An Information Model for a Water Information Platform

Pascal Dihé, Ralf Denzer, and Sascha Schlobinski

Cismet GmbH, Altenkesseler Straße, 17, 66117 Saarbrücken, Germany
Ralf.denzer@enviromatics.org

Abstract. Sharing of open government data is amongst other reasons hindered by incompatibility of data models in different data collections. Only a few areas in the environmental domain have progressed towards commonly used data models. The purpose of this paper is to share with the community a data model which is used in a spatial information platform being built for the purpose of sharing open government data in the domain of water sciences. The objective when building the information model was not to be restricted to one and only one (meta)data standard. The information model therefore uses several standards and extension mechanisms: the ISO19000 series, the Comprehensive Knowledge Archive Network (CKAN), dynamic tag extension and dynamic content extension. The CKAN domain model can also be mapped to semantic-web-compatible standards like Dublin Core and the Data Catalogue Vocabulary of the World Wide Web Consortium.

Keywords: water information platform, water data model, spatial data infrastructure, spatial information platform.

1 Introduction and Related Work

The re-use of publicly funded governmental data has received a lot of attention recently. In the environmental domain, it is clear that improved public services need exchange of data across governments at all levels. Governments keep producing information products at all levels, and some of them are more or less readily available. Reporting obligations in the EC, for instance demanded by the Water Framework Directive (WFD)[1], are direct inputs to European datasets.

Water-related information is progressively made available on-line by a large variety of actors, including water data from operational monitoring and reporting of authorities. In Europe, the EEA plays an important role by providing water information through WISE and the water data centre[2]. Data does not only include monitored or reported data but also fundamental information like basin networks [1]. Over the past years, the Australian government has made large efforts in building water information platforms, information systems and tools [2]. Similarly, in the United States, CUASHI [3] acts as an alliance to

[1] http://ec.europa.eu/environment/water/water-framework/index_en.html

[2] http://www.eea.europa.eu/themes/water/dc

R. Denzer et al. (Eds.): ISESS 2015, IFIP AICT 448, pp. 91–101, 2015.

improve water information and associated ICT. At international level the GEOSS data core also lists many water-related information resources [4].

However, *on-line* means different concepts to different people, from just presenting data to full service integration, with or without registration and authentication [5], and *re-use of data* is not happening at large scale, and platforms, particularly those developed in R&D projects often do not survive long after the project. Recent discussions in the community [6] have identified a fundamental gap between concepts, research and implementation. This gap is given by a) a distance between the modelling community / modelling environments and infrastructure developers / providers, b) a lack of tools supporting the uptake of infrastructures being built and a lack of understanding, how "re-purposing" (re-use under different context) of data can be supported.

As part of the SWITCH-ON project, a Spatial Information Platform (SIP) is being developed, which supports the reuse of information products for the water science domain. This platform is used by water scientists and aims at integrating water related data from various sources including open government data. The re-use (re-purposing) of data is a central element of the platform. The concept of this SIP has been published recently [5] and a first version is on-line[3].

The core platform is implemented with a stack of freely available open source software compiled in the CIDS product suite of CISMET GmbH. This software suite consists of a set of software components, application programming interfaces (APIs), management and development tools, services and applications, with a special focus on interactive solutions which need to integrate geo-spatial systems with databases, sensor networks, document-oriented systems, unstructured information sources and numerical models. CIDS is particularly suited for solutions which have to be built across existing heterogeneous information systems, which may be under control of different organizations. CIDS has been used in numerous projects since 1999. For recent projects see [7,8].

2 Platform Architecture

The architecture of the SIP is separated horizontally into three disparate layers: a *GUI Layer* (graphical user interface), a *Service Layer* and a *Data Layer* (figure 1). Each layer contains several components or sets of components. Interaction between components happens between components of adjacent layers. Apart from a few exceptions, GUI components, for example, do not communicate directly with the data repositories. These three layers are accompanied by a vertical *Tools Layer*.

The *Data Layer* is concerned with the storage and management of data and meta-data (e.g. catalogue data). The most important component of the Data Layer is the Meta-Data Repository, a software component which is responsible for the storage and management of the meta-data compliant with the information model described in this article.

