Open Geospatial Data, Software and Standards

ORIGINAL ARTICLE Open Access

The conceptual schema in geospatial data standard design with application to GroundWaterML2

Boyan Brodaric[1*], Eric Boisvert[1], Peter Dahlhaus[2], Sylvain Grellet[3], Alexander Kmoch[4], François Létourneau[1], Jessica Lucido[5], Bruce Simons[6] and Bernhard Wagner[7]

Abstract

The explosive growth of geospatial data has stimulated the development of many standards aimed at decreasing data heterogeneity and enhancing data use. Well-established design methods for geospatial data standards typically involve the creation of two schemas for data structure, designated here as logical and physical, but this can lead to conceptual inconsistencies and modelling inefficiencies. In this paper we describe a design method that overcomes these issues by incorporating an additional schema – the conceptual schema – and demonstrate its application to the design of GroundWaterML2 (GWML2), a new international standard for groundwater data. Results include not only a new data standard, robustly constructed and tested, but also an enhanced method for geospatial data standard design.

Keywords: Geospatial data standards, Conceptual schema, Groundwater, GroundWaterML2, GWML2

Introduction

The explosive growth of geospatial data in the past two decades has stimulated many international standards efforts. These aim to improve activities involving data by reducing heterogeneity across the data lifecycle, including its creation, discovery, access, use, and archival. Various institutional bodies are responsible for the development and maintenance of such standards, most notably the Open Geospatial Consortium (OGC), the International Standards Organization (ISO), and the World Wide Web Consortium (W3C). From an initial focus on domain-neutral standards, such as those for encoding geometry data and viewing maps [8, 18], community interest within these bodies has expanded to domain-specific data standards, such as those for groundwater (GWML2), geology (GeoSciML), or hydrology (WaterML2) [4, 9, 39].

Such standards incorporate schemas (i.e. data structure templates) to organize various aspects of the data, but there exist many kinds of schemas at different levels of abstraction and for different purposes. Well-established data modelling principles suggest the deployment of three schemas that vary by technological neutrality: physical, logical, and conceptual [27, 37]. Conceptual schemas are technologically neutral and describe the world as it is; logical schemas describe the world from the viewpoint of a specific technological paradigm, such as logical, functional, or object-oriented; and physical schemas are aimed at specific system implementations, such as for a particular database system or data transfer language. Significantly, these schemas are interdependent, as the physical depends on the logical, which depends on the conceptual. A dependent schema then implements all the schemas on which it depends, even if those schemas are not explicitly and separately expressed – i.e. every logical schema reflects some conceptual schema, even if the conceptual schema is not stated; likewise every physical schema reflects some (possibly unstated) logical schema, and thus also some (possibly unstated) conceptual schema.

For the purpose of creating standards, such schema interdependency then implies the explicit development of all schemas on which a specific schema depends. However, OGC and ISO guidelines for geospatial data standards [22, 23] capture this rule only partially, as explicit specification of a logical schema is optional for physical schema design, e.g. for XML encoding, and development of a conceptual schema as understood in data modelling is

* Correspondence: boyan.brodaric@canada.ca
[1]Geological Survey of Canada, 601 Booth St. Ottawa, Ottawa, ON K1A0E8, Canada
Full list of author information is available at the end of the article

not required. Indeed, the conceptual schema is only partially recognized by the major geospatial standards design methods and it is typically conflated with the logical schema: what in essence has many aspects of a logical schema is usually referred to as a conceptual schema by the vast majority of OGC and ISO domain standards. Other standards efforts that build upon OGC/ISO, such as the European INSPIRE initiative [14], mirror this practice. In general, the tri-category segregation of schemas is not a formal part of major geospatial standard design methods, which advocate the development of only one or two schemas that do not fully correspond to a conceptual schema.

The problem with the omission of a conceptual schema is the lack of a conceptual baseline. This causes logical schema designs to contend with both technological and conceptual issues, increasing modelling difficulty and risking poorer results. Consider, for example, a scenario requiring two related logical schemas, a heavyweight version intended for complex data transfer between databases, and a lightweight version intended for simple data visualization on mobile phones. It is unlikely that either could serve as a conceptual baseline given their different technological orientations, hence it would be difficult to ensure alignment between them. The data modelling would then also likely be inefficient due to duplicate efforts in representing the conceptual overlap, which can lead to conceptual inconsistencies, and it would likely be less effective, due to the need to reconcile the conceptual overlap with the technological disparity. However, these shortcomings can be overcome if the logical schemas conform to a distinct conceptual schema, thereby neatly separating technological and conceptual concerns and ensuring conceptual alignment. This would lead to a shared understanding during the modelling process as well as a clear demarcation of what is to be modelled and any associated limits. An enhanced method for domain standards development should therefore include a conceptual schema and describe its interrelation with all other schemas.

In this paper we outline such a method and demonstrate its application to a new groundwater data standard, GroundWaterML2 (GWML2). GWML2 is developed by the OGC's Groundwater Standard Working Group operating under the Hydro Domain Working Group. As this paper focusses on method development for geospatial data standards design, the GWML2 standard is described only minimally (for a full description see [4, 5]), and standards for data access, e.g. for web services, are out of scope. Also for the purposes of this paper, the following terms and meanings are adopted: types are generalizations that encompass notions such as classes, categories, universals and kinds, and relations are links between entities. The instantiation relation holds between a type and another entity, such as between the type Aquifer and its instance Milk River Aquifer; then an instance is something that instantiates a type and a type is instantiated by an instance. The specialization relation holds between types, such that the specialized type has a narrower meaning than the subsuming type; for example, Aquifer specializes HydroGeologicalUnit because an aquifer is a special kind of a hydrogeological rock body. Properties refer to relations that are neither instantiation nor specialization, and encompass links to internal entities such as qualities, parts or constituents, as well as links to external entities such as the containment relation between an aquifer and the fluid body that it holds.

The paper is structured as follows: Section 2 of this paper describes related work, Section 3 outlines the general method, Section 4 describes its application to the design of GWML2, and Section 5 concludes with a brief summary and a discussion of future directions.

Related work

Fully articulated design methods for geospatial data standards are rare, with the most prominent belonging to OGC/ISO and INSPIRE [14, 23]. The OGC/ISO method includes development of an informal or semi-formal description of the domain, establishment of a schema, and integration with other schemas. The INSPIRE method refines these steps to include specification of use-cases, requirements, a data analysis, a logical or physical schema, and testing as well as evaluation guidelines. In each, application schemas that are domain-specific (e.g. for geology) are distinguished from standard schemas that are relatively domain-neutral (e.g. for geometry), with the key difference being that standard schemas are applicable to many domains. Both kinds of schema can be expressed in a conceptual modelling language (CML) such as UML or OWL [22, 24], or a data language (DL) such as GML-XML [32]; additional rules are also established for conversion amongst them [24, 32]. A schema following OGC/ISO protocols and expressed in a CML is referred to as a conceptual schema by OGC/ISO, but to avoid confusion we henceforth refer to such schemas as CML schemas, and use conceptual schema for the more general data modelling notion. We also suggest CML schemas correspond to logical schemas in data modelling, and DL schemas correspond to physical schemas.

The correspondence between CML and logical schemas is partly due to the fact that each entity in a CML schema instantiates an entity from the OGC/ISO General Feature Model (GFM; [23]) resulting in both conceptual and technological implications. Conceptually, the GFM describes a spatial meta-type (called FeatureType) and implies its instances (e.g. WaterWell type) are likely to have spatial properties such as location and shape – it thus implies a limited geographical ontology. Technologically, some instances of GFM entities can best be seen as

technical artefacts that have more to do with the representation of the entity than about the entity itself. These include mandatory role names for relations, a signature for each operation, and a stereotype for each entity: roles do not necessarily exist within domains, operation signatures are computational necessities, and though stereotypes can be seen as denoting the instantiation of a meta-type, their inclusion is also essential to certain OGC technologies, such as a Web Feature Service operating only on feature types. In fact, some CML schemas are developed around implementation targets and are therefore heavily shaped by technological concerns (e.g. GeoSCiML Lite; [9]). Consequently, the instantiation of representational entities as well as the overall influence of related technologies results in CML schemas with varying technological flavour, making them in this respect quite similar to logical schemas.

CML application schemas must also integrate with CML standard schemas, and this forced integration can impose ontological commitments that strongly influence the direction of the CML application schema, increasing the possibility of conceptual bias shaping the data structure. This suggests the need for a more conceptually neutral schema that can defer some ontological commitments, such as geometry representation, to the logical schema when appropriate. Indeed, this is the main point here: regardless of whether CML schemas are understood to be logical schemas (as we suggest and follow in this paper) or hybrid logical-conceptual schemas, there is room and need in the design method for the inclusion of a more general conceptual schema.

Advantages to such a conceptual schema are well documented in data modelling, such as improved consistency, portability, and clarity of design across technologies [27, 37]. Inclusion of a conceptual schema in the specification of any aspect of an information system, such as the data layer, is also implied by the influential Zachman enterprise architecture framework [42], while the potential for application of a tri-schema method to any markup language is also recognized though not commonly applied (e.g. [27] p. 537). Early experiments in testing the tri-schema approach involve groundwater and geological data transfer standard development [2, 33], with results suggesting that the noted advantages do extend to standards design. The advantages are further supported by the recently published GeoSciML data standard for geology [9], which contains two logical schemas, full and "lite" (for map layers), derived from a highly abstracted pre-existing conceptual schema [29]. Evidence is thus mounting for the inclusion of a true conceptual schema in the data standards design process.

