Integrated Groundwater Data Management

26

Peter Fitch, Boyan Brodaric, Matt Stenson, and Nate Booth

Abstract

The goal of a data manager is to ensure that data is safely stored, adequately described, discoverable and easily accessible. However, to keep pace with the evolution of groundwater studies in the last decade, the associated data and data management requirements have changed significantly. In particular, there is a growing recognition that management questions cannot be adequately answered by single discipline studies. This has led a push towards the paradigm of integrated modeling, where diverse parts of the hydrological cycle and its human connections are included. This chapter describes groundwater data management practices, and reviews the current state of the art with enterprise groundwater database management systems. It also includes discussion on commonly used data management models, detailing typical data management lifecycles. We discuss the growing use of web services and open standards such as GWML and WaterML2.0 to exchange groundwater information and knowledge, and the need for national data networks. We also discuss cross-jurisdictional interoperability issues, based on our experience sharing groundwater data across the US/Canadian border. Lastly, we present some future trends relating to groundwater data management.

P. Fitch (✉) • M. Stenson
Commonwealth Scientific Industrial Research Organisation (CSIRO), Canberra, ACT, Australia
e-mail: Peter.fitch@csiro.au

B. Brodaric
Geological Survey of Canada, Calgary, AB, Canada

N. Booth
United States Geological Survey (USGS), Reston, VA, USA

A.J. Jakeman et al. (eds.), *Integrated Groundwater Management*,
DOI 10.1007/978-3-319-23576-9_26

26.1 Introduction

There is a growing recognition that many environmental/hydrological management questions cannot be adequately answered by single discipline studies. This has led a push towards a systems view (Chap. 24), which includes integrating many aspects of the hydrological cycle (Chaps. 1 and 3). The push for integration has significant implications for data management. It requires that data are not only well stored, but also well described, easily discoverable and accessible, and in consistent form for use in the different models in an integrated modeling system. The development of the proto-operational Australian Water Resource Assessments (AWRA) (Van Dijk et al. 2011) system in Australia and a similar system under development by the USGS (Alley et al. 2013) are good examples of this, along with many other studies reported in the literature (Schou et al. 2000; Croke et al. 2006; Krol et al. 2006).

In addition to the focus on integration, new technologies in monitoring and computing, such as advances in computational power and storage, have allowed for an increase in the complexity of studies undertaken. For example, groundwater modeling is increasingly being undertaken at larger scales and groundwater flow is being incorporated into earth system modeling – fully coupled biogeochemical climate models – reflecting the growing awareness of the importance of groundwater systems to society. Therefore, there is a growing need to share data across different jurisdictional and groundwater management areas.

All of these factors mean that groundwater data management, and its support of groundwater modeling, is changing rapidly. It is shifting from discrete standalone data management processes and systems, to connected open and shared data systems that support integrated modeling and decision support (Chap. 25). The chapter is organized as follows: first the concepts of data management are discussed, and then current practices with existing toolsets. This is followed up with case studies and last is some discussion on future directions and trends.

This chapter is not directed at organizations that are responsible for data management; rather it aims to inform the research practitioner who is responsible for an integrated modeling study.

26.2 Data Management Lifecycle

26.2.1 What Is Data Management?

Data management means different things to different practitioners, and often the varying views reflect the differing roles of the actors in the system. The World Meteorological Organization (WMO) Guide to Hydrological Practices (WMO 2008) provides the following definition:

> We define data management as the set of processes or procedures together with a defined workflow and tools, roles and governance arrangements to ensure secure storage ease of discovery and access as well as ensuring the quality and integrity of the data. These data processes and workflows tend to be formally represented in data management models of which there are many examples. In addition, the implementation of a data management model is with a data management plan.

This definition provides the context for following discussion on groundwater data management.

26.2.2 Data Management Models

The task for a data management model is to define the data management workflow and process. It does not necessarily define the governance, nor does it specify how things are to be done. These models are typically defined using graphical representation or formal modeling notation such as Business Process Modeling Notation[1] (BPMN). Here we present two data management models.

The first data management model is presented below in Fig. 26.1, and comes from the WMO Guide to Hydrological Practices (WMO 2008). This model describes a data management scheme where the roles, tools, processes and data products are defined in an abstract manner. This model has been subject to significant input from many practitioners, and is useful as a high-level framework for applications such as integrated groundwater modeling studies. The workflow is described by following the sequence of processes from top to bottom, with the tools used for each of the process connected by dashed lines, and the actors performing particular roles are associated with the tools. In the last column, a range of data inputs and outputs are identified.

The second model is illustrated in Fig. 26.2 using BPMN notation. It is taken from the Data Documentation Initiative (Thomas et al. 2009), which defines a combined cycle including data management processes as well as the associated workflow.

The workflow flows from left to right commencing at the "Start" symbol. Each of the rectangular boxes defines a process and the arrows represent transitions through the workflow from one process to the next.

[1] www.bpmn.org.