[3] `http://www.water-switch-on.eu/sip.html`

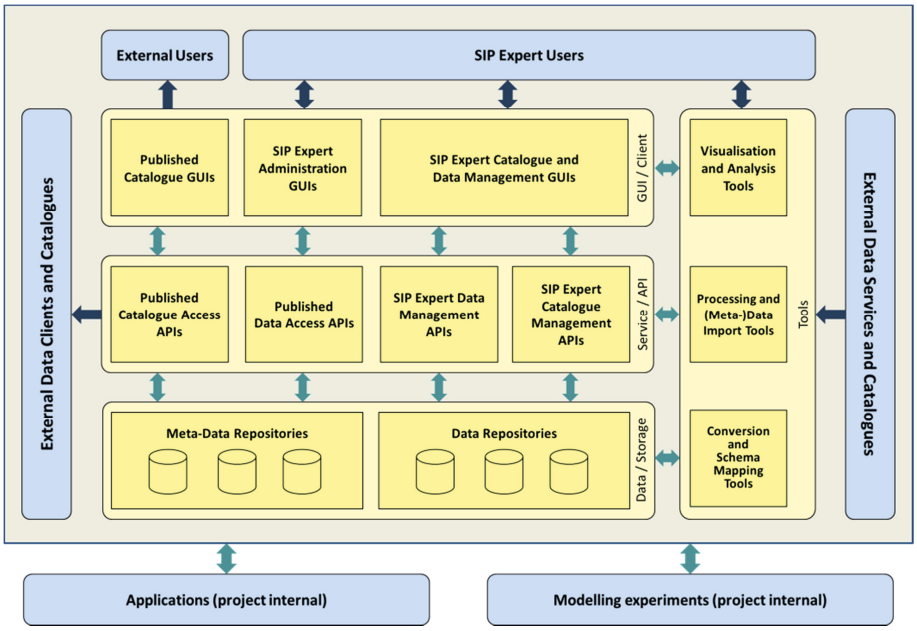


Fig. 1. SIP architecture

The *Service Layer* contains components that offer public service interfaces (APIs) which can be used by public and expert client components of the SIP. Those services also implement most of the (server-side) business logic of the SIP. They are supported by tools which provide additional functionality (e.g. data conversion) and which can be directly embedded in the respective service implementation.

The *GUI layer* of the SIP architecture contains user interface and user interaction components for expert and external users. External users include the general public which is able to use the public GUIs of the SIP without prior registration. Expert users have more access rights than external users, e.g. for the manipulation of data and meta-data.

The Tools Layer contains several supporting components that provide common or specialized functionality which are used by many different components of any other layer. Some tools are implemented as services and can be called from within other services and GUIs, while other tools represent algorithms or processes that can be embedded in services or (expert) GUIs. Some tools provide their own user interfaces.

3 Considerations about Standards and Information Models

The main purpose of the SIP information model is to support an extension of the common *publish-find-bind* pattern towards re-purposing of data (re-using data in a different context, see [6]). This extended pattern is *publish, find, bind, transform* (figure 2). As users re-use data, they not only include data in their analytics, they often also transform data into a new context or frame of reference.

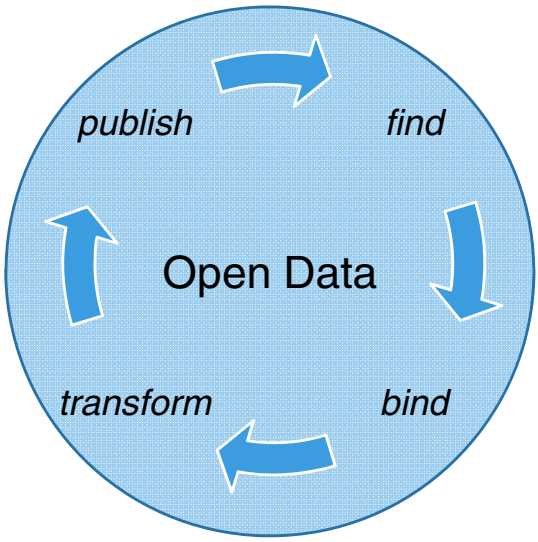


Fig. 2. Extended pattern for re-purposing of data

Although widely adopted information models for the description of data and services exist (e.g. ISO19115 [9] and ISO 19119 [10]) the decision was taken that the information model for the water information platform should not be based on one of these standards alone.