Method for geospatial data standard design

The design method developed in this work is aligned with both OGC/ISO and INSPIRE approaches [14, 23], with some steps aggregated for simplicity and, most significantly, with the addition of a conceptual schema. The method consists of five steps with distinct products, as shown in Fig. 1: (1) usage scenarios, (2) vocabulary, (3) schemas, (4) implementation, and (5) evaluation. Each of the steps implements the results from the previous step. Note these steps are restricted to technical design, with other aspects of standards development deemed out of scope here: e.g. publication, communication, governance, and authority (related to the standing of the developers and the influence of the standard).

Usage scenarios describe significant activities and situations encountered by expected users of the standard. The scenarios can be broadly representative of a class of uses, or be quite specific about particular uses, as dictated by the nature of the domain and purpose of the standard. They are expressed in non-technical terms

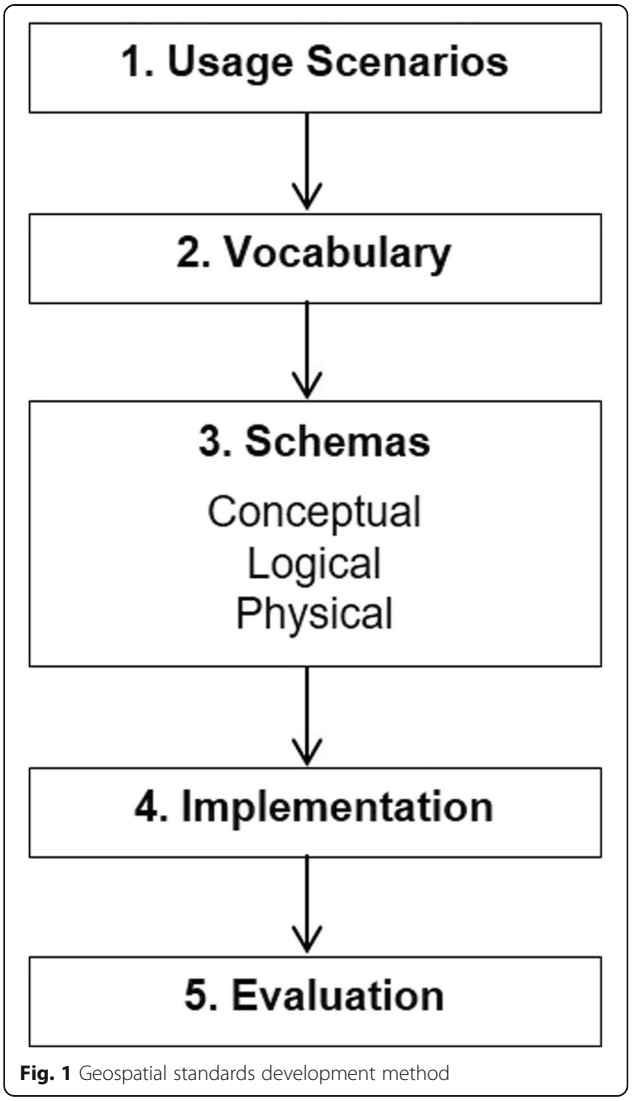

Fig. 1 Geospatial standards development method

using the language of the domain, and are intended to be understood by domain practitioners, e.g. by hydrogeologists for GWML2. In total, the scenarios should cover the breadth of entities to be represented and should explain how the entities are to be used to resolve some significant issue, perform some important task, answer some competency question, or constrain the quality, structure or delivery of the data. They should prioritize conceptual importance within the domain over data availability in information systems, as information concerns are weighed more heavily during logical schema design. They can be seen as encompassing the requirements for the data standard as well as specifying the evaluation criteria that it must meet: if use of the standard can be shown to help resolve the issue, answer the question, or meet the task, then its design is satisfactory. The scenarios can be represented with a variety of approaches for capturing requirements, use-cases, or competencies (e.g. [1, 35]).

Vocabulary development advances scoping by extracting from the use cases the things to be represented in the conceptual schema. Crucial entities are identified, named, and a definition is provided from an authoritative source to ensure common understandings. Entity selection typically involves in-depth consultation with international domain experts, such as hydrogeologists and hydrogeological data managers in the case of GWML2. This vocabulary is usually informally represented, for example, it might lack relationships between terms and each term is likely characterized in an unstructured way, inasmuch as its definition is probably described via narrative text rather than, say, logical conditions. Thus, while the vocabulary identifies conceptual boundaries, it does not provide a fully-structured representation, which in fact is the role of the conceptual schema. However, it can be represented in a semi-structured formalism, such as those used for glossaries and thesauri (e.g. [38]), various catalogs and dictionaries (e.g. [20, 25]), or even ontology languages (e.g. OWL), though a rigorous application of the latter would be quite heavy-handed for such lightly structured vocabularies. Apart from its contribution to the conceptual schema, the vocabulary can also contribute to data discovery by supplying domain terms for keyword-based search over various information systems or even the web.

The conceptual schema organizes the vocabulary into a fully structured knowledge representation, such that each vocabulary item is realized as a relation or type. Each entity is also fully documented with an accompanying narrative, including the retention of definitions from the vocabulary. The conceptual schema minimizes ontological commitments, such as commitments to specific schemas for geometry, coordinate systems, or observations, thus spatiality is not necessarily prioritized. It also defers commitments to technological environments, such as particular schemas for encoding languages, e.g. tuple-based or graph-based, which might affect how and if certain entities are modelled. Pragmatic concerns are also deferred, such as the quantity or quality of information collected about a certain entity, which might affect cardinalities: the number of things to which an entity is related in reality might then differ from the number to which it is related in an information system, largely because an information system's holdings can be partial and incomplete for any variety of reasons. In general, the conceptual schema strives to model an entity as it is in reality, rather than how it exists in an information system.

The conceptual schema can be expressed by a variety of formalisms, for example, logics such as first-order logic or a description logic, various ontology-based languages such as RDF or OWL, or conceptual modelling frameworks such as the Unified Modeling Language (UML; [35]), which is the primary choice for OGC/ISO [22]. Importantly, no representation is completely technology-neutral, as any conceptual modelling language will introduce some representational biases that should be minimized.

While the conceptual schema is conflated with the logical schema in many OGC domain data standards, we consider distinct development of a conceptual schema to be vital to the design of a data standard for several reasons: it is more easily understood by domain specialists than technology-laden logical or physical schemas, making it easier to engage them during design and usage; it provides a stable foundation for representation within the domain, one that is independent of technological changes and minimizes conceptual biases and conflicts, thus reducing both development and maintenance efforts; it eases schema development, by separating domain concerns from technological concerns, allowing focus on a single set of domain problems instead of a mixture of domain issues, technological issues, and varied understandings of existing schemas and their conceptual commitments—as a side-effect, this also allows more focused allotment of domain versus technological expertise; and most importantly, it provides the flexibility to have multiple logical implementations. A negative consequence of this approach is the increased maintenance cost of keeping conceptual and logical schemas aligned, however we consider the benefits to far outweigh the detriments.

The logical schema implements some or all of the conceptual schema in a particular technological framework, taking into account information concerns such as data availability. For OGC data standards this follows the rules for application schema development, which highlight alignment with the GFM [22, 23]. Of note is assignment of a GFM meta-type to each domain type (via an UML stereotype such as <<FeatureType>>), and possibly the specialization of each domain type from some standard schema such as Observations and Measurements

[7], or from some domain application schema such as GeoSciML. While the selected content of the conceptual schema must be fully incorporated into the logical schema, with original meanings intact, the structural correspondence between these schemas is not necessarily isomorphic as technological needs framing the logical schema might lead to some structural divergence. Documentation and definitions are cascaded to the logical schema from the conceptual schema, and development of the logical schema is also optional, i.e. if the goal of the standard is to develop a conceptual schema only, then the logical schema is obviously unnecessary. The logical schema is represented using a conceptual modelling formalism.

The physical schema implements some or all of the logical schema, and typically refers to a schema intended for data transfer and encoded using a specific data language e.g. XML [31, 36]. It could also refer to a schema for data storage, such as a particular relational database design. Out of scope here are schemas for data access, such as for web services, though such standards might mandate the use of certain schemas for data discovery or delivery. The purpose of the physical schema is to specify the pattern for a particular implementation, as well as associated rules and best practices, often following constraints imposed by a standards body, e.g. GML-XML schemas must follow OGC/ISO encoding rules [32]. As with the logical schema, the physical schema is optional if it is not the standard's target.

