Data quality, provenance and IPR Management services: their role in empowering geospatial data suppliers and users

Keiran Millard and Jennifer Herbert (HR Wallingford)

Overview

This poster looks at current experiences of geospatial users and geospatial suppliers and how they have been limited by suitable frameworks for managing and communicating data quality, data provenance and intellectual property rights (IPR). Current political and technological drivers mean that increasing volumes of geospatial data are available through a plethora of different products and services, and whilst this is inherently a good thing it does create a new generation of challenges.

These challenges are that it has become more challenging and time consuming for users to appraise the value of data for their particular purpose. The second is that data suppliers are finding that traditional data licensing models are difficult to apply in a culture that is dominated by public 'open data'.

This paper consider two examples of where these issues have been examined and looks at the challenges and possible solutions from a data user and data supplier perspective. The first example is the EC FP7 IQmulus project that is researching fusion environments for big geospatial point clouds and coverages. The second example is the EU Emodnet programme that is establishing thematic data portals for public marine and coastal data.

Challenges

User Perspective #1: Data Value

Users need to appraise the suitability and value of a data set to solve their particular problem. Traditional formal metadata does not offer a solution; it supports data appraisal in providing information about the data, but the user needs to be able to appraise this information to reach a decision on the data's value. Not all users are in position to do this and also the information contained in metadata alone does not fully support value assessments. Often this can only be obtained from actually using the data.

User Perspective #2: Cost Benefit

Open Data policy provides more data choice, but in itself does not support the cost-benefit assessment of user. Users, especially business and commercial users, have many different value propositions other than the data being 'free' – for example based on factors such as data quality, speed of access and temporal and spatial coverage. This needs to be explicit to users so they can appraise the cost:benefit of 'data A' over 'data B'

Supplier Perspective #1: Value-Add Business

Open Data is meant to empower innovation and growth, but to do that revenue needs to be generated from the data. This can sometimes be at odds with a market where the expectation is that data is free at point of use. Hence the value-add provided needs to be explicit.

Supplier Perspective #2: Open and Non-Open data mix

Value-add organisations need mechanisms to combine open and non-open data to create innovative products to meet user needs. This leads to a mixed licence product and the value may not be apparent to uses. There needs to be approaches where the open:non-open data mix can varied to tune products to different user markets. This leads to big data challenges as many of the source datasets are very large and custom product generation is computationally demanding

Contact

Keiran Millard	k.millard@hrwallingford.com	
Jennifer Herbert	j.herbert@hrwallingford.com	
HR Wallingford	www.hrwallingford.com	@hrwallingford
SeaZone	www.seazone.com	@seazonedotcom
EMODNET	http://www.emodnet.eu/northsea	
IQmulus	http://iqmulus.eu/	

Emodnet: Attributing value to open Geospatial Data

The Emodnet programme delivers harmonised marine data at the EU scale across seven thematic portals. As part of the Emodnet programme a series of 'check points' have been initiated to examine how useful these services and other public data services actually are to solve real-world problems. One key finding is that users have been confused by the fact that often data from the same source appears across multiple platforms and that current 19115-style metadata catalogues do not help the vast majority of users in making data selections.

To address this, we have looked at approaches used in the leisure industry. This industry has established tools to support users selecting the best hotel for their needs from the metadata available, supported by peer to peer rating. We have looked into how this approach can support users in selecting the best data to meet their needs.

Emodnet Data Value Framework

The data value framework for the Emodnet North Sea Checkpoint considers three components: The data itself, a challenge to be solved, and a structure assessment of the value of the data to solve the challenge (Figure 1).

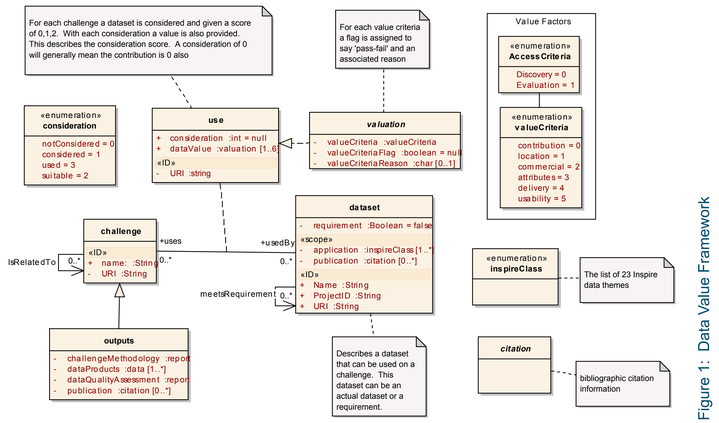


Figure 1: Data Value Framework

Any dataset has three consideration levels for a challenge ('notConsidered', 'considered' and 'suitable'). A 'valuation' is associated with a consideration instance to both qualify and quantify the consideration. The valuation is expressed in terms of an enumerated list of value criteria. There is a pass/fail flag associated with each criteria and an optional narrative reason for this flag. So for example, a data set may be 'consideration::notConsidered' for a challenge because valueCriteria::contribution has valueCriteriaFlag:="False" because valueCriteriaReason::"buildings dataset does not contain offshore structures". A data set used on a challenge would have valueCriteriaFlag:="True" for all valueCriteria.