This is especially true for the requirement to consume and possibly to feed-back open data from existing catalogues. This leads to the imperative to support different (meta)data formats and standards which demand a flexible information modelling approach. Accordingly, the commitment to one (meta)data standard or profile may hinder external providers of open data that are compliant with the Standard Information Model to publish their open data to the SIP. Furthermore, support for the documentation of scientific analyses, the description of software (tools), aspects related to re-repurposing open data, inherent features for cataloguing, and other topics that have to be considered at the level of the information model, are only partially covered by existing (meta)data standards.

For instance, the commitment to one and only one (meta)data standard or profile could hinder external providers of open data to publish their open data or results of their studies to the SIP.

The design aim for the SIP was to support different (meta)data formats and standards in order to provide a flexible framework. This demands a flexible information modelling approach. Thus, instead of defining one fixed information model that is based on a selection of particular meta(data) standards or profiles, the information model for the SIP can be tailored to actual needs.

An interesting approach which goes in the same direction has been adopted by the federal German Open Data portal (GovData.de) [11]. GovData's information model is based on a meta-data structure (domain model) developed by the Open Knowledge Foundation for the Comprehensive Knowledge Archive Network (CKAN) [12].

CKAN is a de-facto standard for catalogues of Open Government Data (OGD). The CKAN domain model distinguishes between resources (e.g. data, services and tools) and their actual ("physical") representation (e.g. database, file, service endpoint, etc.). It defines a fixed set of mandatory or optional attributes to describe resources and their representations (title, description, license, contact, etc.) and allows the extension of the model by introducing arbitrary "extra attributes". GovData has made extensive use of this extension mechanism and created its own CKAN-based OGD meta-data schema [13].

Furthermore, the CKAN domain model can be mapped to semantic-web-compatible standards like Dublin Core [14] and the Data Catalogue Vocabulary (DCAT) of the W3C [15]. While at first glance CKAN seems like the ideal the candidate for the information model for the water information platform, one must consider that catalogue functionality, essentially providing access to a collection of datasets, is just one small part of the SIP. Furthermore, the SIP also has to interface with other types of catalogues like OGC Web Catalogue Service (CS-W) [16] to ensure that resources managed by the SIP can also be exposed in CS-W, among others. Nevertheless, the concepts of the CKAN domain model as well as support for meta(data) standards like Dublin Core, ISO 19115, etc., were considered in the design of the information model for the water information platform.

4 Platform Information Model

The design of the SIP information model follows a graduated approach with three different levels of increasing extensibility and flexibility: a *relational model*, *dynamic tag extensions* and *dynamic content extensions*.

The first layer is a relational information model (figure 3) which simultaneously defines the outline of the two subordinate layers. It is implemented as object relational database model of the CIDS platform and supports the core business processes of the SIP. Besides basic classes for resources, their relationships and representations, the model uses several categories from the ISO 19115 meta-data standard to define a set attributes needed to describe those classes. The relational model is also the basis for both the internal and external catalogues of the SIP. The most important classes of the relational model are

- *Resource*
  Resources are the central entity of the information model and the catalogue. They logically describe a dataset, a service, a tool, etc. Resources have a set of core attributes that have been derived from the basic ISO 19915 metadata categories. Furthermore, arbitrary additional metadata can be associated with a resource.
- *Representation*
  Among other roles, the representation defines how to actually access a particular instance of a logical resource. A resource may have different representations. For example, a dataset may be available in different spatial or temporal resolutions or different formats.

- *Relationship*
  This class is used to specify relationships (e.g. "derived from") between resources, e.g. to track and document data transformations and processing (lineage) of resources. The type of the relationship can be identified by a respective tag and arbitrary additional meta-data can be associated with a relationship.