Implementation and evaluation then ensure the developed schemas are implementable in a satisfactory way. Implementation can vary according to the targeted schema: conceptual and logical schemas are likely implemented in Semantic Web frameworks, e.g. to mediate in data interoperability scenarios or annotate web pages, while physical schemas are likely implemented in data storage or data transfer environments where they are populated with data and used. In all cases evaluation can involve criteria such as deployability, completeness, and usability. Deployability refers to relatively uncomplicated and efficient use within a broad range of information systems. Completeness refers to the degree of implementation of the conceptual schema, as well as to its fit to reference datasets – problems arise if a logical or physical schema does not fully or adequately cover the conceptual schema, or if any schema does not fully or adequately cover the entities identified in the vocabulary or the key contents of target datasets. Usability refers to satisfaction of the usage scenarios and can include a cost/benefit analysis.

Results – application to GWML2

The five-stage data standard design method was successfully applied to the development of GWML2.

GWML2 Usage scenarios

GWML2 usage scenarios include five data delivery cases: commercial, policy-oriented, environmental, scientific, and technological. The commercial scenario searches for water wells and springs within an area to estimate the cost to complete a new nearby water supply well, and involves water wells, related measurements, and hydrogeological units such as aquifers. The policy scenario is aligned with European INSPIRE needs [16] to enable reporting on the state of groundwater in administrative districts, and involves management areas, related hydrogeological units and monitored information. The environmental scenario enables environmental managers, water managers, and legislators to assess threats to groundwater dependent ecosystems, and involves depth to water table, monitored information on groundwater chemistry and biology, and flow between groundwater and surface water. The scientific scenario focusses on the delivery of data for use in groundwater flow modelling and soil-water balance modelling, and involves the hydrogeological and geophysical properties of aquifers and related measurements, as well as the characteristics of water wells, water bodies, and water use. The technologic scenario determines compatibility with other hydrogeological data representations, such as database schemas and exchange formats, via conversion to and from GWML2. This is particularly important to enable data interoperability of all entities from the previous scenarios (i) within a groundwater data network, by converting between local databases and GWML2, and (ii) between different data networks, by converting between GWML2 and local data formats, such as the European-wide INSPIRE hydrogeology standard [15, 16] or the North American GWML1 [2]. A full description of the usage scenarios is available at the GWML2 wiki [12].

GWML2 Vocabulary

The selection of key terms from the usage scenarios as well as agreement about their meanings involved lengthy discussions and voting by majority rule. Authoritative definitions were selected from the scientific literature and assigned to each term. Example terms include aquifer, water well, groundwater basin, porosity, and flow. The terms and their definitions are represented as a list online at the GWML2 wiki [10].

GWML2 Schemas

The GWML2 conceptual and logical schemas are expressed in UML, and the physical schema is represented as a GML-XML schema accompanied by associated rules and examples. All three schemas are available in the GWML2 SVN repository [11], and the physical schema is also available online from OGC [30]. The three schemas are summarized below, and their full

descriptions can be found in the OGC standards specification [4].

GWML2 Conceptual Schema

The GWML2 conceptual schema represents all items found in the vocabulary. It consists of hydrogeological units, fluid bodies, voids, fluid flow, and water wells and associated things. Together, these entities form a simple pattern for water containment, originally outlined in [2] and refined herein: the fluid body is enclosed by a container, such as a hydrogeological unit, and occupies the spaces in the container (i.e. its voids). Fluid flows within and between containers and their spaces, and fluid is added, removed, or observed using natural and artificial artefacts, such as water wells, springs, and monitoring sites. It is particularly important to conceptually differentiate voids from units and fluid bodies to distinguish the unique properties hosted by each, such as their distinct volumes. The main entities and relations of conceptual schema are shown in Fig. 2, separated by colour into five broad categories of entities.

Hydrogeological units, such as aquifers, confining beds, aquifer systems or groundwater basins, are distinct volumes of earth material that serve as containers for subsurface fluids and have physical boundaries typically delineated along natural discontinuities related to fluid flow. This contrasts with management areas, which are terrain volumes typically delineated by social factors such as policy or regulation. Note that several properties usually attributed to hydrogeological units, such as porosity, permeability, and conductivity, are represented more accurately here as properties of the relation of the unit to a fluid body or void. Groundwater fluid bodies are distinct bodies of fluid, either liquid or gas, that fill the voids in hydrogeological units, and are made of biologic (e.g. organisms), chemical (e.g. solutes), or material (e.g. sediment) constituents, have other fluid bodies as parts, such as plumes or gas bubbles, and can host surfaces, such as a water table. Voids are variously sized spaces inside a hydrogeological unit (e.g. an aquifer) or its material (e.g. the sandstone constituting an aquifer), and might contain fluid bodies.

Groundwater flow denotes the process by which a fluid moves within, or between, containers or voids. It includes recharge as flow into a container, and discharge as flow out of a container, with a flow path being a sequenced collection of flows from recharge to discharge, and water budgets being the balance of flow for a container over a time. Water wells are man-made constructions for monitoring, withdrawing, or injecting water from/into a hydrogeological unit, while springs are features where water discharges to the surface naturally.

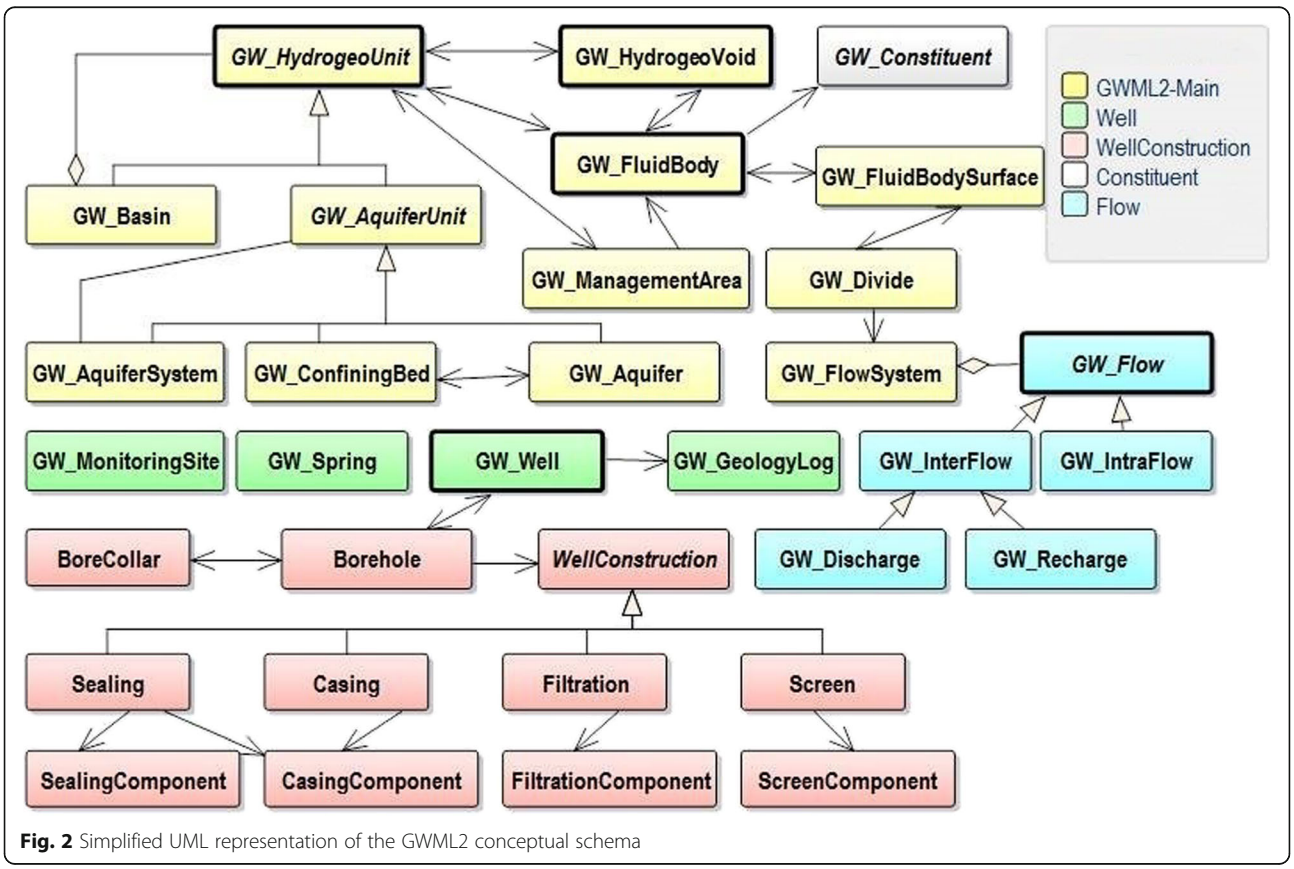

Fig. 2 Simplified UML representation of the GWML2 conceptual schema

Monitoring sites are locations where sensing devices are placed to measure significant hydrogeological properties, such as water level, flow rate, water temperature, or chemical composition, and where samples can be taken.

Figure 3 illustrates a natural setting with relevant GWML2 representations (prefixed with "GW"), and Fig. 4 shows an UML excerpt for the water container pattern. Figure 5 illustrates the movement of water in the subsurface, including its interactions with the ground surface and related water bodies during, for example, recharge and discharge, and the related UML representation is shown in Fig. 6.

GWML2 Logical Schema

The GWML2 logical schema implements all components of the GWML2 conceptual schema and follows its modular structure. It differs from the conceptual schema in two respects: (1) in its re-orientation from representing entities in reality to data in information systems, and (2) in its introduction of technology-laden artefacts from the GFM as well as its adoption of other OGC schemas.

