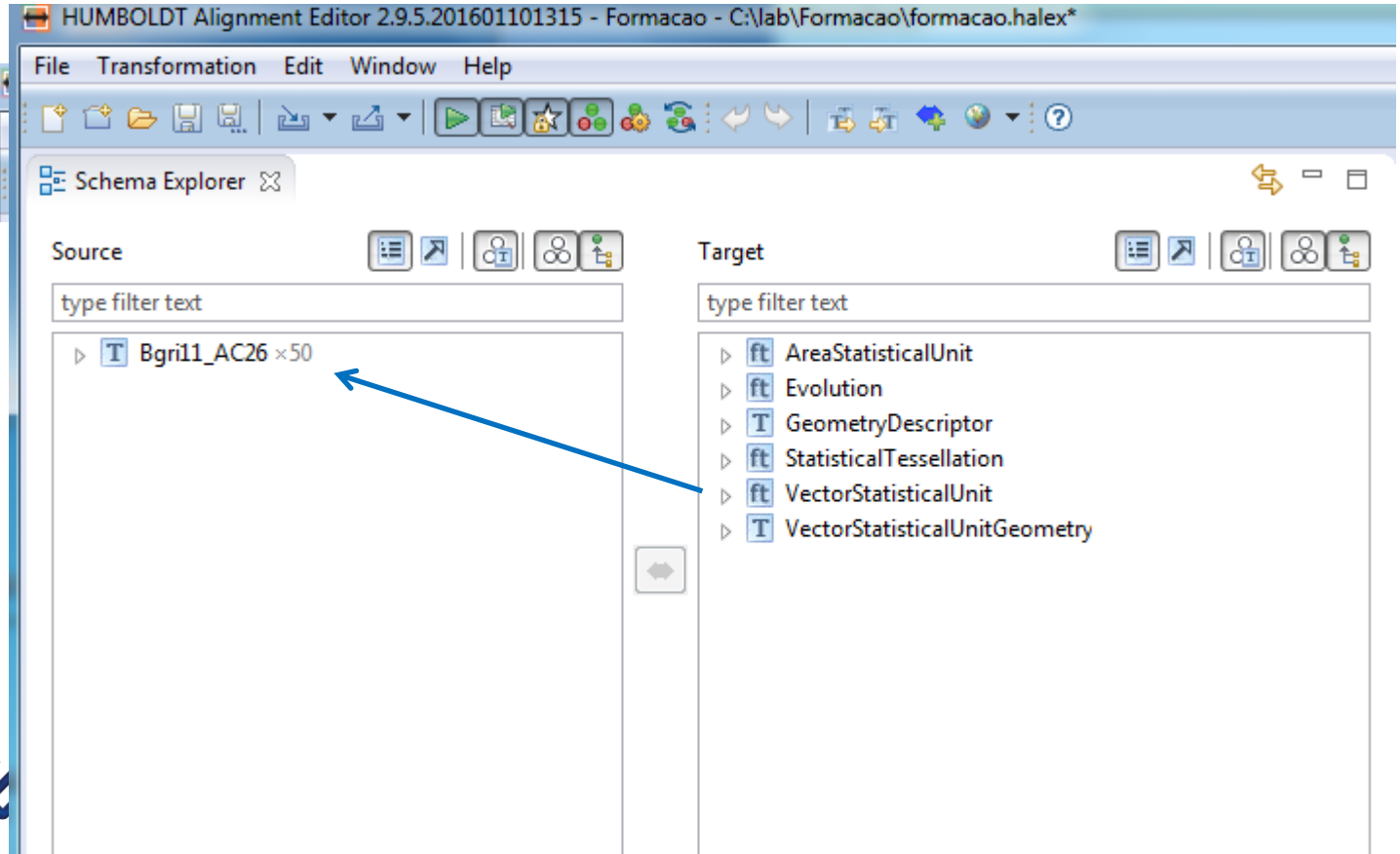


Importar Source data

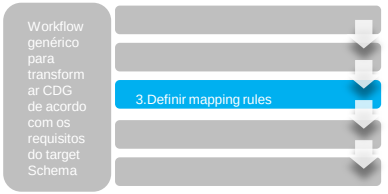
- **Importar *Source Data***
  - Importar cdg's com um número elevado de *features* pode trazer problemas de performance porque o ***HALE*** transforma os dados e valida, por defeito, sempre que existe mapeamento de algum campo.

Importar Source data

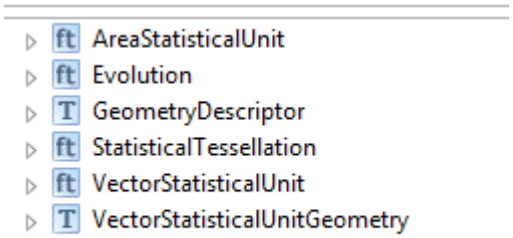
- É possível definir uma seleção de *features* para trabalhar com menos registos.



Definir Mapping rules



- No ***Target Schema*** é necessário identificar tipo relevante para o cdg que pretendemos harmonizar.

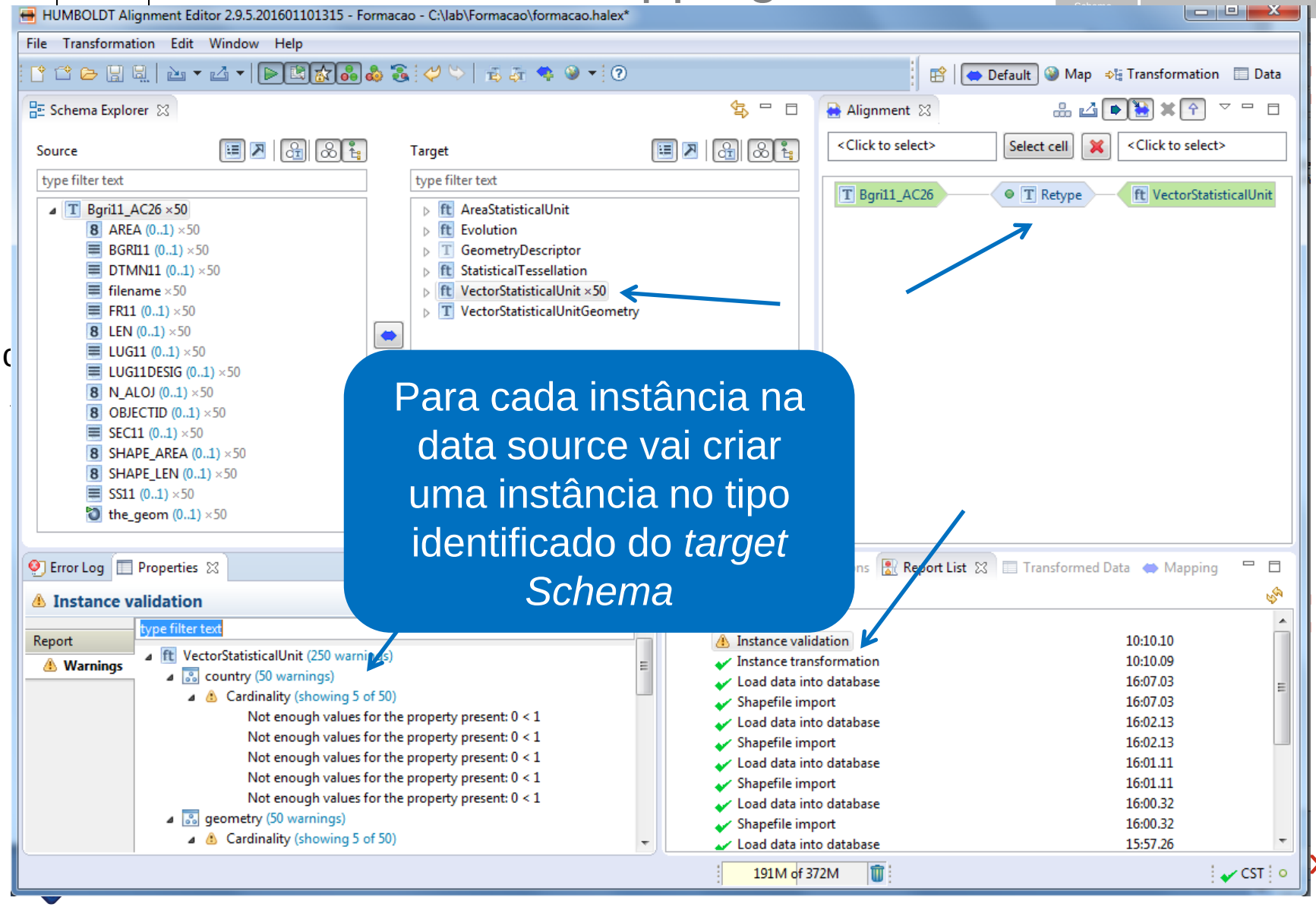


- As data specifications muitas vezes indicam claramente o tipo

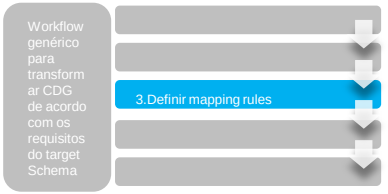
- The application schema on statistical units is composed of different packages:

  - Base: The base package.
  - Grid: Classes to represent statistical grids.
  - Vector: Classes to represent statistical unit having vector geometries (point, line, surfaces).

Definir Mapping rules



Definir Mapping rules



- ***Schema elements***

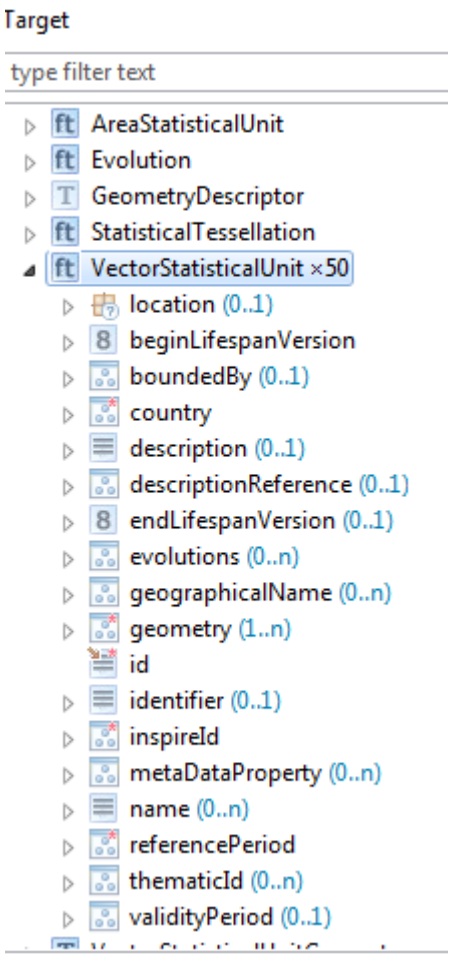
Properties

- String property
- Numeric property
- Geometry property
- Other (complex) property

Groups

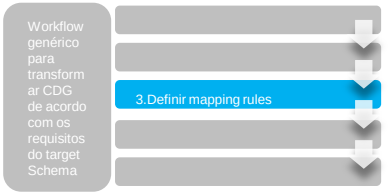
- Normal group containing a set of properties.
- Choice group, where only one of the specified properties is allowed as a child.

- A red asterisk marks properties that are mandatory, i.e. they occur exactly once and must have a value (and the value may not be *null*).
  Please note that if the parent of such a property is a choice ( ) the choice takes precedence, i.e. only one of its children may be present in an object, but the child that is present may still not hold a null value if marked with a red asterisk.
- A small brown error in the top left corner marks a property from a XML schema as being defined as a XML attribute.
- A property that is deemed to hold the main geometry of a type is marked with a small green triangle. Per type, you can set one property as the default geometry property. This property is then used when retrieving geometries for display in the map.