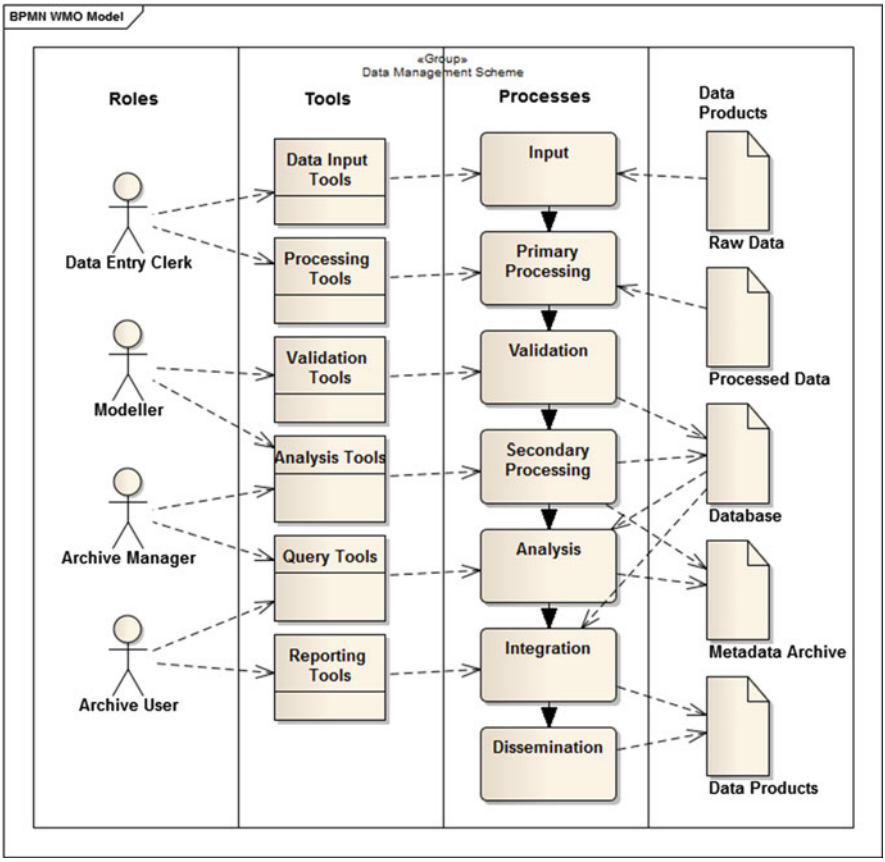

Fig. 26.1 WMO data management scheme

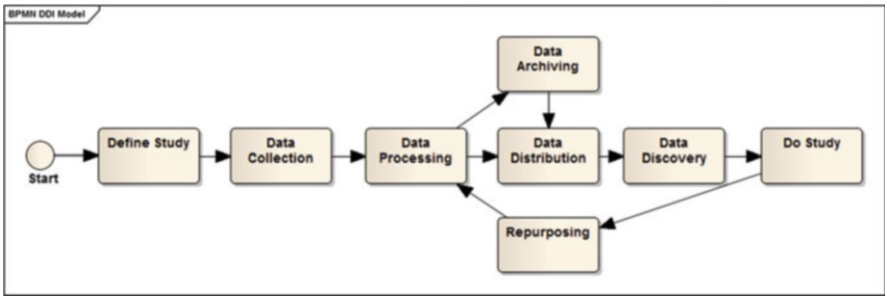

Fig. 26.2 DDI data lifecycle model

This model can be applied to integrated groundwater studies as follows:

Define Study For collection of integrated data, the first goal is to define study objectives, the models to be integrated, and the associated data requirements.

Data Collection The next process involves collection of all the data for the integrated study.

Data Processing In this step, the data is preprocessed into appropriate resolutions and formats such that it is suitable for the integrated models. Typically at this stage, a number of quality assurance and checks are undertaken.

Data Archiving Next, the data is archived in preparation for further distribution and use.

Data Distribution Prior to the study being undertaken, the data are made available through a distribution mechanism. This is very consistent with enterprise data management models where centralised data storage is used, either by way of databases or fileservers. These data stores are then accessed for the study by way of a data discovery process. More contemporary methods of data distribution using web services are now gaining favor.

Data Discovery In this step, the data are located for the groundwater study.

Do Study This is the step in the model where the study is performed. Note groundwater studies, especially modeling studies, almost always are iterative, and this iteration is reflected in the subsequent repurposing of the data.

Repurposing The final step in this workflow, takes the data generated by the groundwater study and repurposes it for another use. This could either be another integrated study, or simply another iteration within the current study.

It is worth noting that this data management model can be modified depending upon the purpose of the study and is provided as a general-purpose model. For example an additional feedback loop can be drawn between 'Do Study' and 'Data Collection' if during the study additional data needs have been identified.

26.2.3 The Data management Challenge

Data management is successful when data are discoverable, available, accessible, understandable, and usable (Robbins 2012). This perspective comes from the ecological community and their long-term ecological research (LTER) program. It recognizes that successful studies depend on the development of integrated databases and data sets, many of which are collected by different teams over

different timescales and are required to be brought together to tackle integrated scientific challenges (Costello 2009), such as integrated groundwater modeling studies. However, while management of data is a core part of the mission of large organizations such as USGS and Bureau of Meteorology in Australia, it is often the case that even within these organizations it is difficult to establish good data management practices in research projects.