The first difference involves the relaxation of existence constraints in which objects that exist in reality do not necessarily exist in information systems, due to data collection practices or other pragmatic factors. This relaxation enables compact syntactical encodings in which missing data has no footprint in a digital file. For example, while it is mandatory that a fluid body occupy a void in the conceptual schema (see Fig. 4), and hence that the void exists, it is optional for it to occupy a void in the logical schema (see Fig. 7), mainly because voids are often neither explicitly identified nor stored in hydrogeological information systems. As such, fluid bodies in syntactical encodings need not be accompanied by voids.

The second difference involves two additions: GFM meta-types are added as OGC stereotypes to each UML class, and several OGC-compliant application schemas are incorporated. The stereotypes most significant to

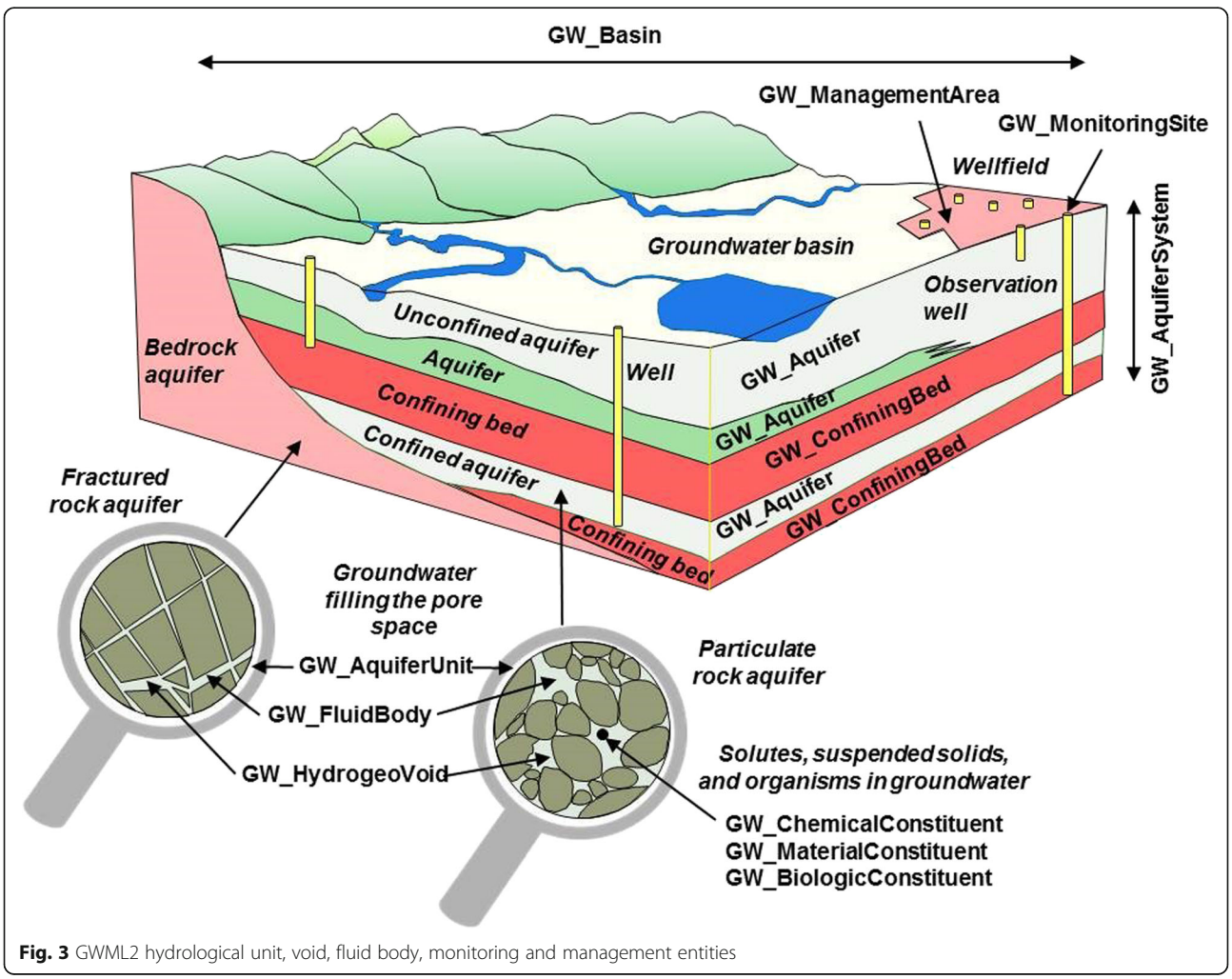

Fig. 3 GWML2 hydrological unit, void, fluid body, monitoring and management entities

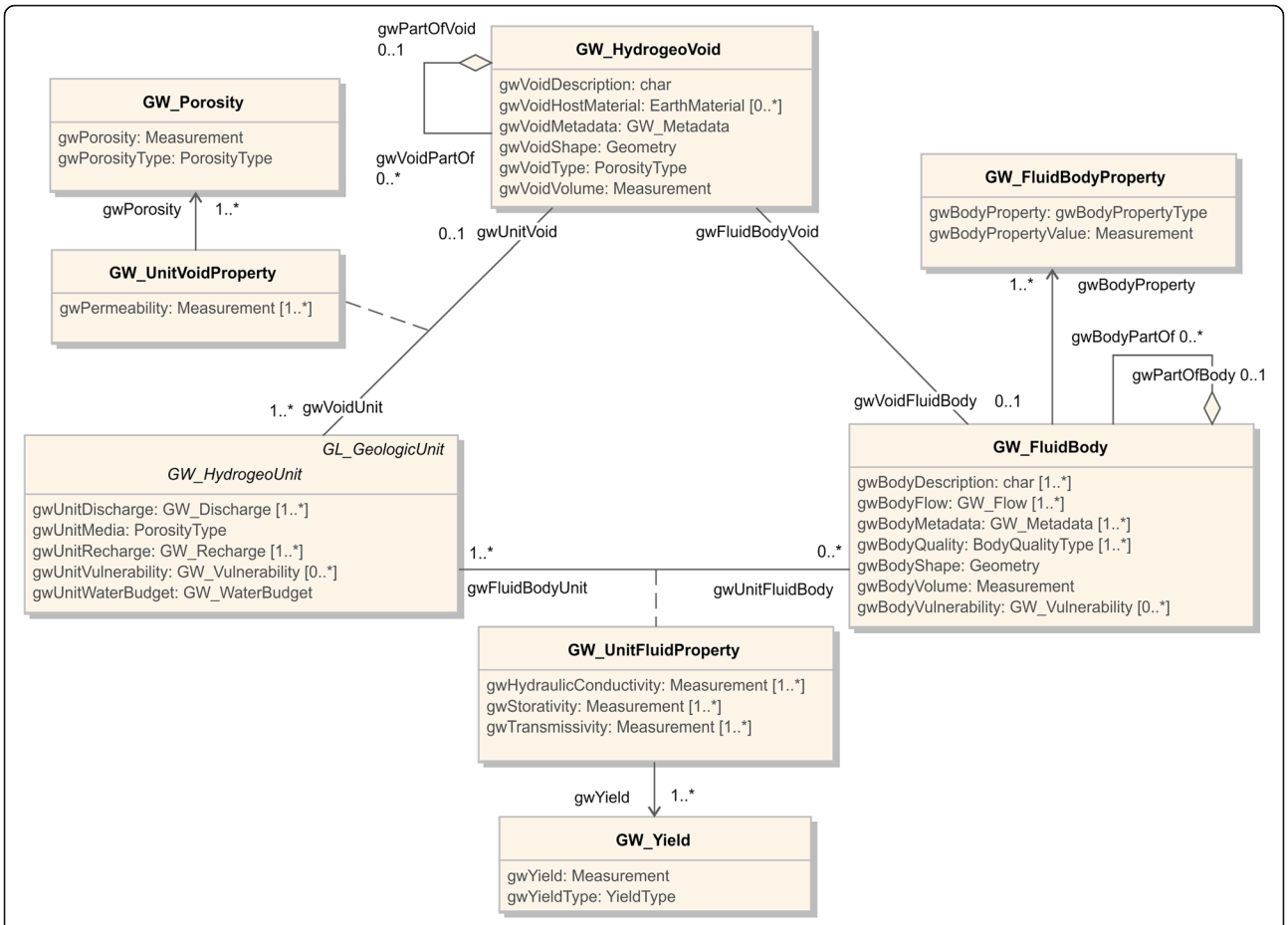

Fig. 4 GWML2 conceptual schema fragment for hydrogeological unit, fluid body, and void

GWML2 include <<FeatureType>>, <<type> > and < <DataType>>. <<FeatureType> > refers to entities that are OGC features, which we interpret for GWML2 purposes to have identity in reality, are geographically located, and have physical unity – i.e. they are a distinct and cohesive physical body – such as GW_HydrogeoUnit in Fig. 7. In addition, an entity stereotyped as <<FeatureType> > can replace name, description and identifier properties from the conceptual schema with equivalents inherited from <<FeatureType> > (e.g. gwBodyDescription in the conceptual schema can be replaced by AbstractFeature::description in the logical schema). <<type> > and < <DataType> > refer to entities that we understand as not necessarily having physical unity such as an amount of material, e.g. sand or water, or groups of properties, such as GW_UnitVoidProperty in Fig. 7.