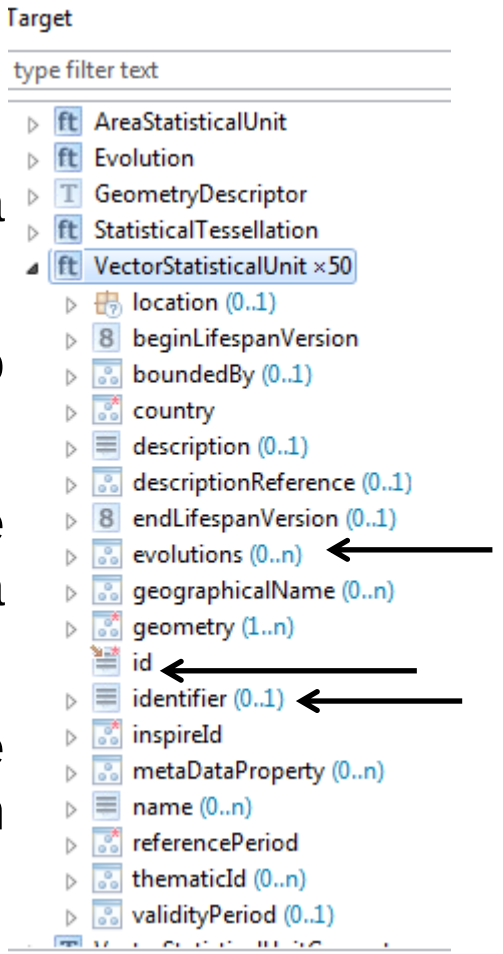


- Propriedades obrigatórias são facilmente identificadas no error log quando não estão preenchidas

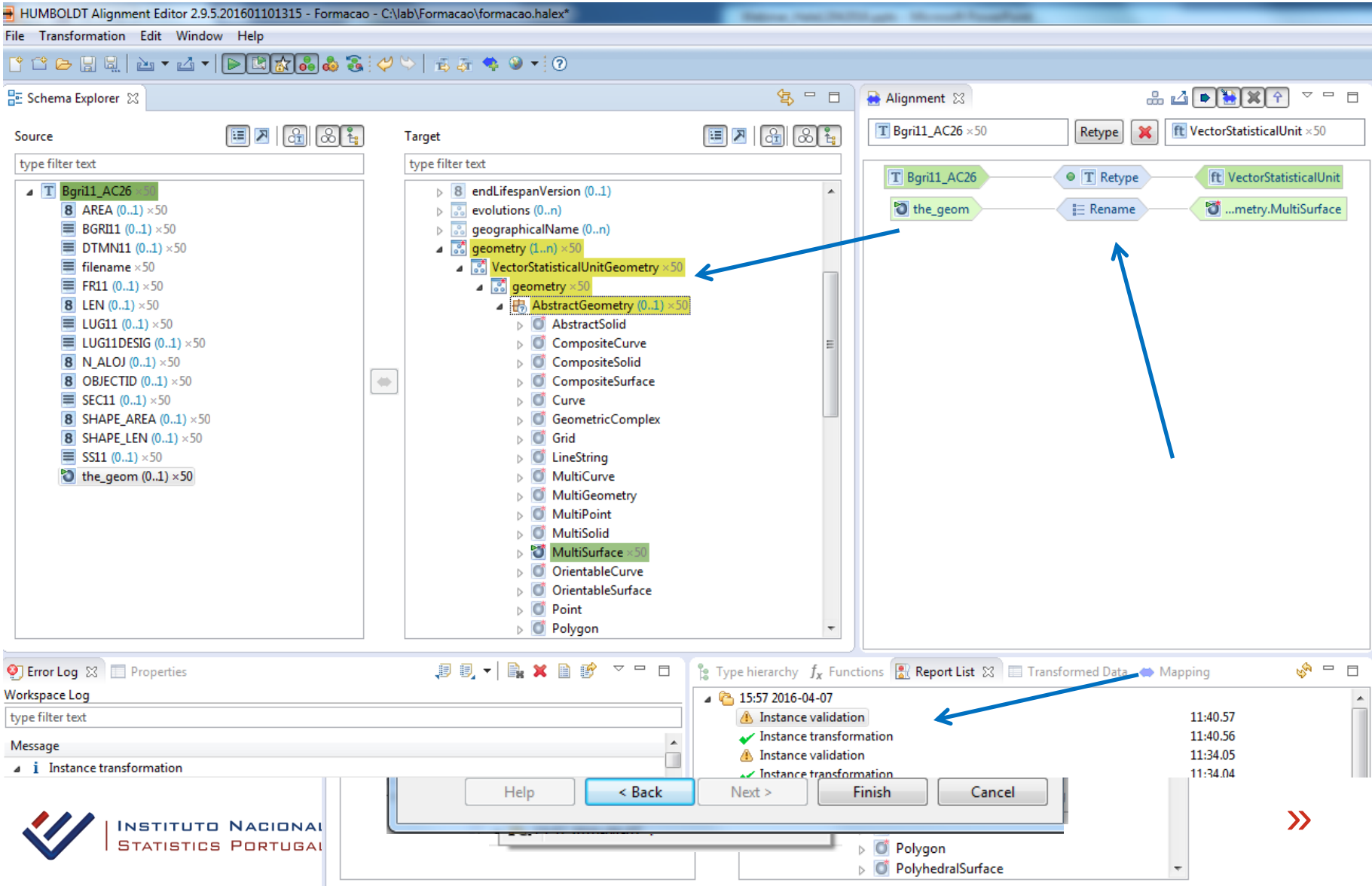
Definir Mapping rules



- **Cardinalidade**
  - Para cada elemento existe informação sobre a sua cardinalidade
  - O valor *n* (0..n) representa um número ilimitado máximo de ocorrências
  - Se não existe informação sobre a cardinalidade significa que a propriedade ocorre apenas uma vez
  - Quando a cardinalidade é (0..1) significa que podem ou não existir ocorrências, ou seja é um elemento opcional.



Definir Mapping rules



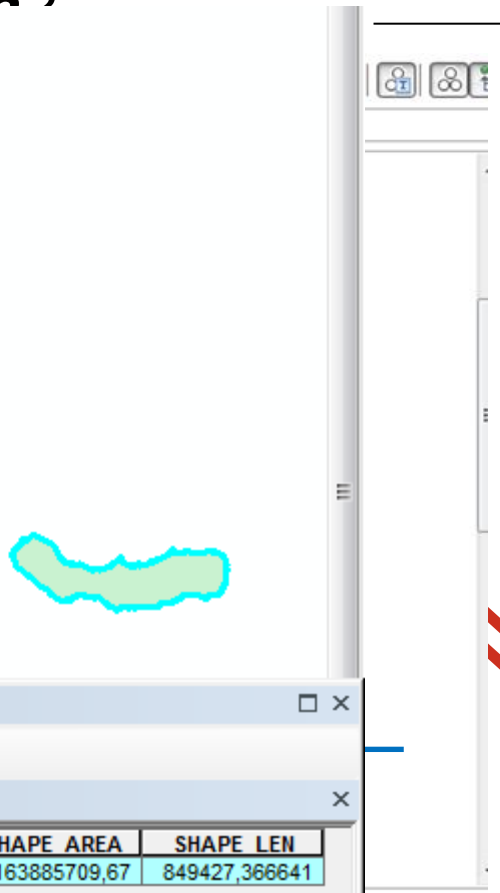
Definir Mapping rules

- ***Geometry***

Tip!

- Porquê ***Multisurface*** e não ***polygon***?

Se na existirem apenas harmoniza ignorados



Table

NUTSI_15_AC26

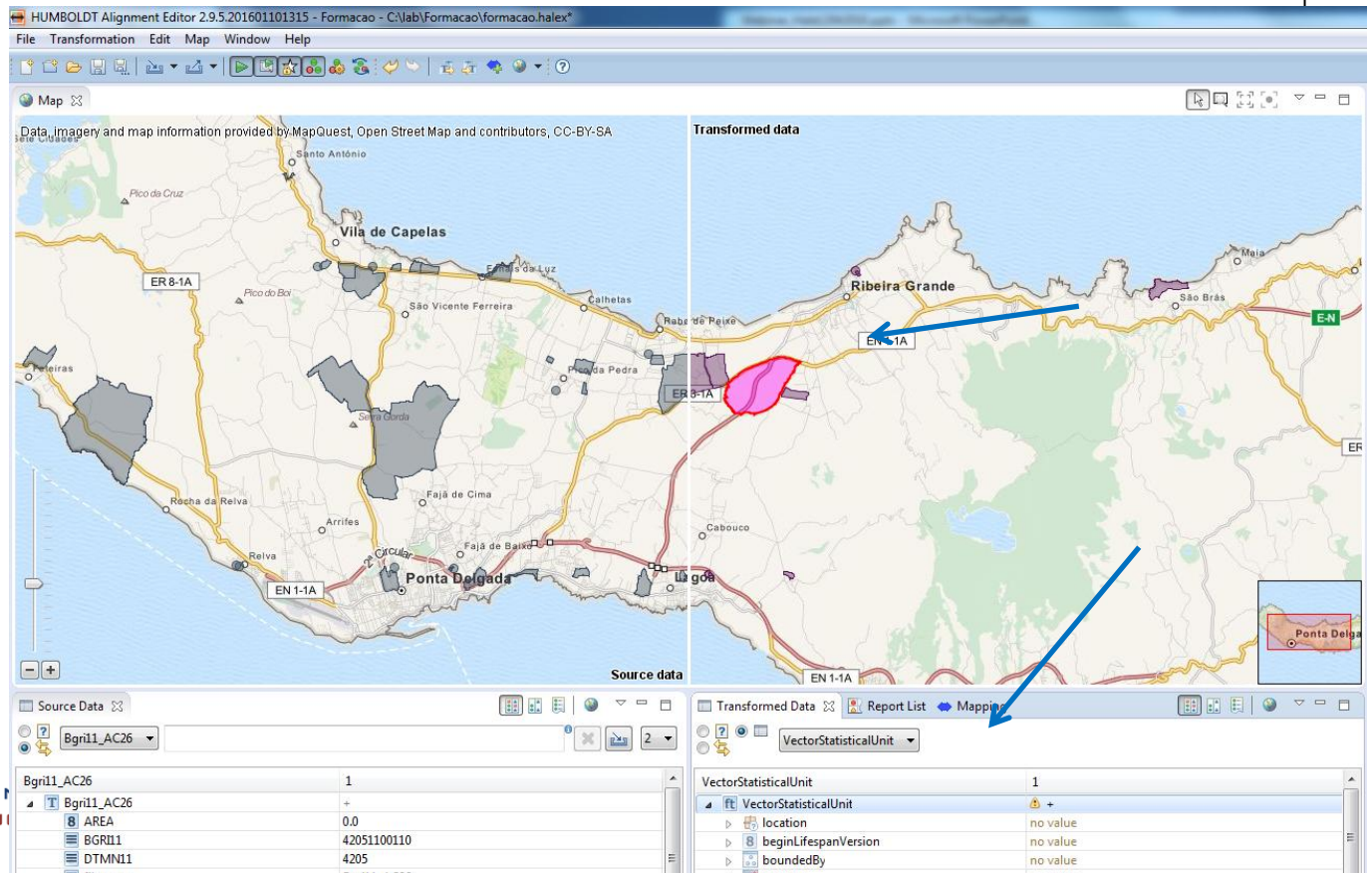
FID	Shape *	OBJECTID	NUTS1_15	NUTS1_15DE	SHAPE_AREA	SHAPE_LEN
0	Polygon	1	PT2	Região Autónoma dos Açores	2163885709,67	849427,366641

1 (1 out of 1 Selected)

