

# Processos de harmonização e agregação de dados no âmbito do projecto GIS4EU.

**Reis, Rui M. P.**

(rui.reis@igeo.pt)

Instituto Geográfico Português, Rua Artilharia Um, 107, 1099-052 Lisboa

As metodologias criadas para a harmonização de dados no âmbito do INSPIRE, nomeadamente, a transformação dos dados dos produtores no modelo de dados INSPIRE, e o processo de agregação de dados e sua exposição a escalas de nível europeu, nacional, regional e local constituem os principais resultados do projecto GIS4EU. Nesta comunicação descrevem-se estas metodologias bem como o papel que o IGP desempenhou no seu desenvolvimento.

O primeiro passo do processo de harmonização de dados consiste numa documentação efectiva dos dados que permita servir de base aos passos posteriores da metodologia. Esta descrição dos dados inclui a criação de metadados, listas com as definições de entidades e atributos e listas codificadas bem como a criação dos diagramas UML descrevendo as relações entre entidades.

Na remodelação de dados dos produtores no modelo de dados INSPIRE trabalharam-se as temáticas hidrografia, rede de transportes e divisão administrativa, estando estas temáticas incluídas no anexo I da directiva. Os resultados deste processo foram usados na fase de testes das especificações de dados INSPIRE. O IGP foi responsável pela edição do relatório do processo de transformação dos dados dos produtores no modelo de dados INSPIRE para a temática hidrografia, tendo participado activamente nos relatórios para as temáticas rede de transportes e divisão administrativa.

O processo de remodelação dos dados disponíveis no projeto no modelo de dados INSPIRE é composto por vários passos iniciando-se pelo estabelecimento de correspondências entre as entidades e atributos dos modelos de dados em comparação. Nem todas as correspondências são do mesmo tipo distinguindo-se três tipos de correspondência possíveis, nomeadamente: (1) Correspondência exata, (2) Correspondência com diferenças dignas de realce e (3) Correspondência complexa mas possível após processamento de dados. Deste modo, identificaram-se os subconjuntos de entidades e atributos dos modelos de dados dos produtores que correspondiam com o modelo de dados INSPIRE. De seguida fez-se o levantamento dos problemas encontrados no processo seguido da identificação dos elementos considerados não relevantes no modelo de dados INSPIRE. Finalmente, propuseram-se novas entidades nos casos em que tal foi julgado conveniente.

Trabalhou-se também o tema altimetria, pertencente ao anexo II, com o objectivo de que os resultados deste processo de harmonização sejam mais tarde usados aquando da criação da especificação de dados para esta temática no âmbito do INSPIRE. Neste caso criou-se de raiz um modelo de dados comum tendo por base os modelos de dados dos conjuntos de dados geográficos desta temática disponíveis no projecto.

A metodologia de agregação e exposição de dados baseou-se na definição de regras e directrizes de agregação e de visualização para os níveis europeu, nacional, regional e local.

O IGP foi responsável pelo grupo de trabalho temático em regras e directrizes de agregação para a temática divisão administrativa tendo também participado activamente nos trabalhos para as temáticas rede de transportes e hidrografia.

Os resultados do projecto constituem um importante contributo para a implementação da directiva INSPIRE, nomeadamente, para o processo de harmonização de dados e para o processo de agregação e visualização de dados a diferentes escalas.

## **PALAVRAS-CHAVE**

GIS4EU, harmonização, INSPIRE, agregação.

## **1. INTRODUÇÃO**

O objetivo do projeto GIS4EU é harmonizar conjuntos de dados geográficos de diferentes produtores de vários países Europeus em quatro temas (hidrografia, divisão administrativa, rede de transportes e altimetria) e assegurar a sua interoperabilidade e acessibilidade a várias escalas em várias línguas e através das fronteiras, tendo em conta os standards internacionais e os requisitos da directiva INSPIRE.

Os parceiros fornecendo dados ao projectos distribuem-se por cinco países e pertencem aos níveis nacional, regional e local. Estes parceiros produtores de dados trabalham em conjunto com parceiros tecnológicos e de investigação de outros cinco países e com uma associação de utilizadores de informação geográfica.

O projecto GIS4EU põe a ênfase nos dados e assim é dada muita importância aos produtores neste projecto. No desenvolvimento dos trabalhos o primeiro passo consistiu na documentação dos dados por forma a formalizarem-se os modelos de dados a usar nas fases seguintes. O objectivo é o de criar um modelo de dados comum ou a adopção do modelo de dados INSPIRE seguido do desenvolvimento de regras e directrizes de harmonização, agregação e exposição dos dados disponíveis no projecto.