The additional schemas that are imported or adapted by GWML2 include: GML [32], MD Metadata [19], Observations & Measurements (O&M; [7]), Sensor Web Enablement (SWE; [34]), TimeSeriesML [40], GeoSciML 4.1 [9], and GWML1 [2]. These are incorporated using the following strategies:

Specialization Some GWML2 entities specialize types from the other schemas. For example, GW_HydrogeologicalUnit specializes GeologicUnit from GeoSciML, recognizing that in its most basic sense a hydrogeological unit is a body of rock (a geological unit) exhibiting some hydraulic properties including possible fluid storage and transfer. In another example, water wells and boreholes specialize SF_SamplingCurve from Observations and Measurements, recognizing that the description of wells and boreholes involves observations along their length.

Property ranges In general, under-defined property ranges in the conceptual schema are replaced with well-defined entities from OGC standard schemas or some other OGC compliant schema. Such ranges refer here to the type for a property internal to an entity, or the type participating in an external relation with an entity. For example, internal properties that denote observed data, or are derived from observed data, have their range replaced by OM_Observation from O&M, e.g. the range for gwPorosity in the conceptual schema (i.e. Measurement) is replaced with OM_Observation.

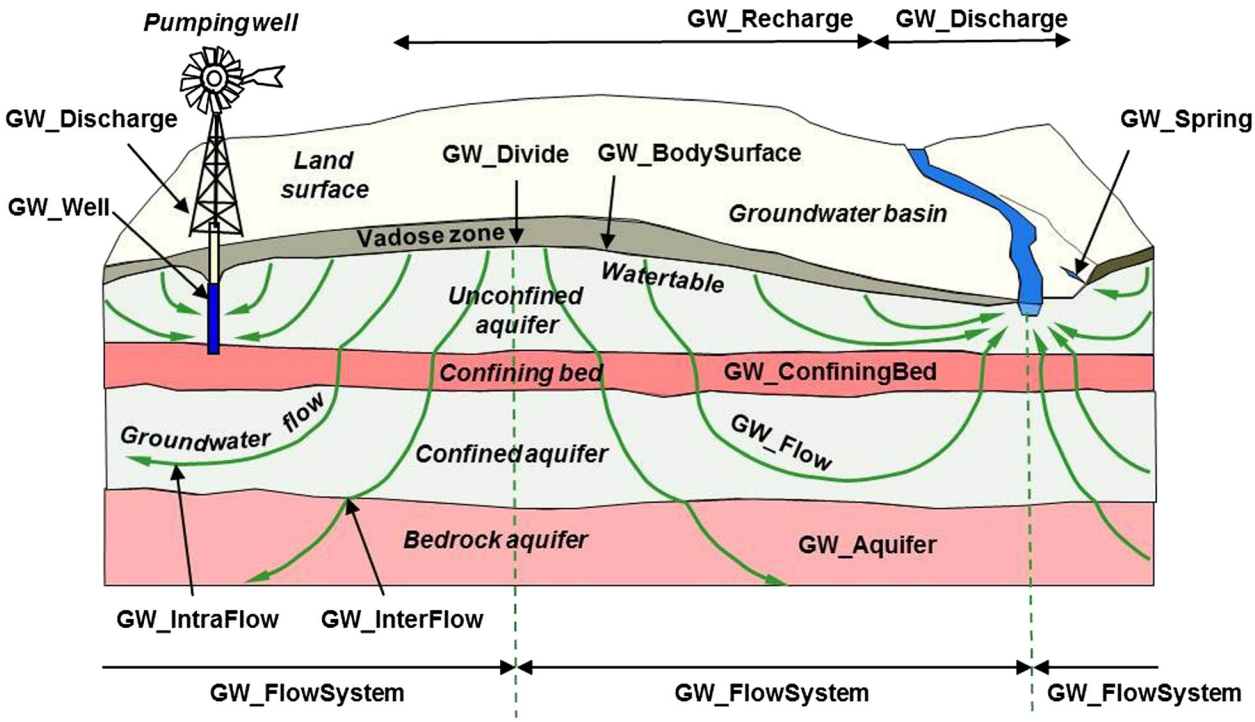

Fig. 5 GWML2 groundwater flow and well entities

Two factors compel this choice: OM_Observation enables method metadata to be added to each observed value, and each property can be soft-typed for greater precision, such as gwPorosity further delineated into effective porosity, primary porosity, or secondary porosity. In some cases internal property range substitutions are made to enable dynamic linkages to schemas that are not pre-determined, such as the substitution of AbstractFeature::GFI_Feature for Feature in various flow entities, to enable the possibility that the sources and targets for groundwater flow, such as rivers, can be specified in other domain schemas. Another example is the substitution of Any for the external document property of GW_ManagementArea, to allow a variety of possible documents to be referenced such as GW_Licence, MD_Metadata, or INSPIRE's DocumentCitation or LegislativeReferences types [17].

Observations In some cases it is not only property ranges that are replaced with OM_Observation, but properties themselves. For example, the gwConcentration internal property in the conceptual schema, which denotes the concentration of some constituent in a fluid body, is represented in the logical schema as an observation related to a constituent such as Arsenic. Such substitutions result in structural, but not conceptual, differences between the logical and conceptual schemas.

GWML2 Physical Schema

The physical schema implements the logical schema as a GML-XML schema with associated encoding rules expressed in Schematron [26]. The schema was generated semi-automatically and iteratively using the Fullmoon software tool to ensure adherence to OGC/ISO encoding standards [21]. Also included are encoding examples for each entity to demonstrate the application of the schema and rules.

GWML2 Implementation

Implementation involved the generation of GWML2 encoding examples and the deployment of web services over existing information systems. Figure 8 shows an example GML-XML encoding of groundwater discharge (GW_Discharge), in which groundwater is flowing from an Australian aquifer to a particular lake. Note the presence of various logical schema artefacts, such as the use of gml:description for descriptive narrative (realized from the <<FeatureType> > stereotype and its AbstractFeature::description property), and the use of OM_Observation to encode the value of various properties, such as gwFlowVelocity, including procedural metadata such as date, time, and unit of measure. Example encodings are available from the public OGC GWML2 repository [30].

Implementation also involved nine organizations deploying twelve OGC standard web services [3]. Each component of GWML2 was implemented against at least