NUTSI_15_AC26

Definir Mapping rules

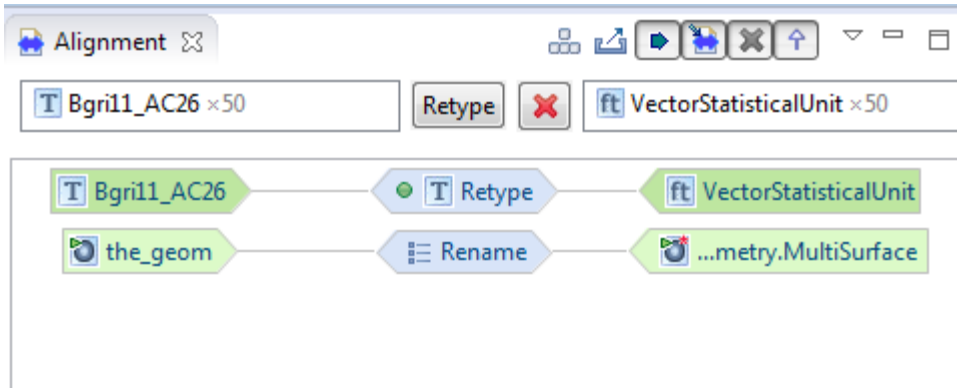
- ***Map View***
  - É possível ver a ***transform data***, geometria e atributos

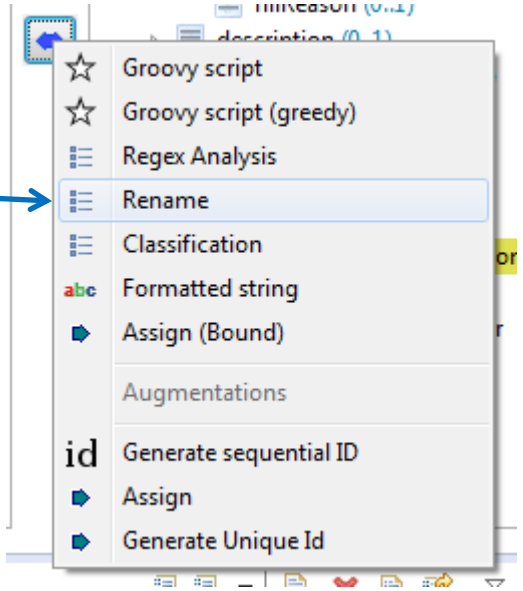


Definir Mapping rules

Geoserver

- **Função *rename***
  - Permite copiar uma propriedade qualquer do ***Source Schema*** para o ***target Schema***
  - Foi utilizada na geometria, mas pode ser utilizada em qualquer propriedade

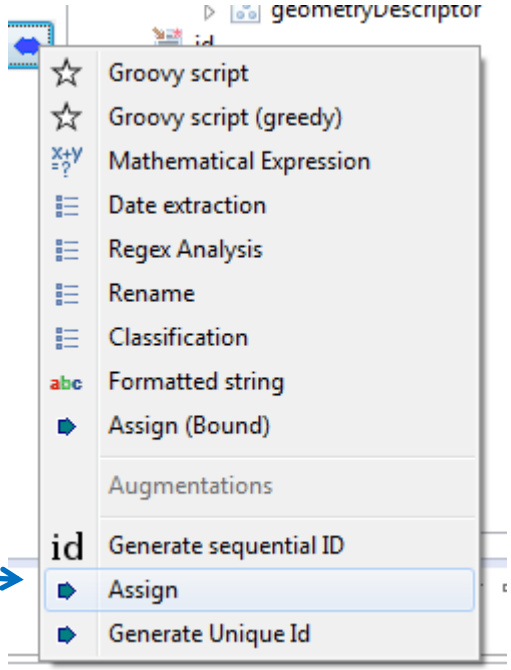




Definir Mapping rules

Geoserver

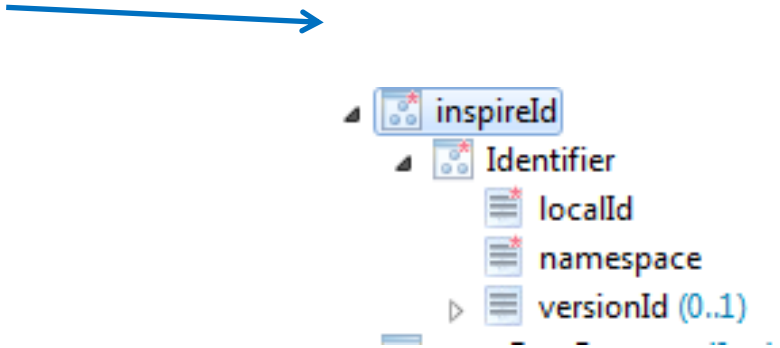
- **Função *Assign***
  - Permite atribuir um valor a uma propriedade do ***target Schema,*** que não exista no ***Source Shema***



Definir Mapping rules

Geoserver

- **Função *Assign***
  - Exemplo da função ***Assign*** com campo ***InspireId***
  - ***InspireId*** é um campo complexo constituído por 3 campos ***LocalId***, ***NameSpace*** e ***VersionId,*** em que os dois primeiros são obrigatórios.



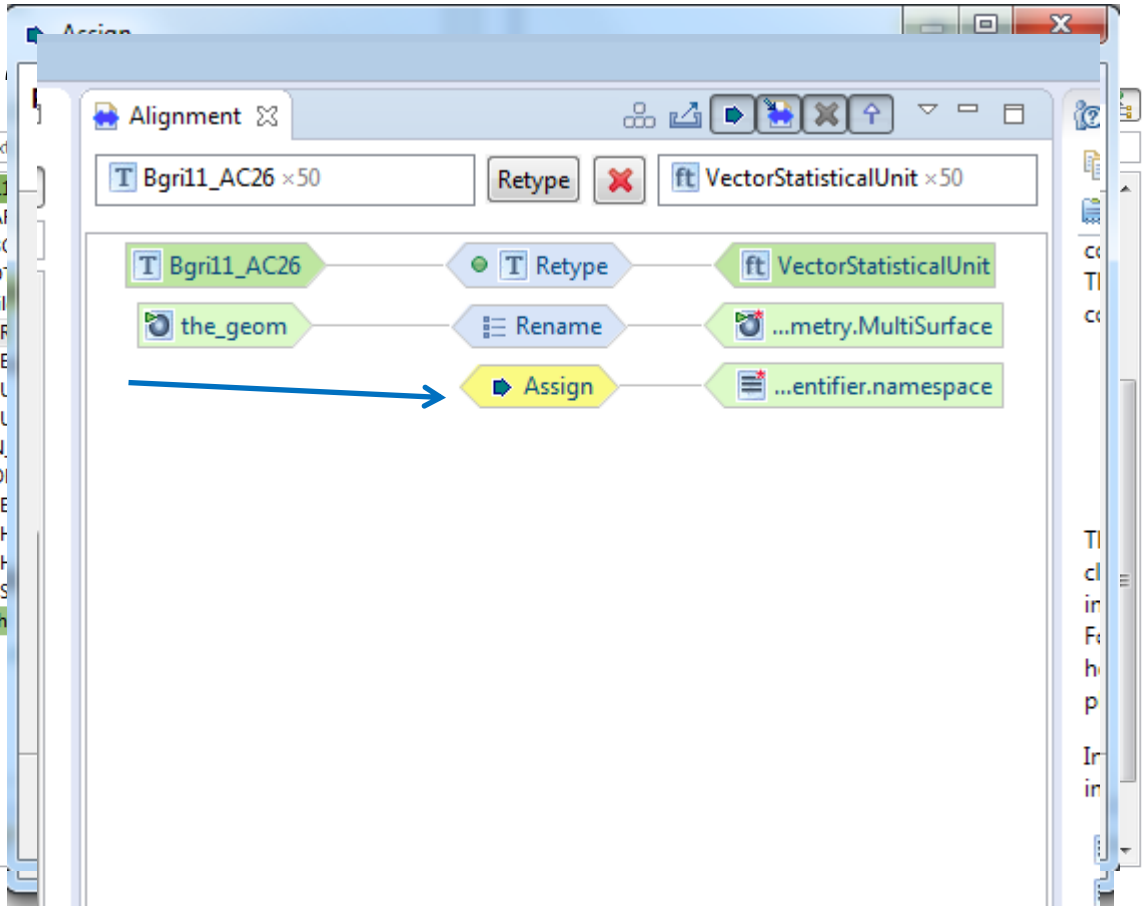
Definir Mapping rules

- **Função *Assign***

Geoserver

- O atributo ***NameSpace*** não existe na ***Source data*** deve ter o valor http://id.igeo.pt/so/SU/VectorStatisticalUnit

- O pa



- ht
sã
- In
- Es
- Si
- Ol
- lo
orig
- ve

http
http

:alId}/{ver

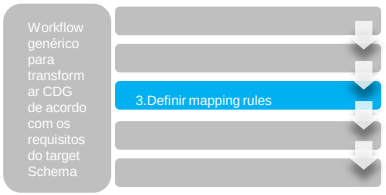
}Inspire}

geográfico

Definir Mapping rules

- **InspireId**
  - Os restantes campos**:**
    - ***LocalId rename*** de um campo com código único ou designação do recurso original + "_" + código do objecto geográfico
      - -- Definition --
      - A local identifier, assigned by the data provider. The local identifier is unique within the namespace, that is no other spatial object carries the same unique identifier.
      - -- Description --
      - NOTE It is the responsibility of the data provider to guarantee uniqueness of the local identifier within the namespace.
    - ***VersionId assign*** de um valor da versão
      - -- Definition --
      - The identifier of the particular version of the spatial object, with a maximum length of **25 characters**. If the specification of a spatial object type with an external object identifier includes life-cycle information, the version identifier is used to distinguish between the different versions of a spatial object. Within the set of all versions of a spatial object, the version identifier is unique.

Definir Mapping rules

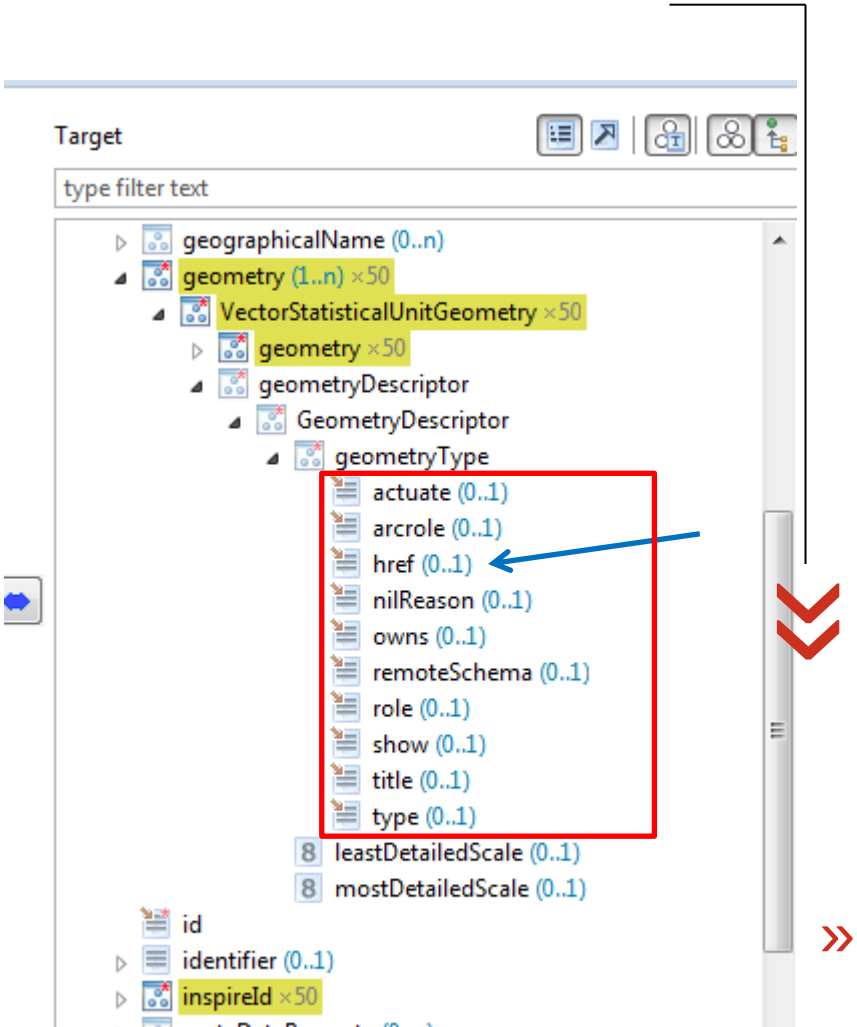


- ***Codelists***
  - São listas de códigos e seus valores, conforme definido nas regras de implementação ***INSPIRE*** relativas à interoperabilidade dos conjuntos e serviços de dados espaciais (Regulamento (UE) no 1089/2010 da Comissão).