This is illustrated further in Tables 1 to 3 opposite. This considers three data sets for use on the wind farm siting challenge where a high resolution bathymetry is needed. In the first case the data set is not considered as it contains no useful information. In the second case the data is considered for a challenge, but it is not used as the data does not meet the requirements of the challenge. In the last case the data does meet the requirements of the challenge.

The result is that the user is present with effective 'star rating' of the data for their purpose.

T1 - Example use record for a data set "not considered" for a challenge

Use	URI (links dataset URI to a challenge URI)	
Consideration	notConsidered	
ValueCriteria	**ValueCriteriaFlag**	**ValueCriteriaReason**
Contribution	F	Dataset contains no information that can be used on the challenge
Location	Null	
Commercial	Nul	
Attributes	Null	
Delivery	null	
Usability	null	

T2 - Example use record for a data set "considered" (but not suitable) for a challenge

Use	URI (links dataset URI to a challenge URI)	
Consideration	Considered	
ValueCriteria	**ValueCriteriaFlag**	**ValueCriteriaReason**
Contribution	T	Dataset contains gridded bathymetry which is required to develop a model of the seafloor for scour assessment
Location	F	Gridded bathymetry outside of study area
Commercial	T	Free to download. Open Data Licence
Attributes	F	Grid resolution of 10arcSec too coarse for modelling
Delivery	T	Web download
Usability	T	Raster Grid Format

T3 - Example use record for a data set "suitable" for a challenge

Use	URI (links dataset URI to a challenge URI)	
Consideration	Suitable	
ValueCriteria	**ValueCriteriaFlag**	**ValueCriteriaReason**
Contribution	T	Dataset contains gridded bathymetry which is required to develop a model of the seafloor for scour assessment
Location	T	Gridded bathymetry within study area
Commercial	T	Licence fee €35.
Attributes	T	Grid resolution of 1arcSec. Better than needed
Delivery	T	Web download
Usability	T	Raster grid format

IQmulus: Provenance and IPR in Big Geospatial Data

IQmulus examines big geospatial data; the data from sources such as LIDAR, SONAR and numerical simulations; these data are simply too big for routine and ad-hoc analysis, yet they could realise a myriad of disparate, and readily useable, information products with the right infrastructure in place. IQmulus is researching how to deliver this infrastructure technically, but a financially sustainable delivery depends on being able to track and manage ownership and IPR across the numerous data sets being processed. This becomes complex when the data is composed of multiple overlapping coverages, however managing this allows for uses to be delivered highly-bespoke products to meet their budget and technical needs.

The Merging licence terms and Associated Value.

In this example the user would like a DEM for a particular area and seabed. The source data that can be used to create this DEM is several TB in size and are also supplied under different licence terms. There is a technical challenge in merging the data to create the right product for the user but merging licensing and re-use terms can be more complex than technically merging data. Very often the most onerous licence terms predominate so the user needs to understand the benefits of additional data with different licence terms.

In IQmulus we have considered how metadata can be used to inform the value decision about which data to merge. In this case we appraise whether the accumulative addition of data make a substantive difference to the values contained in the DEM. For example if it were known that adding some available commercial data to an existing open data DEM gave a 12% variance in depth then the user may decide to include this data in their product. If it gave a 1% difference they may elect to remain with the open data. This approach not only supports users in data selection, it also supports data suppliers by making explicit the value of the data they hold.

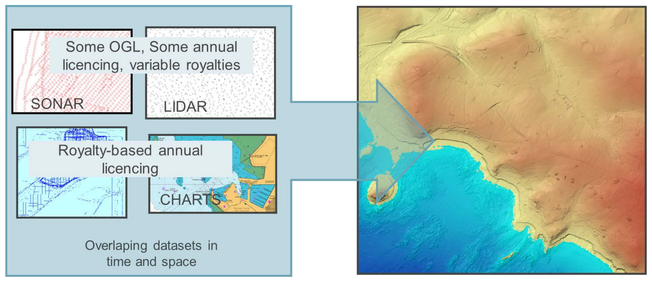
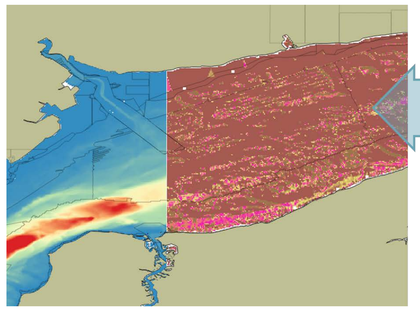


Figure 2: Merging different data sets and data terms to create a DEM



Looking at bathymetry two ways. On the left is a traditional colour ramp related to depth. colour ramp for the section on the right is related to variability in source data.

Figure 3: The effect of merging different data to create the DEM can be made explicit to users enabling them to make value assessments about which data to merge

Acknowledgements

IQmulus is a 4-year Integrating Project (IP) partially funded by the European Commission under the Grant Agreement FP7-ICT-2011-318787. It is positioned in the area of Intelligent Information Management within the ICT 2011.4.4 Challenge 4: Technologies for Digital Content and Languages. IQmulus started on November 1, 2012, and will finish on October 31, 2016.