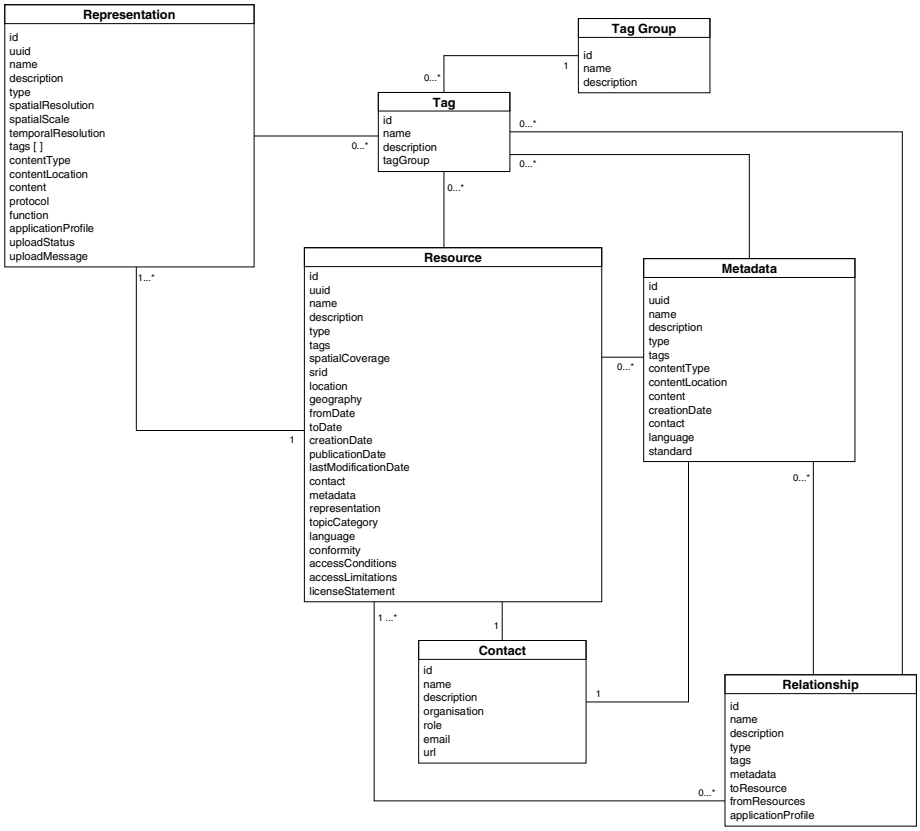


Fig. 3. Overview of relational model

- *Metadata*
  The metadata class represents additional meta-data about a resource or a relationship in a structured or semi-structured format, for example meta-data on data quality.
- *Tag* and *Taggroup*
  Tag and Taggroup represent the dynamic tag extension layer of the Standard Information Model. Taggroups define a general classification for tags and thus can be used to create lists of predefined tags (e.g. code lists).

Dynamic tag extensions provide the possibility to extend the information model for the water information platform without causing changes to the relational model itself. Accordingly, new information-needs of SIP client applications (external catalogues,

tools, etc.) can quickly be fulfilled without the need to change the internal database structure of the Meta-Data Repository. Besides the possibility for extending the relational model by introducing new tag groups, tags and tag groups are mainly used to define fixed value lists like standardized topic categories, INSPIRE compliant keywords lists and so on.

Dynamic content extension is a simple mechanism for further extending the information model by either dynamically injecting arbitrary content encoded in plain text into the model or by providing references (URIs) to externally stored content.