Definir Mapping rules

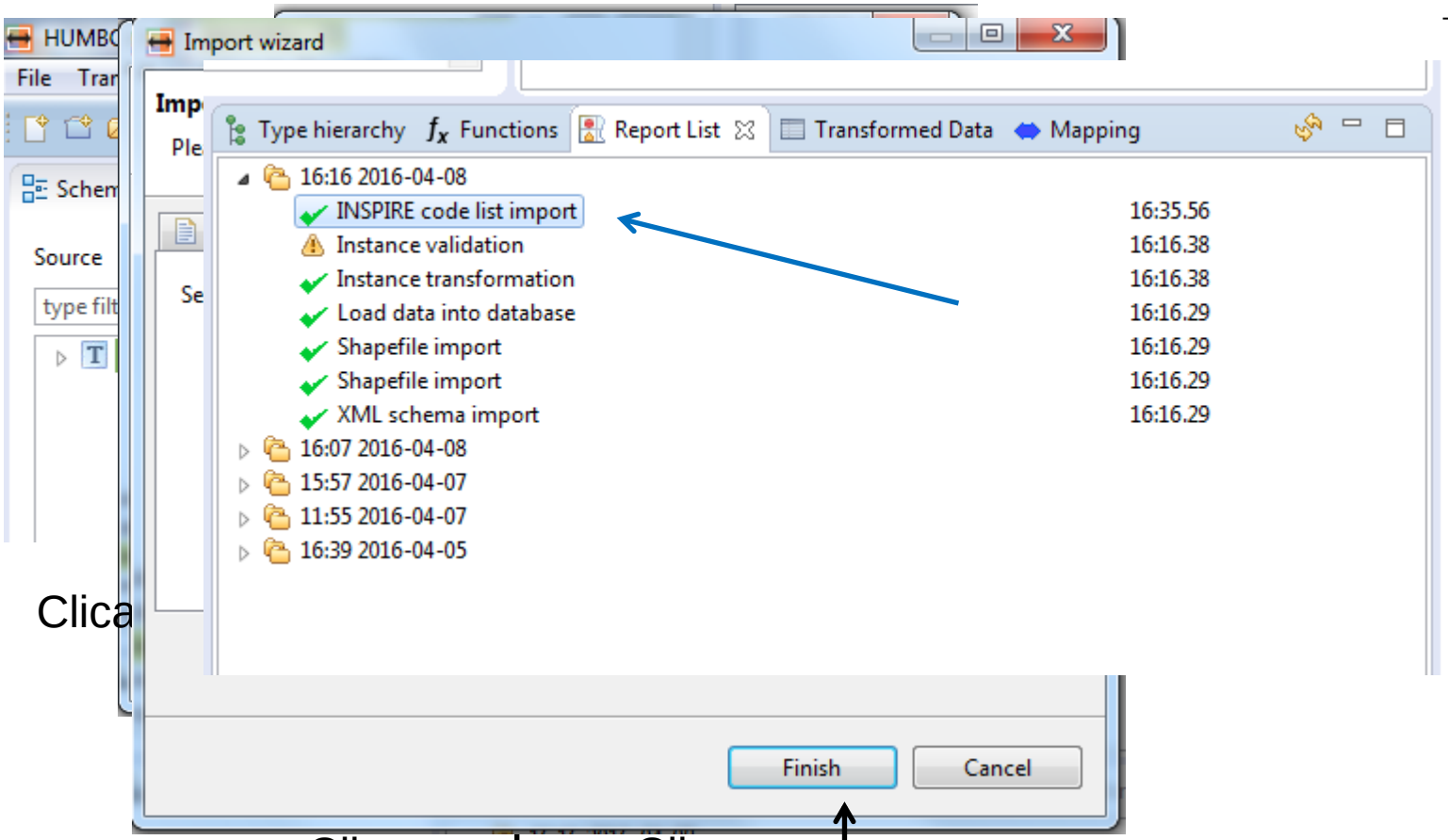
- ***Codelists***
  - No ***target Schema*** esta estrutura de elementos está associada a ***Codelist***
  - O campo ***Href*** irá receber

    os valores



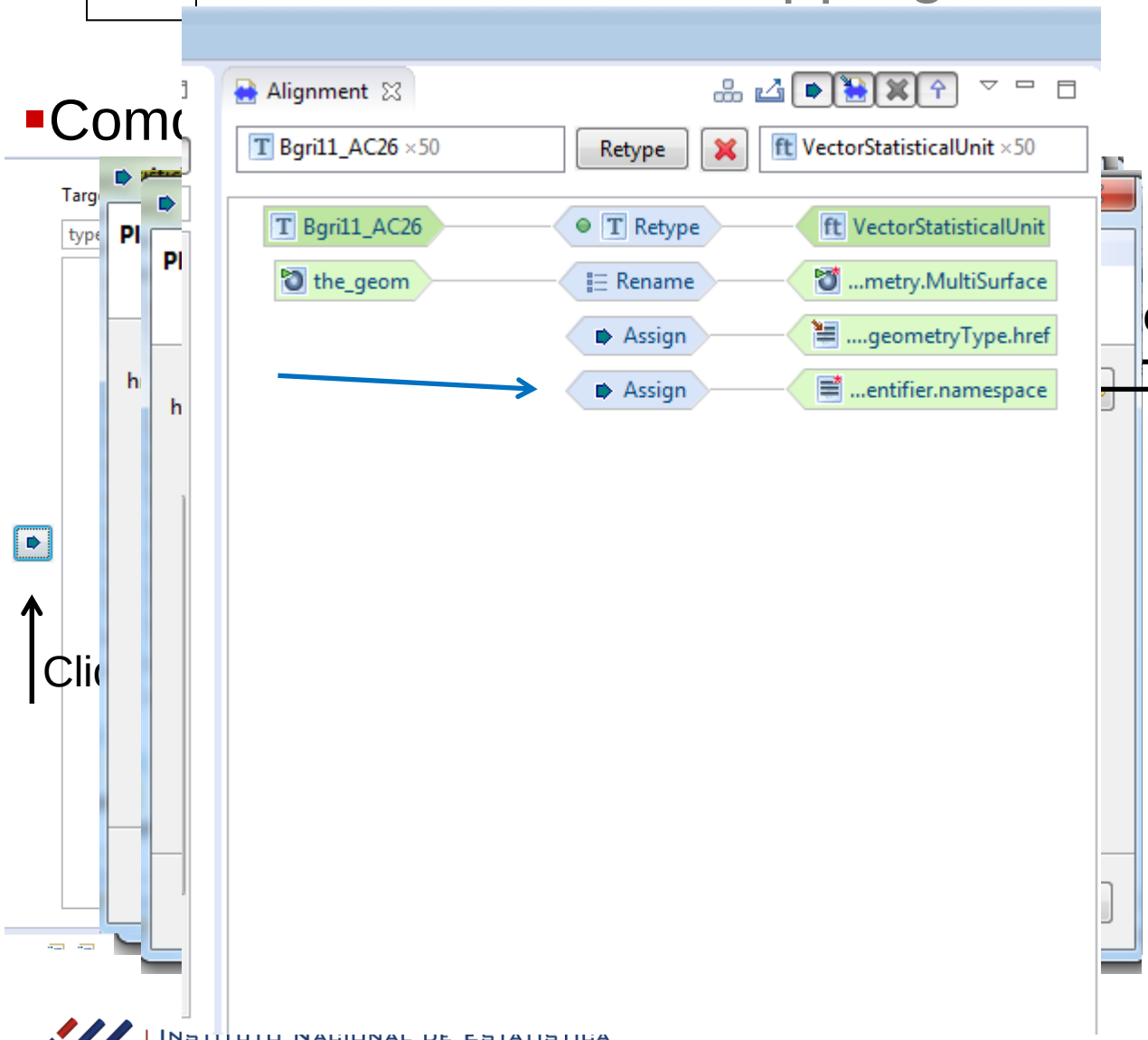
Definir Mapping rules

- Como importar ***Codelists***

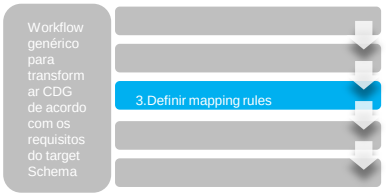


Definir Mapping rules

- Como

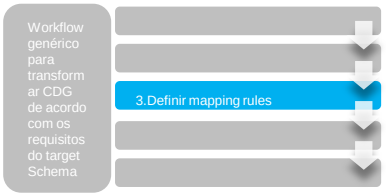


Definir Mapping rules



- ***Codelists Externas***
- Existem codelist mantidas por organizações externas ao Inspire
  - Causas de morte Eurostat
  - Patologia World Health Organization WHO

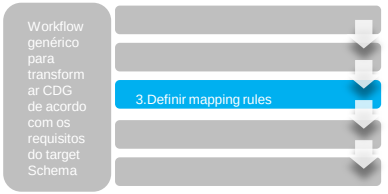
Definir Mapping rules



- ***Codelist ainda sem valores***
- Existem algumas *codelist* sem códigos definidos, por vezes completamente vazias
- *INSPIRE Registry*:
- Padrão em estudo
  - 1. http://registo.igeo.pt/codelist/{NomeListaCodigos}Value
  - 2. http://registo.igeo.pt/codelist/{NomeListaCodigos}Value/{ValorCodigo}

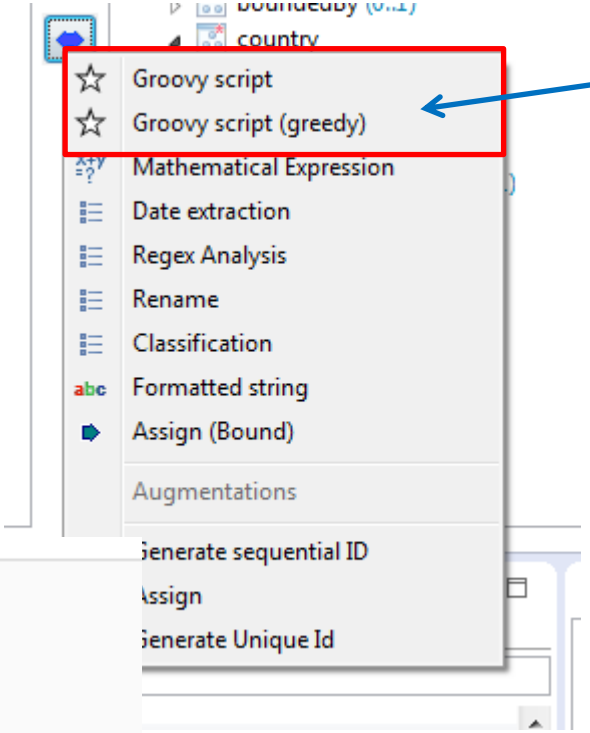
- http://registo.igeo.pt/codelist/AggregationUnitValue/NUTSIII

Definir Mapping rules



Geoserver

- Outras funções
  - ***Groovy Script***
    - Permite o cálculo através de código de um valor de uma propriedade no ***target schema***