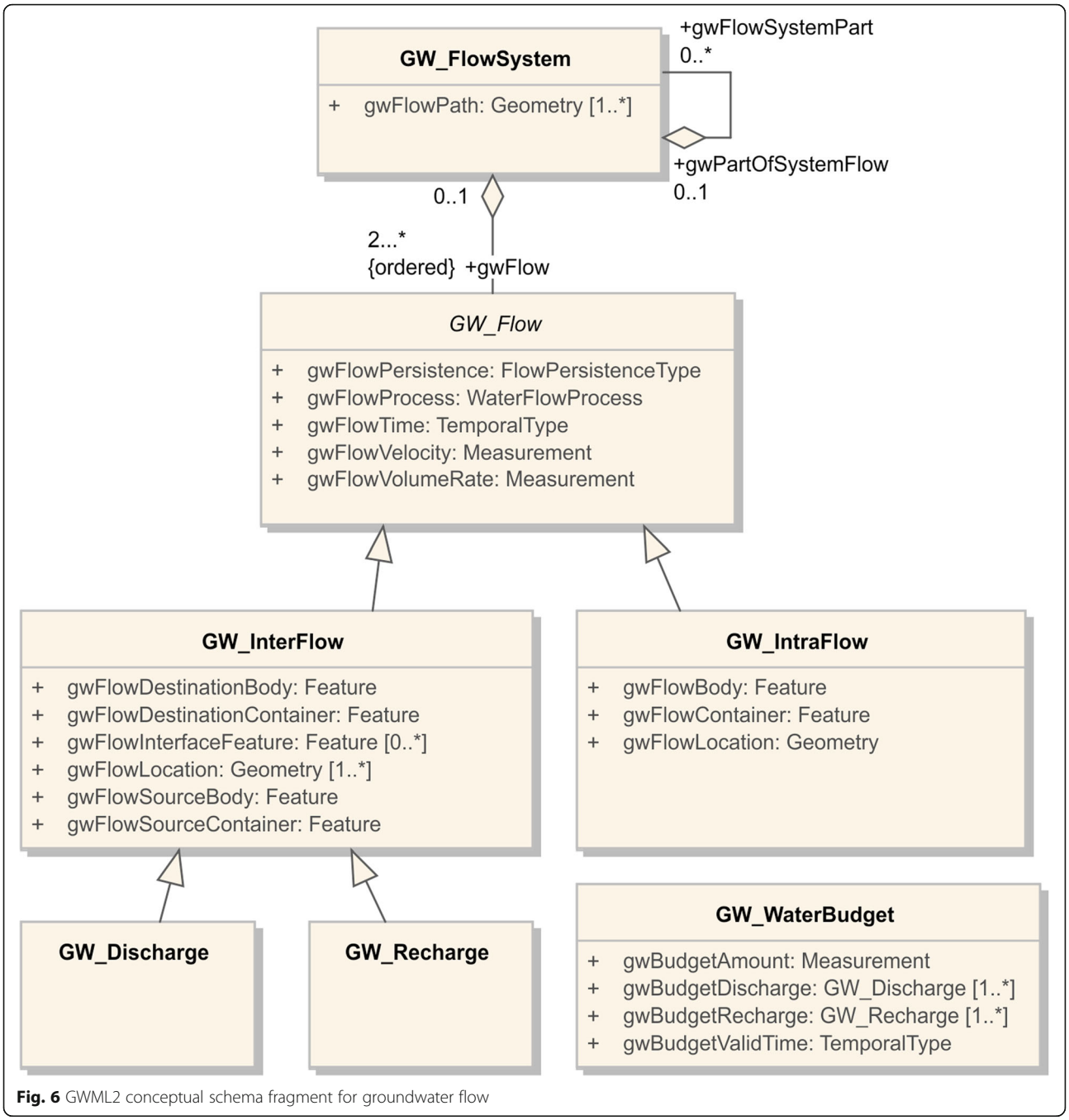

Fig. 6 GWML2 conceptual schema fragment for groundwater flow

one significant groundwater data repository, and many components were implemented against several repositories. Four kinds of web services were deployed: (1) Web Map Service (WMS; [8]), (2) Web Feature Service (WFS; [41]); (3) Sensor Observation Service (SOS; [6]), and (4) Web Processing Service (WPS; [28]). These implementations are summarized in Table 1, including the kind of data delivered by each web service. Participants included the Geological Survey of Canada (GSC), United States Geological Survey (USGS), Bureau of Meteorology (BOM), Commonwealth Scientific and Industrial Research Organization (CSIRO), Bureau de Researches Géologiques et Minières (BRGM), Federation University Australia (FedUni), University of Salzburg (Z_GIS), the Institute of Geological and Nuclear Sciences Institute Ltd. (GNS), and the Geological Survey of Bavaria (LfU).

GWML2 Evaluation

Successful evaluation of GWML2 is demonstrated by satisfaction of the criteria from Section 3: deployability,

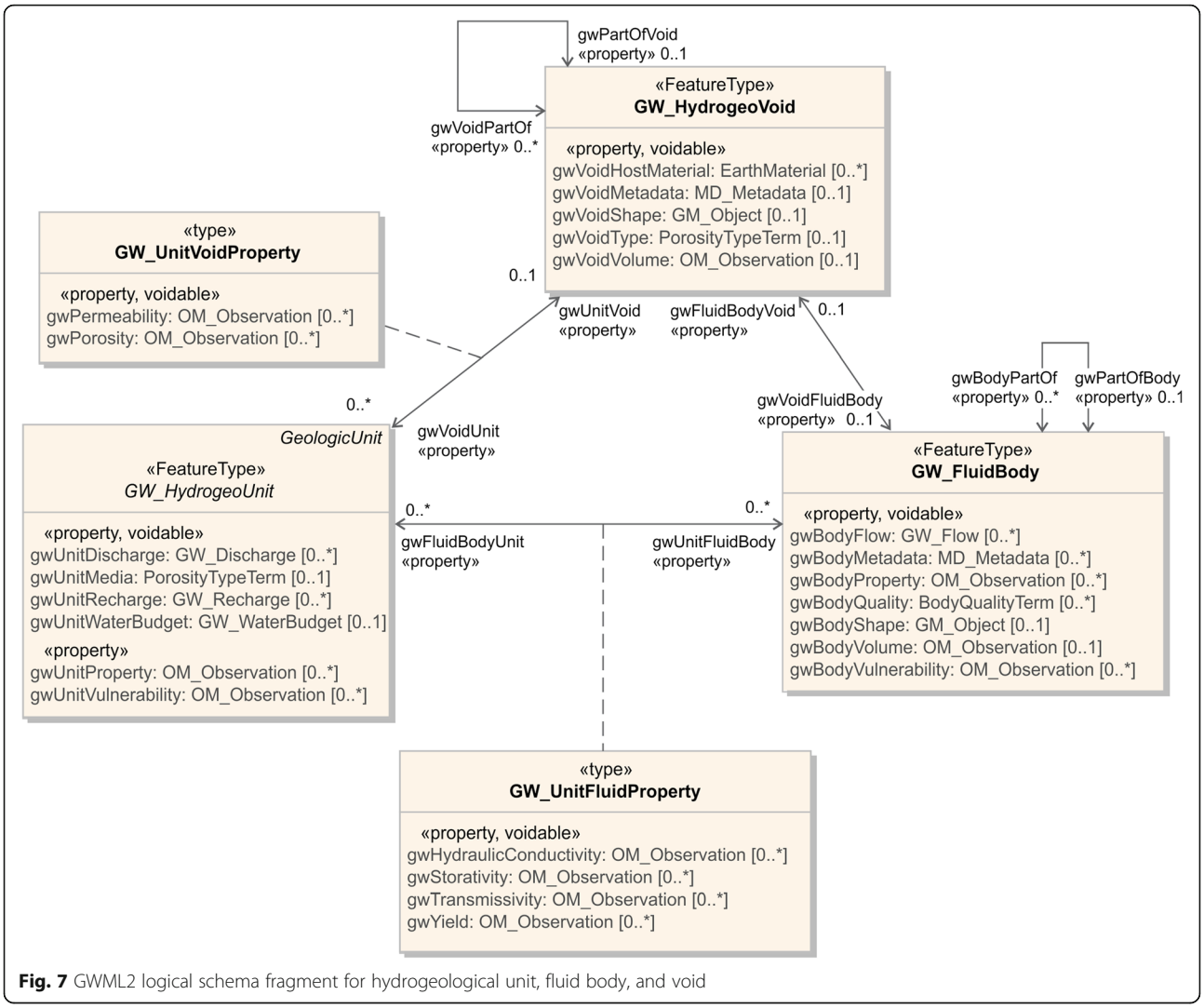

Fig. 7 GWML2 logical schema fragment for hydrogeological unit, fluid body, and void

completeness, and usability. Deployability is shown via schema validation and efficiency: the GWML2 schema validated syntactically as a compliant XML schema in a wide variety of technological environments, and while efficiency measurements were not recorded explicitly, in all cases the web services returned results in times that were deemed acceptable to humans. Completeness is demonstrated by the successful syntactical encoding of all conceptual schema components, and by minimal information loss in mapping GWML2 to the various data repositories. Usability is demonstrated by satisfaction of the five usage scenarios, including the delivery of all data required by the scenarios via the web services and the GWML2 physical schema. Further details about subsequent manipulation of the data to meet the usage scenarios is detailed in [5]. This successful evaluation of GWML2 also implies successful evaluation of its design method, including the addition of a conceptual schema, insofar as application of the method led to appropriate results.

Discussion

Under the tri-schema approach, OGC/ISO CML schemas are understood to be logical schemas, largely because they instantiate representational entities from the GFM, are often shaped by technological concerns, and can impose limiting ontological commitments. However, even if this were not the case and CML schemas are understood to be quasi-conceptual schemas, there is still great value in establishing a more general conceptual schema that minimizes ontological and technological commitments. Indeed, if technological and non-essential ontological commitments are deferred to a logical schema, which derives from a separate conceptual schema, then our experience suggests the resulting standard will be more flexible, adaptable, and conceptually consistent, as well as more efficiently and effectively constructed.