O fluxo de trabalho usado no projeto GIS4EU consistiu em:

1. Criação de descrições homogéneas de dados
2. Criação de MD comum para a altimetria e transformação no modelo de dados INSPIRE para as restantes temáticas de trabalho.
3. (Remodelação) Transformação dos MD dos produtores no MD INSPIRE para a hidrografia, divisão administrativa e rede de transportes.
4. Agregação de dados e sua visualização e exploração a escalas Europeia, Nacional, Regional e Local.

A descrição dos dados consiste na criação de registos de metadados seguindo as especificações INSPIRE para metadados, criação de listas de entidades, listas de atributos e listas codificadas bem como diagramas UML registando as relações entre entidades (Diagramas entidade-associação).

Por forma a descrever as várias fases de trabalho vão-se usar dois exemplos em que o IGP teve especiais responsabilidades no desenvolvimento dos trabalhos. Assim, para a transformação no modelo de dados INSPIRE, que constitui a fase inicial da metodologia de trabalho, vai usar-se como exemplo ilustrativo o tema hidrografia e para a agregação e visualização a várias escalas vai usar-se o exemplo da divisão administrativa.

## **2. TRANSFORMAÇÃO NO MODELO DE DADOS INSPIRE**

No projecto existem onze Conjuntos de Dados Geográficos (CDG) de nove produtores diferentes do tema hidrografia para teste ([4] e[5]) e estes dados podem-se classificar em:

- Dados locais a escalas grandes (1/1000, 1/2000 e 1/50000) da CGE (Comune di Genova) e do MAV (Magistrato alle Acque di Venezia).
- Dados usando as especificações DBPrior10k, fornecidos por produtores de âmbito regional

italianos (Regione Liguria, Regione Piemonte/CSI, Regione Veneto, Regione Friuli Venezia Giulia/INSIEL).

- Dados usados para relatório no contexto da Diretiva Quadro da Água, à escala 1/50 000 fornecidos pelo VUGK - Instituto de Investigação em Geodesia e Cartografia de Bratislava.
- Dados cartográficos genéricos fornecidos pelo ICC - Institut Cartogràfic de Catalunya, a escalas 1/5 000 e 1/50 000 e pelo IGP - Instituto Geográfico Português a escalas 1/250 000 e 1/1 000 000.

Seguindo a metodologia definida em DS-D2.6 (disponível, tal como o texto da diretiva INSPIRE, no arquivo de documentos do site <http://inspire.jrc.ec.europa.eu/>) a versão 2 do MD INSPIRE foi construído com base em 3 casos de utilização:

- Mapping: incluindo a representação das principais entidades físicas relacionadas com a água e objectos relacionados.
- Reporting: Relatório à Comissão da qualidade e quantidade da água, no âmbito da DQA (Directiva Quadro da Água).
- Spatial Analysis and Modelling: inclui a criação de informação com base na rede hidrográfica usando técnicas SIG.

Esta versão (versão 2) do modelo de dados INSPIRE é composto por 4 pacotes:

- -ManagementAndReporting: A entidade base é WFDWaterBody,
- -Network: é composto por nós, lados, lados agregados (formando caminhos) e áreas.
- -PhysicalWaters: é composto por 11 entidades (CatchmentArea, Basin, SurfaceWater, Watercourse, StandingWater, LandWaterBoundary, Sea, Foreshore, Riverbank, Wetland e GlacierSnowfield).
- -RelatedObjects: 24 elementos, alguns estão reservados para temas do anexo III, por ex: MonitoringPoint.

Está dividido em 3 sub-pacotes: HydroFacility, HydroPointOfInterest e ManMadeObject.

O caso de utilização "Reporting" dá origem ao pacote "Management and Reporting", o caso de utilização "Spatial Analysis" dá origem ao pacote "Network" e o caso de utilização "Mapping" dá origem aos pacotes "Physical Waters" e "Related Objects".