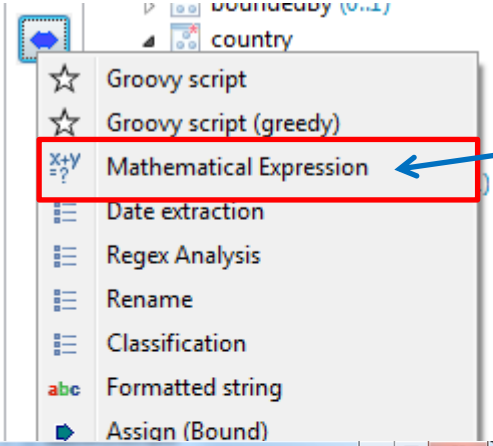


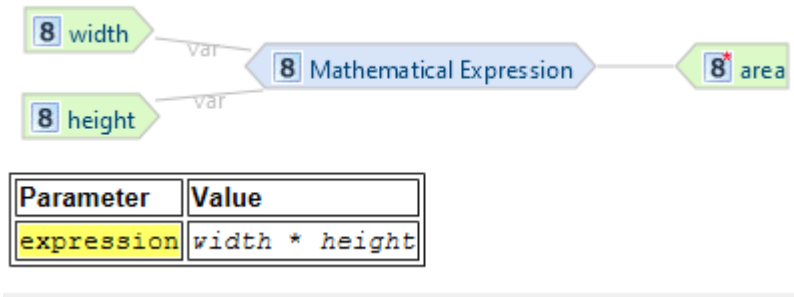

```
// retrieve all values of "result" in the merged object
def list = _source.p.result.values()

// concatenate the values, separated by comma
def str = list.join(',')

// assign the string to the property "resultString" in the target object
_target {
  resultString(str)
}
```

Definir Mapping rules

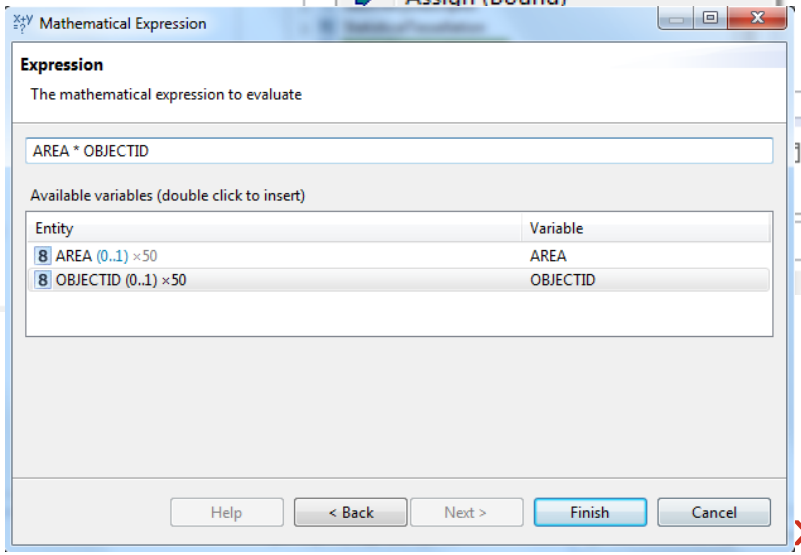
Geoserver

- Outras funções
  - ***Mathematical Expression***
    - Define um valor através de uma expressão matemática
    - Pode utilizar variáveis





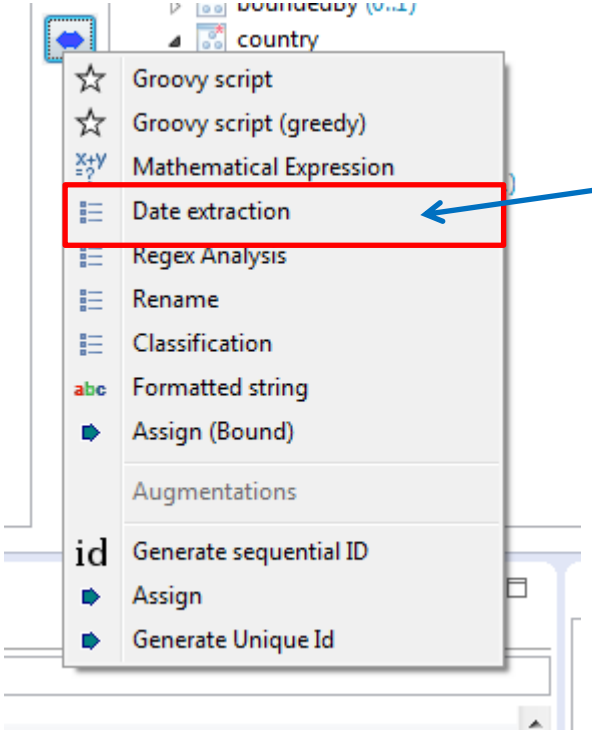
Parameter	Value
expression	width * height



Definir Mapping rules

Geoserver

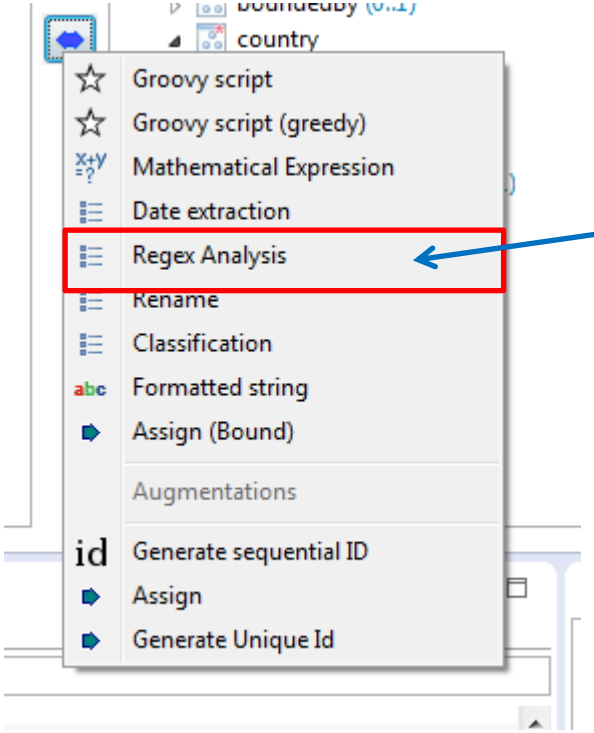
- Outras funções
  - ***Date Extraction***
    - Extrai uma data de uma string
    - **yyyy-MM-dd HH:mm:ss**



Definir Mapping rules

Geoserver

- Outras funções
  - ***Regex Analisys***
    - Aplica uma *regular expression* a uma *string*



Definir Mapping rules

Let us assume that the source property represents a date of the format:

20081209

and that the target property needs a date in the format :

YYYY-MM-DD HH:MM:SS

We can use a regular expression like:

([0-9]{4})([0-9]{2})([0-9]{2})

and an output format:

{1}-{2}-{3} 00:00:00

to achieve our goal.

In fact the result of the above example will result in:

2008-12-09 00:00:00

Explained

The regular expression used defines 3 groups, which are separated by the round brackets:

Regex Group	Explained	Catched part of **20081209**
([0-9]{4})	catches 4 numbers between 0 and 9	2008
([0-9]{2})	catches 2 numbers between 0 and 9	12
([0-9]{2})	catches 2 numbers between 0 and 9	09

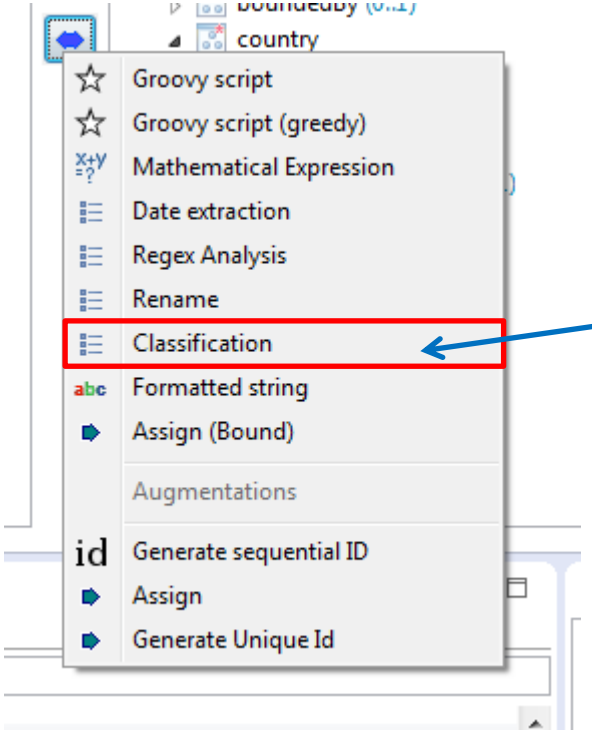
The output can then be formatted concatenating groups with any string.

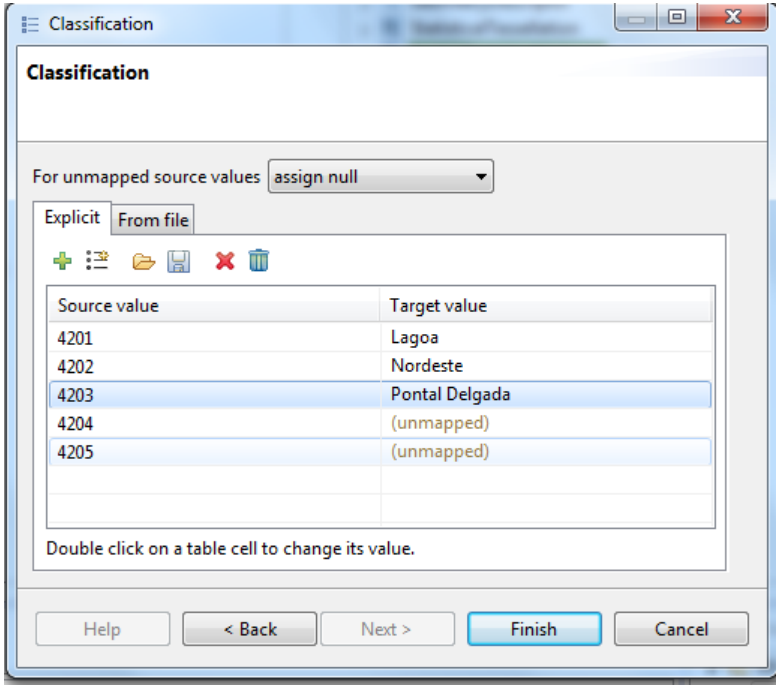
The format **{1}-{2}-{3} 00:00:00** simply defines that the first three groups will be concatenated by a dash and that a default time will be added at the end of the resulting string.