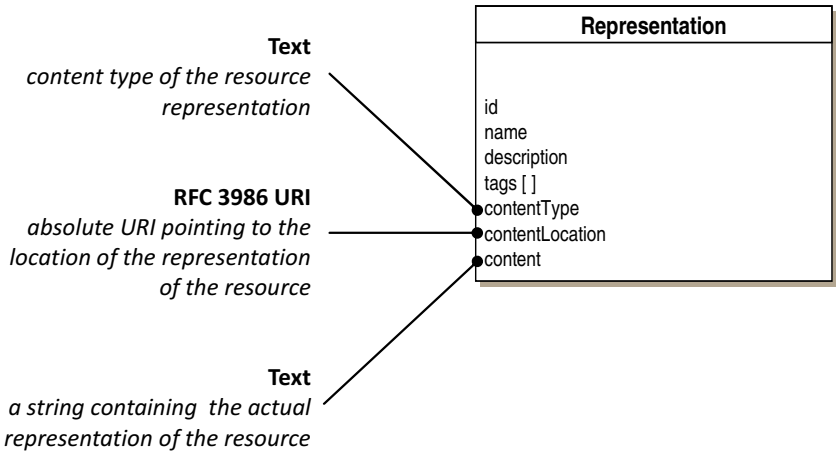


Fig. 4. Dynamic Content Extension

In the relational model, dynamic content extensions are represented by triples of attributes "contentType", "contentLocation" and "content". The "content" attribute contains the actual content encoded in plain text, for example a JSON or XML document. Although this type of content is not directly part of the relational data base model, it can be used to store structured or semi structured information that can be processed by content-aware tools or clients. The attribute "contentLocation" defines the location of the content when the content is not dynamically injected but is referenced. The type of this attribute is a URI (RFC 3986).

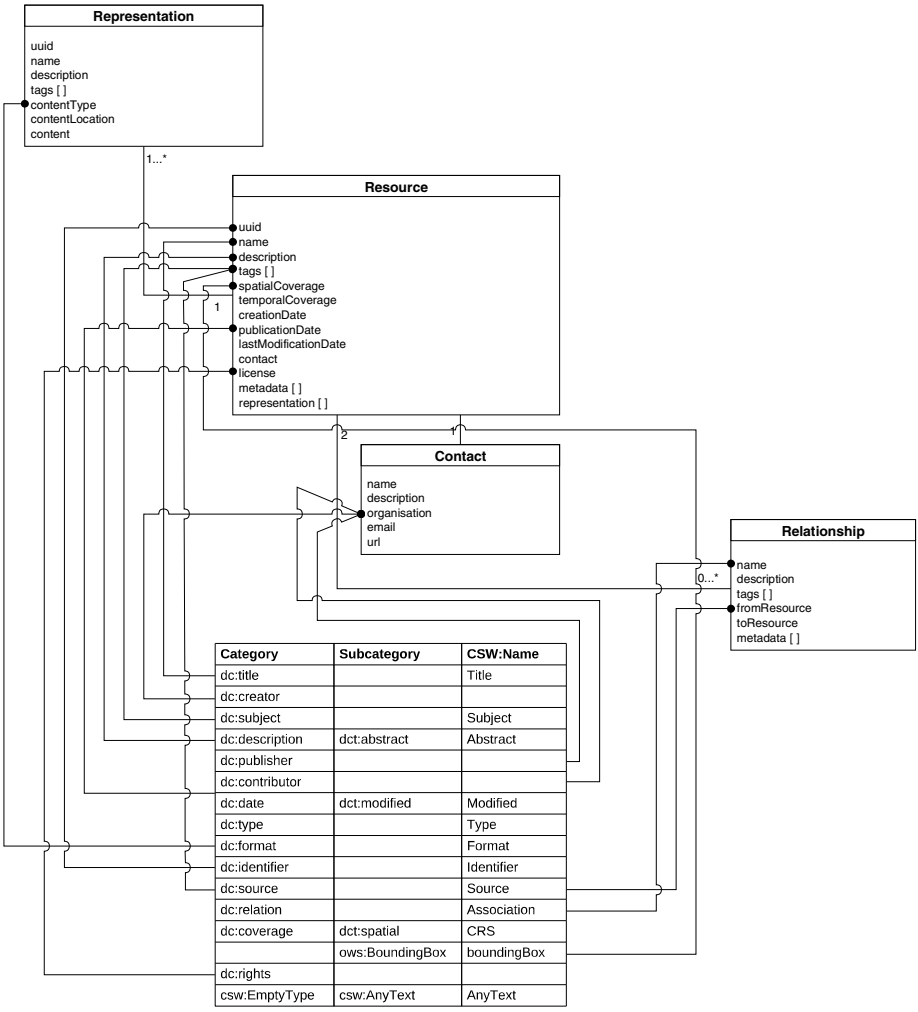
The "contentType" attribute refers to standardised Internet Assigned Numbers Authority (IANA) media types (image/png, text/plain,...) as well as to custom industry standard media types (application/x-netcdf, application/gml+xml,...).