These benefits were all realized in GWML2 development. Of particular note was the optimal deployment of diverse expertise: domain scientists were more engaged

```
▼<gwml2f:GW_Discharge gml:id="lake-murdeduke-discharge-zone_19930101">
 ▼<gml:description>
    Example Interflow instance of fluid flow from Tertiary-Quaternary Basalt Aquifer
    to Lake Murdeduke based on data from Jane Coram MSc 1996
  </gml:description>
 ▶<gml:identifier  codeSpace="http://www.vvg.org.au/">...</gml:identifier >
  <gml:name codeSpace="http://www.vvg.org.au/">Lake Murdeduke western discharge zone
  January 1993</gml:name>
 ▼<gwml2f:gwFlowTime>
   ▶<gml:TimePeriod gml:id="lake-murdeduke-discharge-zone_19930101.timeperiod">...</gml:TimePeriod>
 ▼</gwml2f:gwFlowTime>
 ▼<gwml2f:gwFlowVelocity>
   ▼<om:OM_Observation gml:id="lake-murdeduke-discharge-zone_19930101.flowvelocity ">
      <om:phenomenonTime xlink:href="#lake-murdeduke-discharge-
      zone_19930101.timeperiod"/>
      <om:resultTime xlink:href="#lake-murdeduke-discharge-
      zone_19930101.timeperiod"/>
      <om:procedure
      xlink:href="http://www.vvg.org.au/def/procedure/flowvelocity/V=Ki/n "
      xlink:title="V=Ki/n"/>
      <om:observedProperty xlink:href="http://www.opengis.net/gwml-
      flow/2.2#gwFlowVelocity " xlink:title="gwFlowVelocity"/>
      <om:featureOfInterest xlink:href="#lake-murdeduke-discharge-zone_19930101"/>
     ▼<om:result>
      ▼ <swe:Quantity>
          <swe:uom code="m/mth"
          xlink:href="http://www.vvg.org.au/def/flowvelocity/metrespermonth "
          xlink:title="metres per month"/>
          <swe:value>2.2</swe:value>
        </swe:Quantity>
      </om:result>
    </om:OM_Observation>
  </gwml2f:gwFlowVelocity>
 ▼<gwml2f:gwFlowVolumeRate>
   ▼<om:OM_Observation gml:id="lake-murdeduke-discharge-
    zone_19930101.flowvolumerate ">
      <om:phenomenonTime xlink:href="#lake-murdeduke-discharge-
      zone_19930101.timeperiod"/>
      <om:resultTime xlink:href="#lake-murdeduke-discharge-
      zone_19930101.timeperiod"/>
      <om:procedure
      xlink:href="http://www.vvg.org.au/def/procedure/flowvolumerate/V=Ki/n "
      xlink:title="V=Ki/n"/>
      <om:observedProperty xlink:href="http://www.opengis.net/gwml-
      flow/2.2#gwFlowVolumeRate " xlink:title="gwFlowVolumeRate"/>
      <om:featureOfInterest xlink:href="#lake-murdeduke-discharge-zone_19930101"/>
     ▼<om:result>
      ▼ <swe:Quantity>
          <swe:uom code="m3/mth"
          xlink:href="http://www.vvg.org.au/def/flowvolumerate/cubicmetrespermonth "
          xlink:title="cubic metres per month"/>
          <swe:value>80427.0</swe:value>
        </swe:Quantity>
      </om:result>
    </om:OM_Observation>
  </gwml2f:gwFlowVolumeRate>
  <gwml2f:gwFlowDestinationContainer
  xlink:href="http://www.vvg.org.au/feature/gde/gde_cutdown_region.2080"
  xlink:title="Lake Murdeduke"/>
  <gwml2f:gwFlowInterfaceFeature xlink:href="http://www.vvg.org.au/uri-
  cgi/feature/gwml2/spring/feduni.spring.126" xlink:title="Lake Murdeduke western
  discharge zone"/>
  <gwml2f:gwFlowSourceBody xlink:href="http://www.vvg.org.au/feature/gwml2/fluid-
  body/feduni.fluidbody.13 " xlink:title="Upper Tertiary-Quaternary Basalt Fluid
  Body"/>
  <gwml2f:gwFlowSourceContainer
  xlink:href="http://www.vvg.org.au/feature/gwml2/aquifer/feduni.aquifer.13"
  xlink:title="Upper Tertiary-Quaternary Basalt Aquifer"/>
</gwml2f:GW_Discharge>
```

Fig. 8 GWML2 GML-XML encoding example for GW_Discharge

Table 1 GWML2 implementation via OGC web services

GWML2 Feature Type	OGC Service	Provider	URL
GW_Well GW_Aquifer GW_AquiferSystem GW_ConfiningBed GW_MonitoringSite	WMS	GSC USGS	http://gin.gw-info.net/service/gin/wms/mediator/gin_en?service=WMS&version=1.1.0&request=GetCapabilities http://cida.usgs.gov/ngwmn-geoserver/ows?service=wms&version=1.3.0&request=GetCapabilities
GW_Aquifer GW_AquiferSystem GW_ConfiningBed GW_Discharge GW_FluidBody GW_ManagementArea GW_MonitoringSite GW_Recharge GW_Well GW_Spring Borehole	WFS	GSC BOM/ CSIRO BRGM USGS FedUni GNS/Z_GIS	http://gin.gw-info.net/GinService/wfs/gwie?REQUEST=GetCapabilities&ACCEPTVERSIONS=2.0.0&SERVICE=WFS http://gwservices.it.csiro.au:8080/geoserver/ows?service=wfs&version=1.1.0&request=GetCapabilities http://geoserverref.brgm-rec.fr/geoserver/ows?service=wfs&version=2.0.0&request=GetCapabilities http://cida.usgs.gov/ngwmn_cache/wfs?version=1.1.0&service=wfs&REQUEST=GetCapabilities http://data.vvg.org.au:8080/geoserver/wfs?version=1.1.0&request=GetCapabilities http://portal.smart-project.info/gs-smart/wfs?request=GetCapabilities&service=WFS(deprecated)
WML2::TimeSeries	SOS	GSC BRGM USGS GNS / Z_GIS	http://gin.gw-info.net/GinService/sos/gw?REQUEST=GetCapabilities&ACCEPTVERSIONS=2.0.0,1.0.0&SERVICE=SOS http://ressource.brgm-rec.fr/service/sosRawPiezo/service=SOS&version=2.0.0&request=GetCapabilities http://cida.usgs.gov/ngwmn_cache/sos?request=GetCapabilities&service=SOS&AcceptVersions=2.0.0 http://portal.smart-project.info/sos-smart/service/kvp?request=GetCapabilities&service=SOS
Same as WFS and SOS above	WPS	GNS/Z_GIS	http://portal.smart-project.info/wps/WebProcessingService?Request=GetCapabilities&Service=WPS(deprecated)

in early stages (up to and including conceptual schema development) and technological experts more engaged in subsequent stages, enabling more focussed and relevant contributions. Data modelling issues were also easier to resolve through the well-defined modelling structure and process: the tri-schema segregation allowed problems to be isolated to specific schemas and associated expertise, which further enabled solutions to be cascaded to all dependent schemas and pertinent experts. The long-terms costs of maintaining the three schemas is yet to be determined, but maintenance has not proven to be onerous during the early life of GWML2.

Conclusion and future directions

This work includes both product and methodological innovations: (1) as a product, GWML2 is the first global groundwater data standard designed to work with open geospatial technologies, and (2) methodologically, it is the first product within the eco-system of OGC/ISO geospatial data standards in which all schemas from the tri-schema approach are explicitly developed during the data standard design process.

Future work on GWML2 includes the development of additional syntactical encodings with associated physical schemas. The most notable of such encodings are JSON and RDF/OWL, which would facilitate data exchange with web applications and Semantic Web initiatives, respectively. A foray into RDF/OWL raises many questions. For instance, which GWML2 schema is to be targeted as a prospective RDF/OWL ontology? A superficial analysis suggests if the intended purpose is data exchange, then the logical schema is the natural target for an RDF/OWL ontology, given that the logical schema is intrinsically constructed with data transfer in mind. However, if the intended purpose is semantic interoperability, e.g. to interoperate with other domain ontologies such as those emerging for hydrology, then the optimal target for RDF/OWL ontology representation is likely the GWML2 conceptual schema (as begun in [13]). These and other GWML2 efforts will continue within the OGC Groundwater Standards Working Group.

Abbreviations

BOM: Bureau of Meteorology; BRGM: Bureau de Researches Géologiques et Minières; CSIRO: Commonwealth Scientific and Industrial Research Organization; FedUni: Federation University Australia; GeoSciML: Geoscience Markup Language; GFM: General Feature Model; GML: Geography Markup Language; GNS: Institute of Geological and Nuclear Sciences Institute Ltd.; GSC: Geological Survey of Canada; GWML1: Groundwater Markup Language 1; GWML2: Groundwater Markup Language 2; INSPIRE: Infrastructure for Spatial Information Europe; ISO: International Standards Organization; LfU: Geological Survey of Bavaria; OGC: Open Geospatial Consortium; OWL: Web Ontology Language; RDF: Resource Description Framework; SKOS: Simple Knowledge Organization System; SOS: Sensor Observation Service; TimeSeriesML: Time Series Markup Language; UML: Unified Modeling Language; USGS: United States Geological Survey; W3C: World Wide Web Consortium; WaterML2: Water Markup Language 2; WFS: Web Feature Service; WMS: Web Map Service; WPS: Web Processing Service; XML: Extensible Markup Language; Z_GIS: University of Salzburg

Acknowledgements

The Federation University of Australia also acknowledges the support of several contributors in the development and testing of GWML2, most notably Andrew MacLeod.

Funding

Funding for CSIRO and Bureau of Meteorology work was provided by Water Information Research and Development Alliance (WIRADA). Funding for Federation University Australia work by the 'Centre for eResearch and Digital

Innovation' and the 'Victorian Broadband Enabled Innovation Program'. Funding for BRGM work is provided by its joint IT Research Center 'INSIDE' dedicated to innovation in Environmental Information Systems with the French National Agency for Biodiversity (AFB), French Museum of Natural History (MNHN) and French Marine Agency (IFREMER). Funding for the SMART Aquifer Characterisation (SAC) programme at GNS Science, New Zealand, and the Interfaculty Department for Geoinformatics (Z_GIS), University of Salzburg, is provided by the Ministry of Business, Innovation, and Employment (MBIE), New Zealand. Funding for Natural Resources Canada (NRCan) participants is provided by the NRCan Groundwater Geoscience Program. Funding for U.S. Geological Survey participants is provided by the USGS Groundwater and Streamflow Information Program.

Availability of data and materials

All published GWML2 artefacts, such as schemas, reports, and encoding examples, are found online at OGC's permanent repository: http://www.opengeospatial.org/standards/gwml2. Additional materials are also available online from the OGC Groundwater Standards Working Group [10–12]. Online web services delivering GWML2 data are listed in Table 1.