Definir Mapping rules

Geoserver

- Outras funções
  - Classification
    - Reclassifica valores



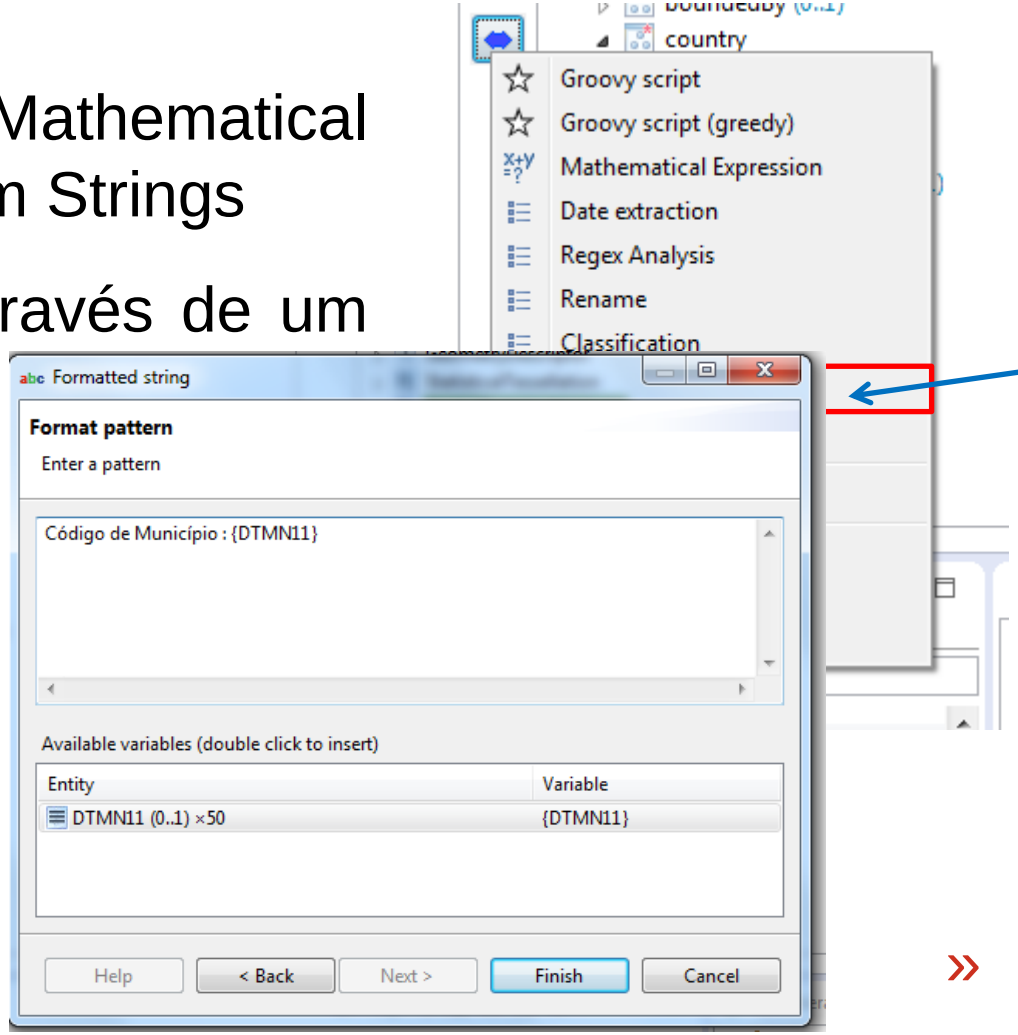


Definir Mapping rules

- Outras funções

Geoserver

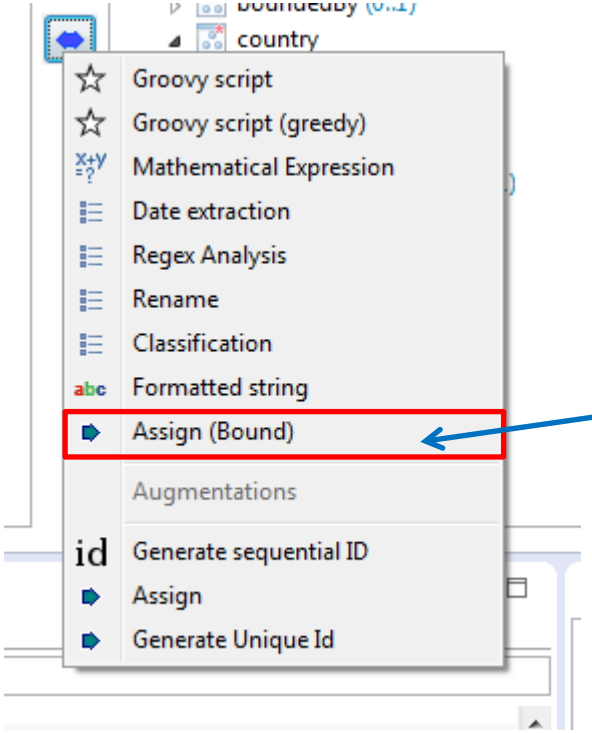
- ***Formatted String***
  - Semelhante ao Mathematical Expression mas com Strings
  - Cria uma string através de um padrão e variáveis



Definir Mapping rules

- Outras funções
  - ***Assign Bound***
    - Atribui um valor a uma propriedade (***target***) se uma propriedade na fonte estiver presente.

Geoserver

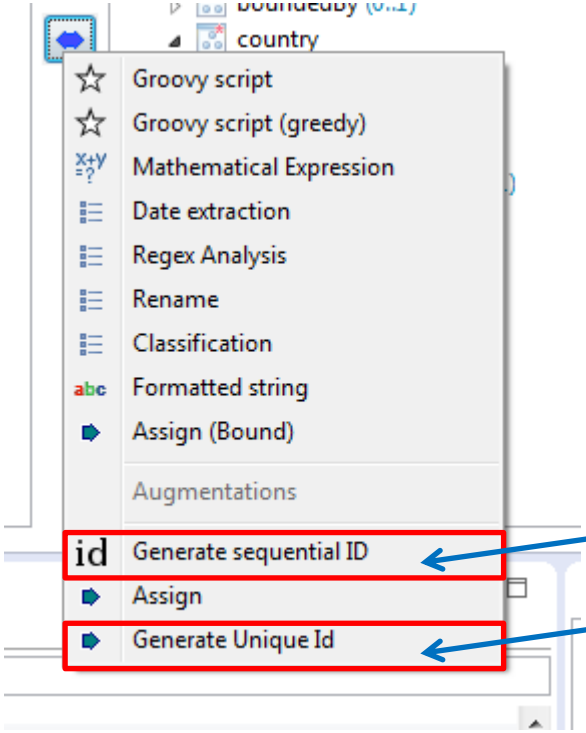


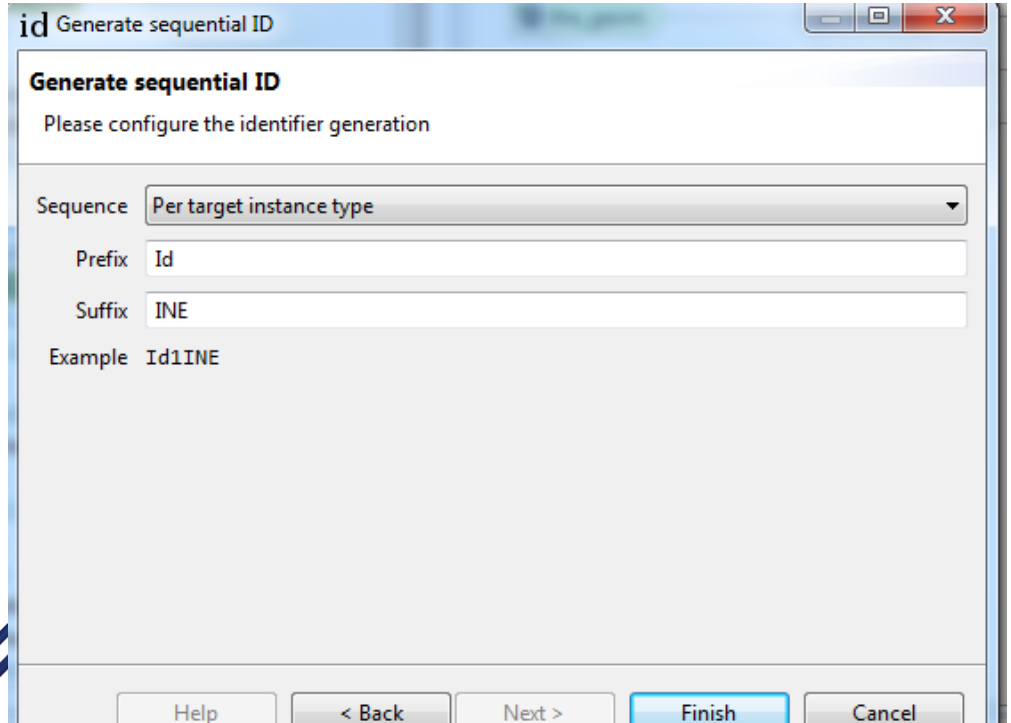
Definir Mapping rules

- Outras funções

Geoserver

- ***Generate Sequential ID*** e ***Generate UniqueID***
  - Cria um identificador sequencial numérico

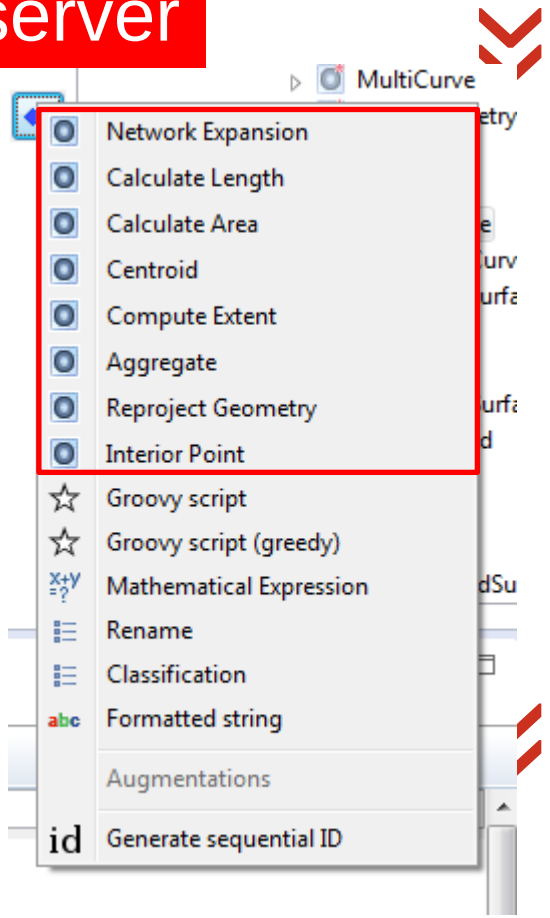




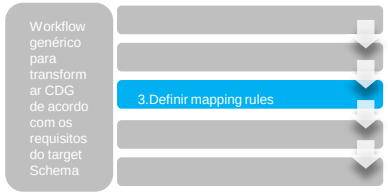
Definir Mapping rules

- **Funções de Geometria**
  - Ficam ativas se clicar na geometria do ***source schema***
  - Funções de geometria simples na maioria.
  - ***Network expansion*** permite converter para um diferente tipo de goemetria (linhas -> Pontos)
  - ***Aggregate*** - Junta geometrias

Geoserver



Definir Mapping rules

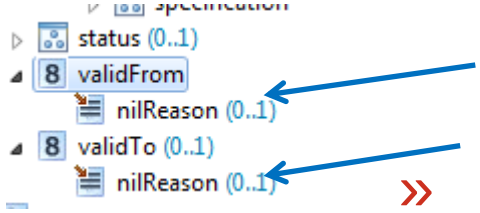


- ***Void***

- Algumas propriedades no ***target Schema***, especialmente as obrigatórias podem receber o valor de ***void***

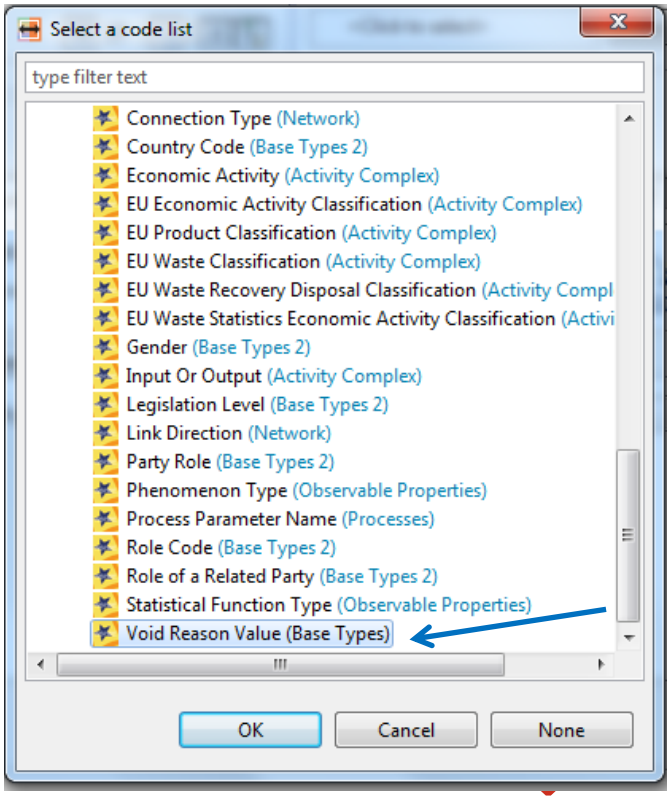
- As propriedades ***voidable,*** podem ter uma subpropriedade ***nilReason***

Para caracterizar a ***nilReason*** podemos utilizar uma ***codelist (void reason value)*** está no folder ***Others***.