While the content type is in general sufficient to identify the data stored in the content field of the representation, a URL that is stored in the contentLocation field may however not directly point to actual data. Instead, the link may lead to an online form for accessing the data (in different formats) or a service endpoint. Therefore additional information about the handling, access and processing of content must be provided.

For this purpose, dynamic tag extensions can be used to introduce new tag groups like "function" and "protocol". "Function" defines for example the function that can

be performed when following the “contentLocation” link to the resource representation. Examples of such functions may include “download” which enables the user to directly download the resource representation or “order” where the link value is an URL of a web application that requires user interaction to order/request access to the resource representation.

Classes of the relational model with support for dynamic content extension, and thus containing the three aforementioned attributes, are “Representation” and “Metadata”. In the case “Representation”, content is generally provided by reference. Thus the “contentLocation” attribute contains a URI that points to actual resource data.



Category	Subcategory	CSW:Name
dc:title		Title
dc:creator		
dc:subject		Subject
dc:description	dct:abstract	Abstract
dc:publisher		
dc:contributor		
dc:date	dct:modified	Modified
dc:type		Type
dc:format		Format
dc:identifier		Identifier
dc:source		Source
dc:relation		Association
dc:coverage	dct:spatial	CRS
	ows:BoundingBox	boundingBox
dc:rights		
csw:EmptyType	csw:AnyText	AnyText

Fig. 5. Mapping to Dublin Core / CSW

5 Mapping to Related Standards

The SIP also has to interface with other types of catalogues like OGC Web Catalogue Service (CSW) or Open Knowledge Foundation for the Comprehensive Knowledge Archive Network (CKAN) catalogues. For this purpose, Published Catalogue Access APIs have been introduced which provide public and standards based access to meta-data stored in the SIP meta data repositories.

Because OGC CSW is one of the most commonly used data catalogues, the open source CSW implementation pyCSW (which is the reference implementation of CSW) has been selected as one realization of a Published Catalogue Access API. Since pyCSW retrieves its complete data from a relational database model, a mapping of the relational model of the SIP to the Dublin Core encoding of CSW Core Metada-ta schema could also been defined.

This mapping is shown in figure 4. The tables represent an excerpt of the Dublin Core meta-data profile as adopted by the CSW standard. The boxes represent the respective tables of the data model. The connections represent the mapping.

6 Future Work

Since provenance of re-used data and modelling experiments is important, it should be considered also in the relational model of the SIP. Up to now provenance informa-tion has not yet been considered systematically in the project. The "Relationship" class can be used to describe the relationship (lineage) between resources and Dynam-ic content Extensions can be used to represent the actual lineage meta-data.