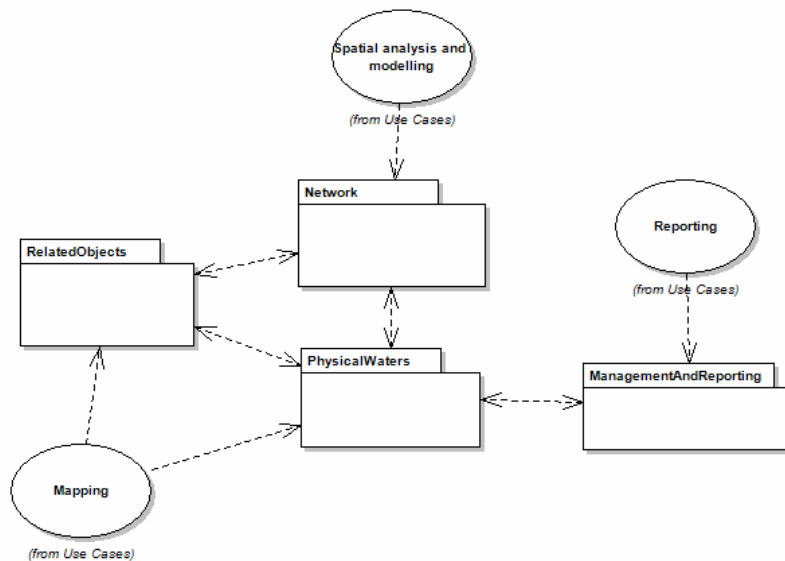


Figura 1 - Relações entre os casos de utilização e o modelo de dados para a hidrografia (fonte: portal INSPIRE)

Os objectivos da análise comparativa são:

- a identificação do subconjunto do modelo de dados e catálogo de entidades INSPIRE aos quais correspondem os conjuntos de dados do GIS4EU
- levantamento dos problemas que podem ocorrer nesta fase do processo de harmonização
- Propor novas entidades para o modelo de dados INSPIRE se for caso disso
- Identificar as entidades e atributos do modelo de dados INSPIRE que possam ser consideradas como não relevantes.

O processo de análise comparativa é composto pelas seguintes fases:

- identificação dos conjuntos de dados e documentação relevante para o tema hidrografia
- tomar como referência as especificações técnicas do modelo de dados INSPIRE para hidrografia
- Análise detalhada de cada conjunto de dados disponível no projecto, inclui:
  - Identificação de entidades correspondentes entre modelos de dados
  - identificação de atributos correspondentes para as entidades seleccionadas no passo anterior
- Analisar o tipo de correspondência entre entidades e/ou atributos
- seleccionar o subconjunto do modelo de dados INSPIRE para ser usado no projecto GIS4EU
- Identificar elementos não existentes no modelo de dados INSPIRE
- Identificar elementos não relevantes no modelo de dados INSPIRE
- Relatar os eventuais problemas encontrados nesta fase do processo de harmonização

O relatório D2.2 [2] contendo a caracterização dos dados do projecto foi usado na identificação dos conjuntos de dados relevantes e na análise detalhada do conteúdo de cada conjunto de dados. As especificações técnicas do modelo de dados INSPIRE para hidrografia usadas foram as da segunda versão, especialmente o modelo UML consolidado.

Na análise das correspondências entre cada um dos conjuntos de dados do projecto com o modelo de dados INSPIRE usou-se uma tabela de correspondências (matching table). Como resultado desta análise obtém-se uma lista das entidades e atributos correspondentes nos dois modelos de dados comparados bem como a classificação desta correspondência em:

A - Entidades e atributos para as quais há correspondência no modelo da dados INSPIRE

A.1 - Correspondência directa

A.2 - Correspondência com diferenças semânticas ou na aquisição de dados que devem ser realçadas

A.3 - Correspondência complexa

B - Entidades e atributos não incluídas no modelo da dados INSPIRE

B.1 - Entidades e/ou atributos que podem ser relevantes para a directiva INSPIRE

B.2 - Entidades e/ou atributos que podem NÃO ser relevantes para a directiva INSPIRE

C - Entidades e atributos do modelo da dados INSPIRE que não estão incluídos no conjunto de dados em análise

C.1 - Entidades e/ou atributos considerados relevantes para a directiva INSPIRE

C.2 - Entidades e/ou atributos considerados NÃO relevantes para a directiva INSPIRE

Foi pedido aos produtores que procedessem a esta análise comparativa sendo os resultados apresentados por meio de uma tabela de correspondências para as entidades e atributos de classe A e um relatório crítico do processo onde as entidades e atributos das classes B e C são referenciados.