Definir Mapping rules

- ***Void reason value***
- Pode conter 3 valores:
- ***Unknown***
  - The correct value for the specific spatial object is not known to, and not computable by, the data provider. However, a correct value may exist.
  - Example When the elevation of the water body above the sea level of a certain lake has not been measured, then the reason for a void value of this property would be 'Unknown'.



Definir Mapping rules

- ***Void reason value***

- ***Unpopulated***

The characteristic is not part of the dataset maintained by the data provider. However, the characteristic may exist in the real world.

EXAMPLE When the "elevation of the water body above the sea level" has not been included in a dataset containing lake spatial objects, then the reason for a void value of this property would be 'Unpopulated'.

- ***Withheld***

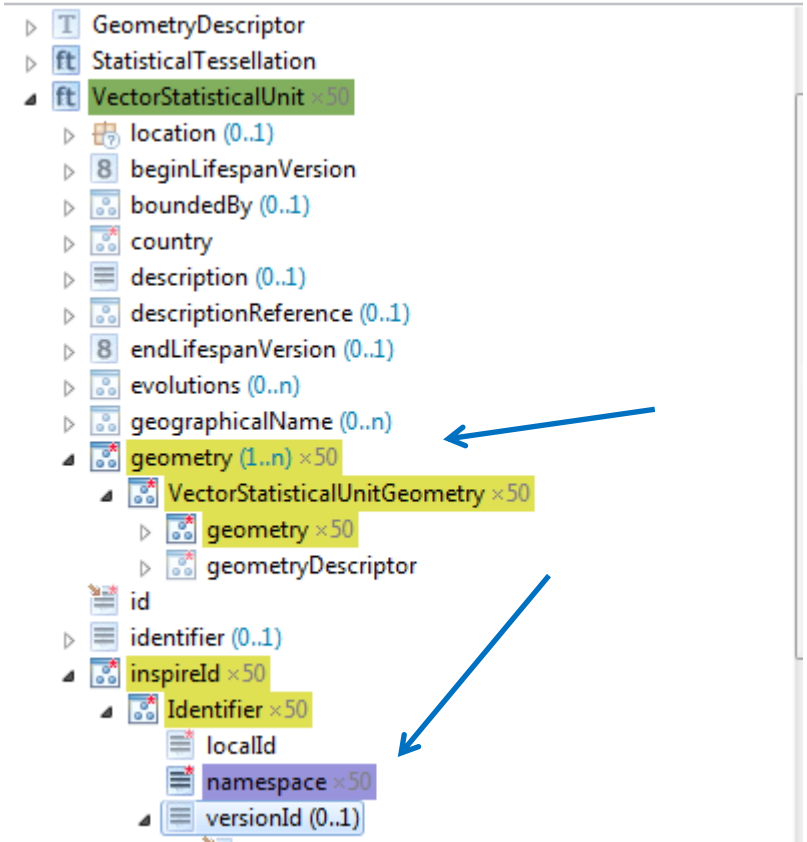
The characteristic may exist, but is confidential and not divulged by the data provider.

Definir Mapping rules

Sistema de cores dos elementos

- Para identificar os elementos do ***schema*** já mapeados o ***Hale*** utiliza um sistema de cores que nos permitem ter uma noção do trabalho já realizado

city	Not mapped
street	Mapped explicitly with a relation
address	Mapped implicitly due to the mapping of a sub-property
type	Value assignment independent of the source schema



Definir Mapping rules

- **Sistema de validação**
  - Verifica se existem inconsistências no mapeamento nomeadamente atributos obrigatórios no ***target schema*** não mapeado, ou restrições ignoradas.

Live transformation button

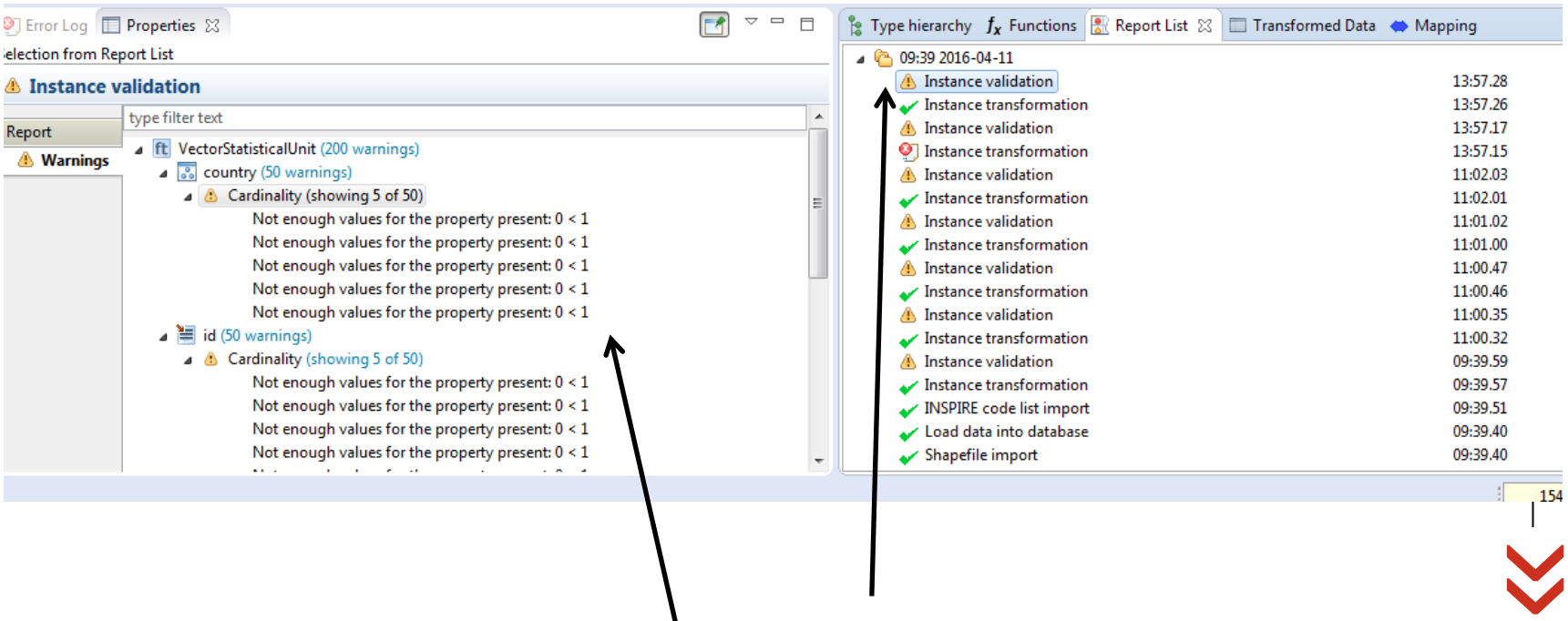
Permite que as instancias carregadas no projeto sejam transformadas sempre que haja um mapeamento.

Validate transformed instances automatically

Executa uma validação das instâncias carregadas sempre que haja um mapeamento.