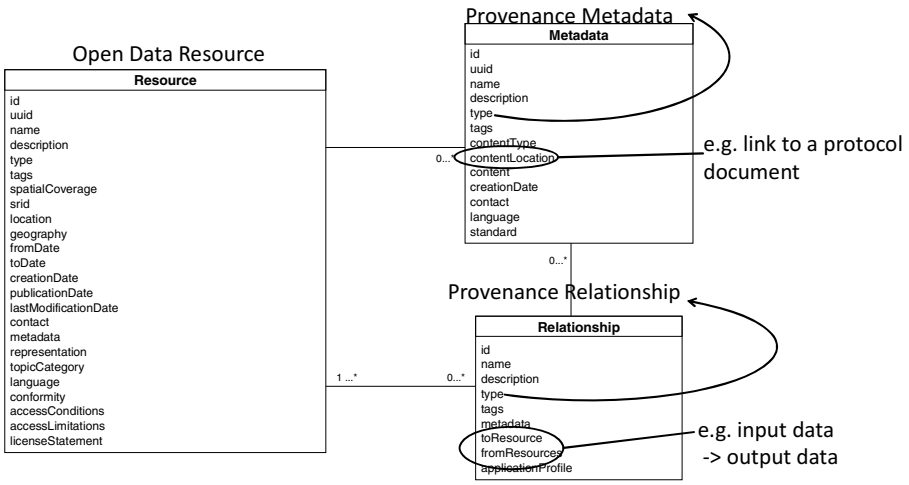


Fig. 6. Potential provenance extension

This simple relationship (fromResource / toResource) can for example be used to describe the I/O of an experiment (model run), a script for repurposing, etc. The type of the relationship can be identified by a respective tag. Interestingly, arbitrary additional meta-data can be associated with a relationship. This is where the actual lineage meta-data, e.g. a link to the protocols of an experiment, or a description could be stored. These considerations are part of future work.

Acknowledgements. The SWITCH-ON project is funded under the European Framework Program FP7 (contract number 603587).

References

1. ECRINS EEA, EEA Catchments and Rivers Network System, ECRINS v1.1, ISSN 1725-2237 (2012), http://www.eea.europa.eu/publications/eea-catchments-and-rivers-network
2. BOM. Austrian Government, Bureau of Meteorology, Improving Water Information Programme, Progress Report, Advances in water information made by the Bureau of Meteorology in 2013 (2013), http://www.bom.gov.au/water/about/publications/document/progress_report2013.pdf
3. CUASHI
4. GEOSS. Group on Earth Observation, GEO-IX, 22-23 November 2012, Report of Data Sharing Working Group, Document 13 (2013), https://www.earthobservations.org
5. Denzer, R., Schlobinski, S., Boot, G., Keppel, F., De Rooij, E.: An information platform fostering re-use of water data. In: International Congress on Environmental Modelling and Software (2014), http://www.iemss.org/sites/iemss2014/proceedings.php ISBN: 978-88-9035-744-2
6. Denzer, R.: Hydroinformatics: Interoperability, standards and governance of water information infrastructures. In: Proceedings of the WIRADA Science Symposium, pp. 120–124 (2012), http://www.csiro.au/WIRADA-Science-Sympsosium-Proceedings
7. Denzer, R., Schlobinski, S., Gidhagen, L., Hell, T.: How to Build Integrated Climate Change Enabled EDSS. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli, A.E., et al. (eds.) ISESS 2013. IFIP AICT, vol. 413, pp. 464–471. Springer, Heidelberg (2013)
8. Hell, T., Kohlhas, E., Schlobinski, S., Denzer, R., Güttler, R.: An information system supporting WFD reporting. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli, A.E. (eds.) ISESS 2013. IFIP AICT, vol. 413, pp. 403–413. Springer, Heidelberg (2013)
9. ISO 19115 (2002), ISO/TC 211 Geographic information/Geomatics. ISO reference number: 19115 (2002)
10. ISO 19119 (2007), ISO/TC 211 Geographic information/Geomatics. ISO reference number: 19119 (2007)
11. Marienfeld, F., Schieferdecker, I., Lapi, E., Tcholtchev, N.: Metadata aggregation at GovData.de - An experience report, Association for Computing Machinery. In: ACM 9th International Symposium on Open Collaboration 2013, Proceedings, Hong Kong, China, August 5-7, pp. 638-6. WikiSym + Opensym (2013)

12. CKAN, Comprehensive Knowledge Archive Network, CKAN Domain Model, http://docs.ckan.org/en/ckan-1.8/domain-model.html
13. OGD-METADATA, Fraunhofer FOKUS, Schema and documentation to be used by the German Open Data Portal, https://github.com/fraunhoferfokus/ogd-metadata
14. Dublin Core Metadata Initiative, http://dublincore.org/
15. Maali, F., Erickson, J., Archer, P.: Data Catalog Vocabulary (DCAT), World Wide Web Consortium (W3C), W3C Recommendation 16 January 2014, http://www.w3.org/TR/vocab-dcat/
16. Nebert, D., Whiteside, A., Vretanos, P. (eds.): OpenGIS® Catalog Services Specification, Version 2.0.2, OGC 07-006r1, Open GIS Consortium Inc. 218 p. (2007)