Authors' contributions

All authors participated fully in this work from inception to completion. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Geological Survey of Canada, 601 Booth St. Ottawa, Ottawa, ON K1A0E8, Canada. [2]Federation University Australia, Ballarat, Australia. [3]Bureau de Recherches Géologiques et Minières (BRGM), Paris, France. [4]University of Salzburg (Z_GIS), Salzburg, Austria. [5]U.S. Geological Survey (USGS), Reston Virginia, USA. [6]CSIRO Land and Water, Canberra, Australia. [7]Geological Survey of Bavaria (LfU), Bavaria, Germany.

Received: 2 October 2018 Accepted: 8 November 2018
Published online: 23 November 2018

References

1. Alexander I, Beus-Dukic L. Discovering requirements: how to specify products and services: Wiley; 2009. 457 pp.
2. Boisvert E, Brodaric B. GroundWater markup language (GWML) – enabling groundwater data interoperability in spatial data infrastructures. J Hydroinf. 2012;14(1):93–107.
3. Brodaric B (ed) (2016a) OGC GroundWaterML2 – GW2IE Final Report, Open Geospatial Consortium Engineering Report 15-082, v2.1, 171 pp. http://www.opengis.net/doc/PER/GW2IE.
4. Brodaric B (ed) (2016b) OGC WaterML2: Part 4 – GroundWaterML2 (GWML2), Open Geospatial Consortium Standard 16-032r2, v2.2.0, 158 pp.
5. Brodaric B, Boisvert E, Chery L, Dahlhaus P, Grellet S, Kmoch A, Letourneau F, Lucido J, Simons B, Wagner B. Enabling global exchange of groundwater data: GroundWaterML2 (GWML2). Hydrogeology Journal, 26(3):733–741. 2018; https://doi.org/10.1007/s10040-018-1747-9. Accessed 07 Aug 2018.
6. Broring A, Stasch C, Echterhoff J, editors. OGC sensor observation service Interface standard. Open Geospatial Consortium Standard. 2012;12-006:v2.0 163 pp.
7. Cox S (2013) Geographic Information – Observations and Measurements v2.0, Open Geospatial Consortium Standard 10-004r3, 54 pp.
8. De la Beaujardiere (ed) (2004) Web Map Service. OGC Standard 03-109r1, version 1.3.0.
9. GeoSciML Modelling Team (ed) (2017) OGC geoscience markup Language 4.1 (GeoSciML). Open Geospatial Consortium Standard 16–008, v4.1, 247 pp.
10. GWML2 (2016a) GWML2 Vocabulary. http://external.opengis.org/twiki_public/HydrologyDWG/Gwml2FinalFeaturesList, Accessed 07 Aug 2018.
11. GWML2 (2016b) GWML2 SVN Repository. https://xp-dev.com/svn/gwml2. Accessed 07 Aug 2018.
12. GWML2 (2016c) GWML2 Use-cases. http://external.opengis.org/twiki_public/HydrologyDWG/GroundwaterInteroperabilityExperiment2. Accessed 07 Aug 2018.
13. Hahmann T, Stephen S. Using a hydro-reference ontology to provide improved computer-interpretable semantics for the groundwater markup language (GWML2). Int J Geogr Inf Sci. 2018;32:1138–71.
14. INSPIRE (2008) D2.6_v3.0 Drafting Team "Data Specifications" – deliverable D2.6: Methodology for the development of data specifications. D2.6_v3.0, 123 pp. http://inspire.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.6_v3.0.pdf, Accessed 07 Aug 2018.
15. INSPIRE (2013a) D2.8.II.4 INSPIRE Data Specification on Geology, Technical Guidelines. D2.8.II.4_v3.0. http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_ER_v3.0.pdf, Accessed 07 Aug 2018.
16. INSPIRE (2013b) D2.8.III.11 INSPIRE Data Specification Area Management/Restriction/Regulation Zones and Reporting Units, D2.8.III.11_v3.0. https://inspire.ec.europa.eu/id/document/tg/am, Accessed 05 Sep 2018.
17. INSPIRE (2013c) D2.5: Generic Conceptual Model, Version 3.4, D2.5_v3.4. https://inspire.ec.europa.eu/documents/inspire-generic-conceptual-model, Accessed 21 Sep 2018.
18. ISO. ISO 19107:2003 geographic information – spatial Schema. International standards organization, ISO/TC211 19103:2015. Switzerland: Geneva; 2003. 166 pp.
19. ISO. Geographic information – metadata – XML Schema implementation. International standards organization. In: ISO 19139:2007. Switzerland: Geneva; 2007.
20. ISO. ISO 19118:2011 geographic information – feature concept Dictionairies and registers. International standards organization, ISO/TC211 19103:2015. Switzerland: Geneva; 2009. 40 pp.
21. ISO. ISO 19118:2011 geographic information – encoding. International standards organization, ISO/TC211 19118:2011. Switzerland: Geneva; 2011. 69 pp.
22. ISO (2015a) ISO 19103:2015 Geographic Information – Conceptual Schema Language International Standards Organization, ISO/TC211 19103:2015, Geneva, Switzerland, 81pp.
23. ISO (2015b) ISO 19109:2015 Geographic Information – Rules for Application Schema. International Standards Organization, ISO/TC211 19109:2015, Geneva, Switzerland, 91pp.
24. ISO (2015c) ISO 19150-2:2015 Geographic information – Ontology – Part 2: Rules for developing ontologies in the Web Ontology Language (OWL). International Standards Organization, ISO/TC211 19109:2015, Geneva, Switzerland, 101 pp.
25. ISO (2016a) ISO 19110:2005 Geographic information — Methodology for feature cataloguing. International Standards Organization, ISO 19139:2007, Geneva, Switzerland, 70 pp.
26. ISO (2016b) ISO 19757-3:2016 Information technology – Document Schema Definition Languages (DSDL) – Part 3: Rule-based validation – Schematron. International Standards Organization, ISO/IEC 19757–3:2016(en), Geneva, Switzerland.
27. Lemaitre J, Hainaut J-L. Quality evaluation and improvement framework for database schemas – using defect taxonomies. In: Mouratidis H, Rolland C, editors. CAiSE 2011, LNCS, vol. 6741. Berlin: Springer-Verlag; 2011. 536–50.
28. Mueller M, Pross B (2015) OGC WPS2 Interface Standard Corrigendum 1. Open Geospatial Consortium Standard, v2.0.1. http://www.opengis.net/doc/IS/wps/2.0. Accessed 07 Aug 2018.
29. NADM (2004) NADM Conceptual Model 1.0, A Conceptual Model for Geologic Map Information. U.S. Geological Survey Open-File Report 2004–1334 and Geological Survey of Canada Open File 4737. http://pubs.usgs.gov/of/2004/1334/2004-1334.pdf, Accessed 07 Aug 2018.
30. OGC (2016) Open Geospatial Consortium GWML2 physical schema. http://schemas.opengis.net/gwml. Accessed 07 Aug 2018.
31. Peterson D, Malhotra A, Shudi G, Sperberg-McQueen CM, Thompson HS (2012) W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes. 5 April 2012. https://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/.
32. Portele C. Geography markup language (GML) encoding standard. Open Geospatial Consortium Standard. 2007;07-036(v):3.2.1 437 pp.
33. Richard S, CGI Interoperability Working Group. GeoSciML – a GML application for geoscience information interchange. In: Soller D (ed) Digitial mapping techniques '06 – workshop Proceddings, U.S. Geological Survey Open-File Report. 2006;20017-1285:47–59.
34. Robin A (ed) (2011) OGC SWE Common Data Model Encoding Standard. Open Geospatial Consortium Standard 08-094r1, v2.0.0, 207 pp.
35. Rumbaugh J, Jacobson I, Booch G. The unified modeling language reference manual. 2nd ed: Addison-Wesley Professional; 2004. 752 pp.
36. Shudi G, Sperberg-McQueen CM, Thompson HS (2012) W3C XML Schema definition language (XSD) 1.1 Part 1: Structures 5 April 2012. https://www.w3.org/TR/2012/REC-xmlschema11-1-20120405/.

37. Simsion GC, Witt GC. Data Modeling Essentials. 3rd ed. San Francisco: Morgan Kaufmann; 2005. 562 pp.
38. SKOS (2009) Simple knowledge organization system reference. W3C Recommendation 18 August 2009, https://www.w3.org/TR/skos-reference/. Accessed 07 Aug 2018.
39. Taylor P (2012) OGC WaterML 2.0: Part 1-Timeseries. Open Geospatial Consortium Implementation Standard 10-126r3, 149 pp.
40. Tomkins J, Lowe D (2016) TimeseriesML 1.0 – XML Encoding of the Timeseries Profile of Observations and Measurements. Open Geospatial Consortium Standard 15-042r3, v0.9.9, 61 pp.
41. Vretanos PA (ed) (2014) OGC Web Feature Service 2.0 Interface Standard – With Corrigendum. Open Geospatial Consortium Standard, 09-025r2, version 2.0.2, 254 pp.
42. Zachman J. A framework for information systems architecture. IBM Syst J. 1987;26(3):276–92.