Definir Mapping rules

- **Sistema de validação, com erros**



Total e lista de erros

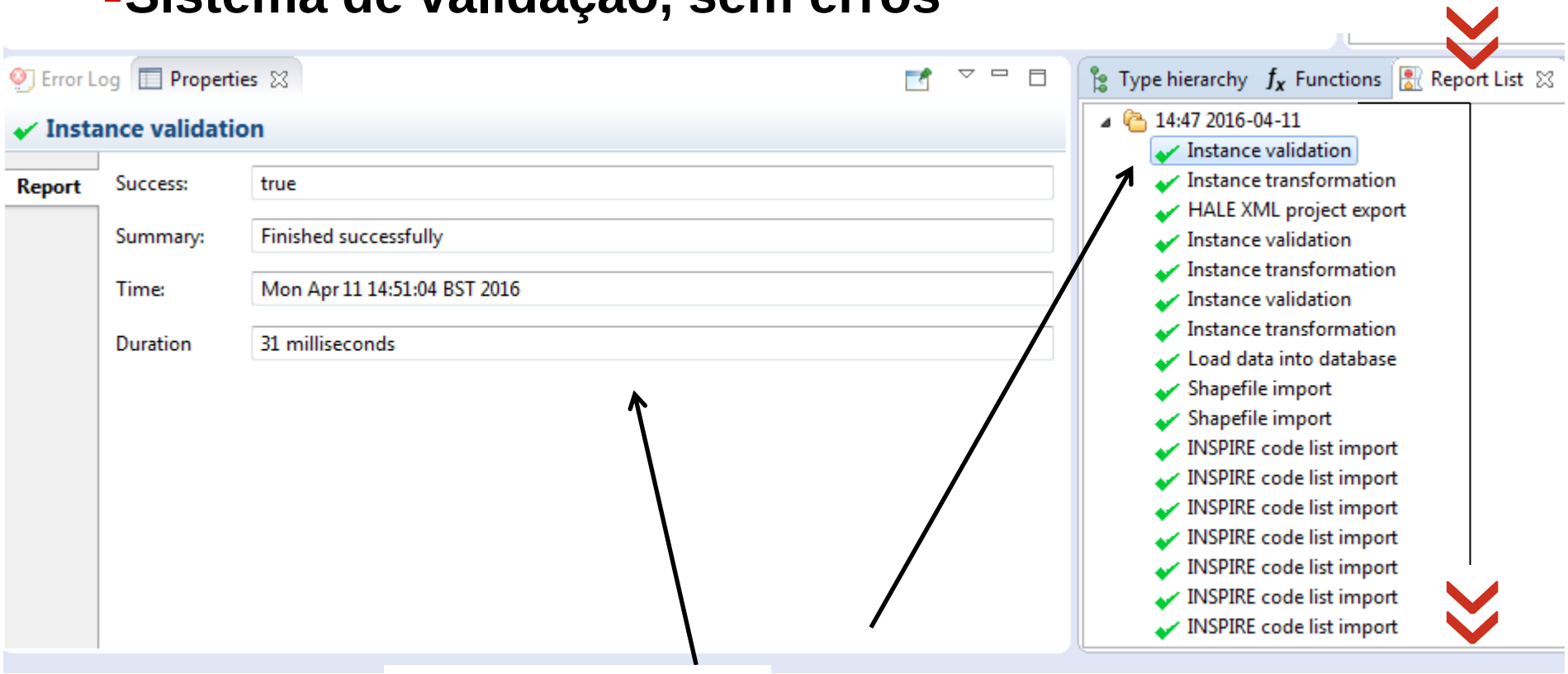
Instance validation

Símbolo *Warning*

Significa que existem erros

Definir Mapping rules

- **Sistema de validação, sem erros**



Lista de erros deixa de existir

Instance validation

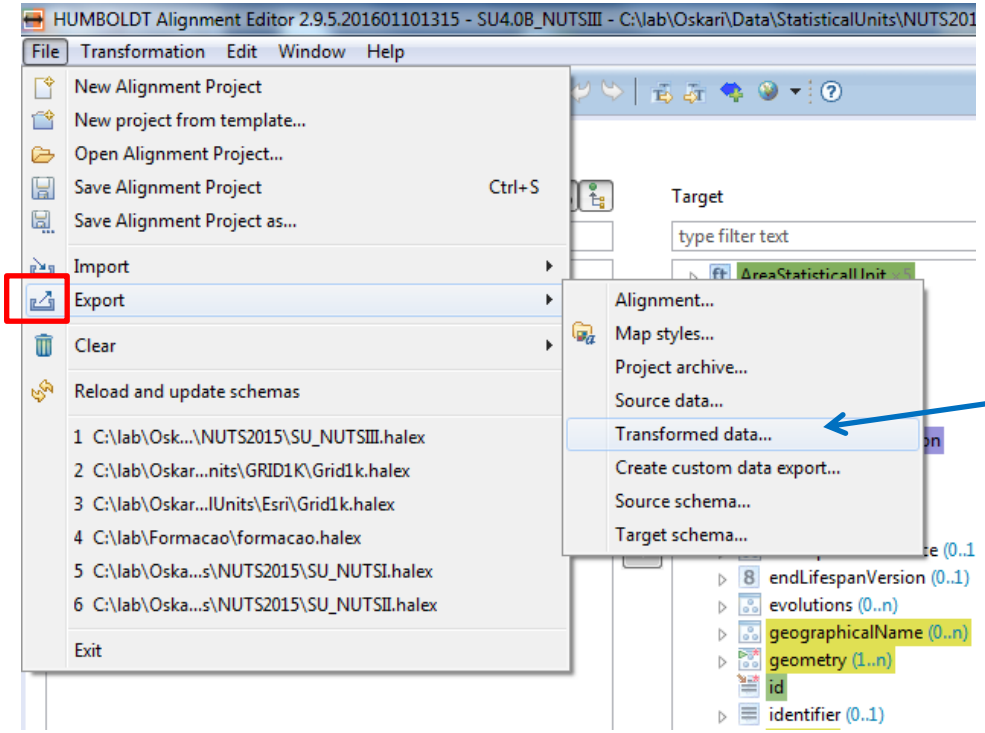
Símbolo = visto (*checked*)

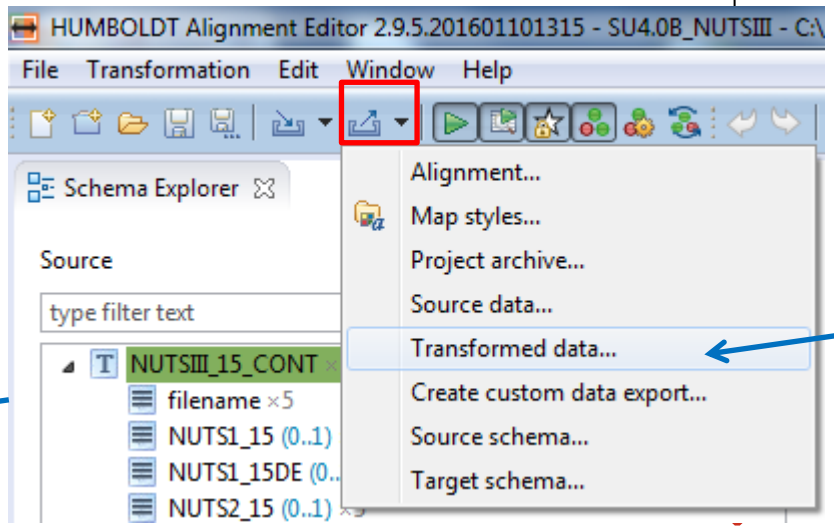
Exportar dados transformados

- **Exportar dados quando:**
  - não existem erros reportados no ***Instance validation***
  - não existe mais nenhum mapeamento para realizar
  - não existir ***instance sampling***

Exportar dados transformados

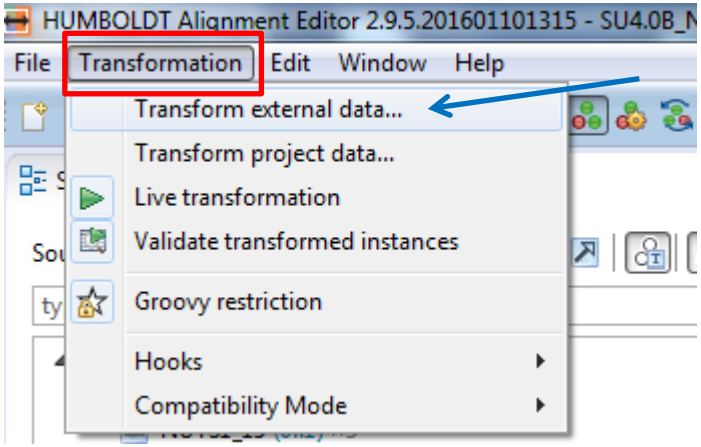
- **Existem duas formas para transformar dados**
  - **1.**Transformar as instâncias carregadas no projeto





Exportar dados transformados

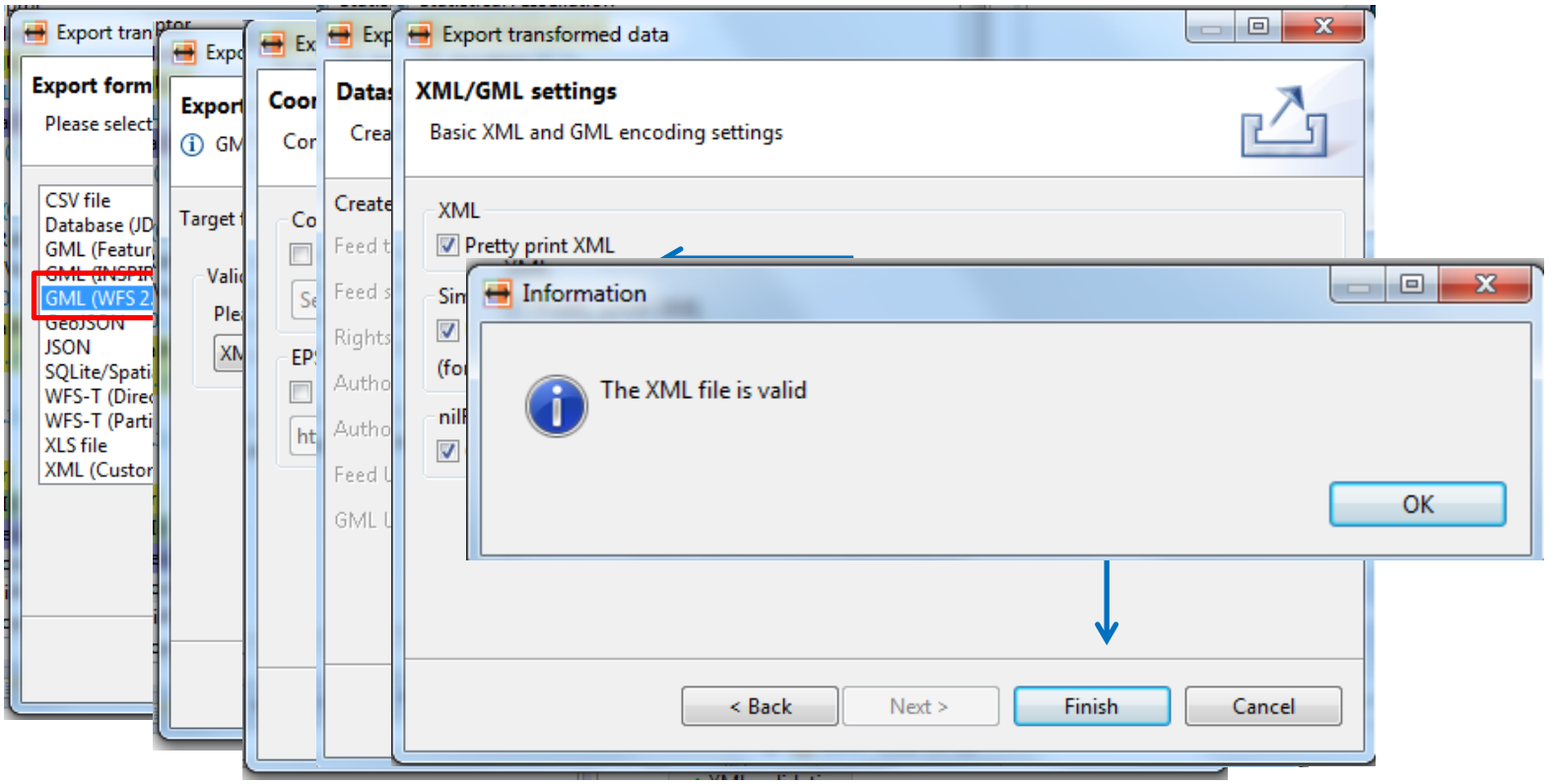
- **Existem duas formas para transformar dados**
  - **2.**Transformar dados externos



- Não se recomenda este método para dados com muitas instâncias: no caso da BGRI (~200 000 registos) não chega a terminar.
- Utiliza-se para partilhar o projeto Hale com dados de Portugal continental e R.A.'s

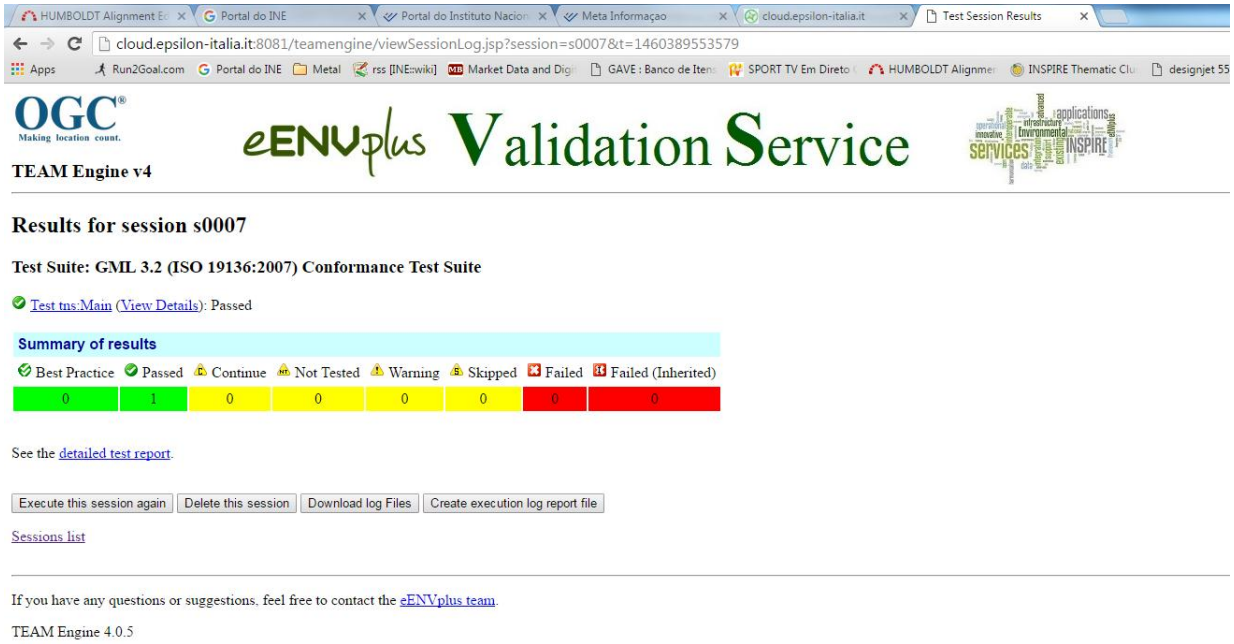
Exportar dados transformados

- **Tipo de formato e outras opções**



Validação dados transformados

- ***eENVplus Validation Service***
- http://cloud.epsilon-italia.it/eenvplus_new/
- Só 4 temas validam com ***Schematron***



Validação dados transformados

- ***Oxygen***
- *Software* com custos