## 2.1. RESULTADOS

Package	Matched feature types	Not matched
Network	WatercourseLink, WatercourseNode (2)	4
ManagementAndReporting	WFDSurfaceWaterBody (1)	5
PhysicalWaters	StandingWater, Watercourse, LandWaterBoundary, Sea, Foreshore, Wetland, Riverbank, CatchmentArea (8)	3
RelatedObjects	DamOrWeir, Embankment, ShorelineConstruction, DischargePoint, SpringOrSeep, Pipe, Ford, AbstractPoint, VanishingPoint, SubsurfaceCrossing, SurfaceCrossing (11)	13

Os pacotes "Network" e "Management and Reporting" pertencentes ao modelo de dados INSPIRE têm apenas algumas entidades com correspondência com entidades dos MD dos produtores.

Globalmente, podemos afirmar que foi possível estabelecer correspondências com a maior parte das entidades dos modelos de dados dos produtores com entidades do modelo de dados INSPIRE pertencentes aos pacotes Physical Waters e Related Objects.

Foi possível estabelecer correspondência com cerca de metade das entidades do pacote Related Objects, pertencente ao MD INSPIRE, com entidades dos modelos de dados dos produtores.

INSPIRE Attributes Matched	
<ul style="list-style-type: none"> <li>• <u>Network package:</u> <ul style="list-style-type: none"> <li>– WatercourseLink               <ul style="list-style-type: none"> <li>• id,</li> <li>• geographicalName,</li> <li>• flowDirection ,</li> <li>• centerlineGeometry,</li> <li>• Length</li> </ul> </li> <li>– WatercourseNode               <ul style="list-style-type: none"> <li>• id,</li> <li>• hydroNodeCategory</li> </ul> </li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• <u>PhysicalWaters package:</u> <ul style="list-style-type: none"> <li>– StandingWater               <ul style="list-style-type: none"> <li>• hydroid ,</li> <li>• id,</li> <li>• geographicalName,</li> <li>• Origin,</li> <li>• Elevation,</li> <li>• surfaceArea,</li> <li>• localType</li> </ul> </li> </ul> </li> </ul>
<ul style="list-style-type: none"> <li>• <u>PhysicalWaters package:</u> <ul style="list-style-type: none"> <li>– Watercourse               <ul style="list-style-type: none"> <li>• beginLifespanVersion,</li> <li>• endLifespanVersion,</li> <li>• Origin,</li> <li>• Condition,</li> <li>• Fictitious,</li> <li>• Level,</li> <li>• id,</li> <li>• geographicalName,</li> <li>• length,</li> <li>• LevelOfDetail ,</li> <li>• localType,</li> <li>• waterCourseHierarchy,</li> <li>• persistence</li> </ul> </li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• <u>PhysicalWaters package:</u> <ul style="list-style-type: none"> <li>– LandWaterBoundary               <ul style="list-style-type: none"> <li>• Origin,</li> <li>• waterLevelCategory,</li> <li>• id</li> </ul> </li> <li>– CatchmentArea               <ul style="list-style-type: none"> <li>• area,</li> <li>• hydroid,</li> <li>• geographicalName</li> </ul> </li> <li>– Foreshore               <ul style="list-style-type: none"> <li>• geographicalName</li> </ul> </li> <li>– Sea</li> <li>– Wetland</li> </ul> </li> </ul>
<ul style="list-style-type: none"> <li>• <u>RelatedObjects:</u> <ul style="list-style-type: none"> <li>– DamOrWeir               <ul style="list-style-type: none"> <li>• condition</li> </ul> </li> <li>– SpringOrSeep</li> <li>– Embankment</li> <li>– ShorelineConstruction</li> <li>– AbstractPoint</li> <li>– DischargePoint</li> <li>– Pipe</li> <li>– Ford</li> <li>– SubsurfaceCrossing</li> <li>– SurfaceCrossing</li> <li>– VanishingPoint</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• <u>ManagementAndReporting:</u> <ul style="list-style-type: none"> <li>– WFDSurfaceWaterBody               <ul style="list-style-type: none"> <li>• id</li> </ul> </li> </ul> </li> </ul>

## 2.2. AVALIAÇÃO DE RESULTADOS

As principais conclusões do processo de estabelecimento de correspondências entre os modelos de dados dos produtores com o modelo de dados INSPIRE, para o tema hidrografia são as seguintes:

Foi possível encontrar correspondências para todas as entidades (menos uma) em dois CDG.

Foi possível encontrar correspondências para a maior parte das entidades noutros dois CDG.

Outros CDG têm um número mais restrito de correspondências, isto é:

Para os CDG da Regione Liguria, Regione Piemonte e Regione Veneto foi possível estabelecer correspondências com cinco entidades em cada um deles.

Nos CDG das regiões Italianas que seguem a especificação de dados DBPrior10K, nos dados da Comune di Genova, Magistrato alle Acque di Venezia e VUGK não há correspondências com entidades do pacote RelatedObjects.

Alguns pacotes quase não têm correspondências, isto é, o pacote ManagementAndReporting tem apenas uma correspondência e o pacote Network tem apenas duas correspondências.

A maior parte das correspondências envolve entidades de pacotes derivados do caso de utilização "Cartografia/Mapping", nomeadamente os pacotes PhysicalWaters e RelatedObjects.

Alguns produtores referiram a necessidade de inclusão de entidades específicas, por exemplo o Magistrato alle Acque di Venezia referiu a necessidade do atributo "speed limit" nos canais da lagoa.

Nenhum dos CDG analisados inclui a totalidade do catálogo de entidades do INSPIRE.

### 3. METODOLOGIA DE AGREGAÇÃO DE DADOS E SUA EXPOSIÇÃO A ESCALAS EUROPEIA, NACIONAL, REGIONAL E LOCAL

Por forma a ilustrar o desenvolvimento das regras e directrizes de agregação e visualização de dados nas escalas de nível Europeu, nacional, regional e local vamos usar o exemplo da divisão administrativa [1].

Os dados disponíveis são constituídos por nove CDG de níveis nacional (Eslováquia, Hungria e Portugal), regional (quatro regiões Italianas e a Catalunha) e local (Comune di Genova).

As contribuições dos produtores traduziram-se essencialmente em três tópicos que foram elaborados com base no documento [3]. O primeiro consistiu no levantamento e solução para os problemas que surgem nas fronteiras entre CDG que se tocam, conhecidos por "cross-border issues", o segundo tópico consiste na definição dos problemas de representação e visibilidade de entidades de acordo com a escala e, finalmente, o terceiro consiste nas necessidades de generalização, isto é na definição das entidades que necessitam de ser generalizadas para as 4 escalas de referência e aquelas que não são visíveis a determinadas escalas.

No caso dos CDG da divisão administrativa existiam quatro pares de CDG que se tocavam na fronteira, nomeadamente: Eslováquia/Hungria, Regione Veneto/Regione Friuli Venezia Giulia, Regione Liguria/Regione Piemonte, Regione Liguria/Comune di Génova. Para estes CDG foi efectuado o levantamento dos problemas que ocorrem e concluiu-se que soluções que não tenham em conta as especificidades e razões por detrás de cada um dos CDG não são em geral aceitáveis. Por exemplo, em geral não é possível usar a linha média para resolver o caso em que duas fronteiras não coincidem pois esta solução não é aceitável para nenhuma das partes. Como existe sempre uma boa razão para que a fronteira esteja delimitada de determinado modo a solução terá, regra geral, que ser negociada entre os dois produtores envolvidos.

O resultado do levantamento das inconsistências na fronteira entre CDG que se tocam revelou que as inconsistências podem ser topológicas, devidas a diferentes identificadores ou a nomes diferentes. As inconsistências topológicas detectadas foram sobretudo sobreposições e lacunas (overlaps, gaps) ou inconsistências devidas a diferentes escalas dos CDG que se tocam. Distinguiram-se as lacunas e sobreposições entre polígonos, entre curvas e entre curvas e polígonos.

De seguida definiu-se o conteúdo dos dados de acordo com a escala de visualização (Europeia, Nacional, Regional ou Local). Estas escalas foram definidas tema a tema e, no caso da divisão administrativa as escalas seleccionadas foram: Europeia - 1/40 000 000, Nacional - 1/10 000 000, Regional - 1/1 000 000 e Local - 1/200 000.

Para cada escala foram definidas as características de representação das entidades, por exemplo, limite cinzento, espessura de 1 pixel, sem preenchimento de polígono, etc.

As entidades de tipo poligonal são visualizadas ou transformadas em pontos de acordo com a sua área, assim, na escala europeia as entidades poligonais são visualizadas se a área for > 6400Km<sup>2</sup>, são transformadas em pontos se a área for menor que 6400 km<sup>2</sup> e se já forem representadas por pontos são eliminadas. A área de referência para decidir se uma entidade poligonal é visualizada ou transformada em ponto é, para o nível nacional 400km<sup>2</sup>, para o nível regional 4km<sup>2</sup> e para o nível local 0,16km<sup>2</sup>.

Scale	Visualized	Transform to point	Eliminate
European 1/40 000 000	If AREA > 6400 km <sup>2</sup>	If AREA ≤ 6400 km <sup>2</sup>	YES if it is a point
National 1/10 000 000	If AREA > 400 km <sup>2</sup>	If AREA ≤ 400 km <sup>2</sup>	YES if it is a point
Regional 1/1 000 000	If AREA > 4 km <sup>2</sup>	If AREA ≤ 4 km <sup>2</sup>	
Local 1/200 000	If AREA > 0.16 km <sup>2</sup>	If AREA ≤ 0.16 km <sup>2</sup> SIMPLIFY	

As entidades da divisão administrativa dos CDG remodelados são "Administrative Unit" e "Administrative Boundary". Para estas entidades definiram-se as necessidades de generalização para determinadas escalas, sobretudo simplificação (simplify) e adoçamento (smoothing).



Data provider	Dataset	Feature	Operation	Level
RPIE	CTR10.000/Census	Administrative Unit	SIMPLIFY	Regional, Local
		Administrative Boundary	SIMPLIFY	Regional, Local
RLIG	DBPrior10K	Administrative Unit	SIMPLIFY, SMOOTH	European, National
FÖMI	52-424_t_region	Administrative Unit	SIMPLIFY, SMOOTH	European, National, Regional, Local
VUGK	SK50 - Slovak Administrative boundaries	Administrative Unit	SIMPLIFY, SMOOTH	Regional, Local
		Administrative Boundary	SIMPLIFY, SMOOTH	Regional, Local
IGP	CAOP	Administrative Unit	SIMPLIFY	Regional, Local
ICC	BM-50M	Administrative Unit	SIMPLIFY, SMOOTH	European, National, Regional, Local
		Administrative Boundary	SIMPLIFY, SMOOTH	European, National, Regional, Local
RVEN	DBPrior10K	Administrative Unit	SIMPLIFY	Regional, Local
		Administrative Boundary	SIMPLIFY	Regional, Local
INSIEL	DBPrior10K	Administrative Unit	SIMPLIFY	Regional, Local
		Administrative Boundary	SIMPLIFY	Regional, Local

Foram também tratados os problemas devidos a diferentes línguas e as soluções consistiram na tradução de elementos do modelo de dados comum (catálogos de entidades e atributos) bem como dos atributos alfanuméricos e texto, dos elementos dos metadados e da interface humana do Geoportail do GIS4EU.

#### 4. CONCLUSÕES

As principais contribuições do projecto GIS4EU consistiram na criação de metodologia para transformar os dados dos produtores no modelo de dados INSPIRE e na subsequente participação na fase de testes das especificações de dados do anexo I da directiva INSPIRE, na criação de um modelo de dados comum para o tema altimetria, bem como na criação de metodologia para agregação de dados e sua exposição a escalas europeia, nacional, regional e local.

Dada a diversidade dos CDG disponíveis no projeto provenientes de produtores de diversos níveis o projecto GIS4EU constituiu um cenário real da situação que a implementação da directiva INSPIRE enfrentará quando as directrizes de execução forem obrigatórias. Por outro lado a metodologia usada no projecto e a experiência adquirida constituem boas bases de trabalho para a adopção das directrizes INSPIRE pelos produtores de dados.

#### 5. REFERÊNCIAS

- [1] Biotto, G., De Zorzi, S., De Zorzi, S., Reis, R. (eds.): Guidelines for Data Aggregation: Administrative Units Theme. GIS4EU project report D 5.1 (2009)
- [2] Hobona, G., Jackson, M.(eds.) Data descriptions, GIS4EU project report D 2.2 (2008)
- [3] Hobona, G., Jackson, M.(eds.): Merging Rules - Degradation Rules, GIS4EU project report D 3.6 (2009)

- [4] Reis, R.(ed.): Common Data Model: Hydrography. GIS4EU project report D 3.3 (2009)
- [5] Reis, R., Barrot, D.: Testing Cartographic Data Harmonization in the INSPIRE Context: The Case for Hydrography. In: Actas da VI Conferência Nacional de Cartografia e Geodesia, Lidel, Lisboa, pp.452--461 (2010)

## 6. AUTORES

Rui REIS  
Rui.reis@igeo.pt  
IGP  
DSIGIG