Data management is beset with multifaceted problems characterized by social, cultural, and technical dimensions. The social and cultural issues associated with data management are often overlooked and can often be the reason why organizations, research project teams, and individuals, struggle with it.

Leadership heavily influences the culture of an organization, by modeling and defining behavior and values. This is particularly evident in many research projects and integrated modeling studies. It therefore follows that perhaps the most important single driver for good data management within an organization, project or study is the priority placed on it by leadership. This begins with individual practitioners recognizing the value of data, and its management, and cascades to project leaders and senior managers, who include and enforce data management in project plans through policies and adequate resourcing (Costello 2009). Efforts in this area are also augmented by leadership from national agencies such as the US National Science Foundation (NSF) and UK National Environment Research Council (NERC), which now require a data management plan to be prepared with all research funding applications.

26.2.4 Data Quality Assurance and Quality Control

The concepts of data Quality Assurance (QA) and Quality Control (QC) are profoundly critical any study. This topic is mentioned here because of its importance, but the reader is referred to WMO 2008 for a detailed treatment of the practical issues and approaches to ensuring QA/QC of hydrological data. In this section we will provide definitions of QA and QC, illustrating the differences, which are not always well understood.

QC is defined as a procedure or set of procedures intended to ensure that data adheres to a defined set of quality criteria, typically accuracy and reliability. These checks are usually done post data acquisition. QA is a more systematic approach to ensuring that the data will meet quality requirements, typically undertaken prior to data acquisition. To illustrate these differences, we will use a manufacturing example. Say a plastic part is manufactured with specific dimensions and tolerance of 10 mm square plus or minus 0.1 mm. A quality control is to check these dimensions with a micrometer to confirm that the part meets specification. In this case the dimension and tolerances are the quality criteria. For data quality control, checks could include bounds checking (not exceeding known maximum or minimum criteria) and that it conforms to some expected distribution and so on.

QA is defined as a procedure or a systematic set of procedures intended to ensure quality controlled data. These are procedures undertaken before data acquisition,

intended to improve/ensure quality once checked for. In our manufacturing example, these might include regular maintenance of the machine that manufactures the part, training for the operator, etc. Examples of this for data measurement systems can include instrument calibration procedures, operator training and so on.

QA and QC are usually bundled together as QA/QC without a good understanding of the differences and are commonly now tackled together by organisations implementing a quality management framework such as ISO 9001.[2]

For more information, the reader is directed to WMO (2008, Chap. 9) for details on data processing and quality control.

26.2.5 Data Licensing

There is a growing push towards the idea of open data across the research and government sectors, particularly for data supported by publically funded programs. Opendefinition.org provides the following definition: "a piece of data or content is open if anyone is free to use, reuse, and redistribute it – subject only, at most to the requirement to attribute and/or share – alike." Examples of the growing interest in open data are the open data agendas of the United States, Canada, United Kingdom and Australia. These are manifest in data discovery and access portals such as data.gov, data.gov.au, and others. Many of these data initiatives use open data licensing such as Open Data Commons (opendatacommons.org) and Creative Commons (creativecommons.org.au). The intent of all of these open license formats is to maintain copyright with the data creator, ensure attribution, and to transfer risk of use to the user. The interest in Opendata is driven by the assumption that making data freely available generates greater value to society. The authors of this chapter subscribe to this view.

Much data used in integrated studies are subject to a restrictive data license. This is particularly the case in environmental studies where there has been significant cost to collect hydrogeological data, lithological data, and so on. There are potentially other concerns that may limit availability such as commercial interests (eg. storage levels within a hydro-electricity scheme) or potential security concerns. In our work with large scale integrated surface and groundwater modeling, the majority of data have come from state jurisdictions and water management authorities, and is subject to strict licensing conditions. It is often the case for the data to be licensed for a particular study, and in some cases with conditions stipulating deletion once the study is complete (Hartcher and Lemon 2008). Any data management initiative thus needs to be fully cognizant of the many and varied and often strict data licensing requirements.

[2] http://www.iso.org/iso/iso_9000.

26.2.6 Data Management and Analysis Tools

Integrated groundwater studies have a specific set of requirements for data types and their specific data management needs. For integrated groundwater modeling studies, these are well described by Refsgaard et al. (2010). Typical data include borehole data containing general descriptions, location, lithology, borehole geophysics, water level and water chemistry. This is supplemented with surface geophysical data, which might include seismic, electromagnetic and electrical data from which the hydrogeology and conceptual models of the groundwater systems can be developed. Most groundwater data management systems have separate tools, processes, and mechanisms for storage of time series, GIS, and spatial data, metadata, and conceptual models.

26.3 Time Series Data Management

There exist many commercial time-series data management systems, which specialize in the storage, dissemination and management of surface and groundwater data (e.g. WISKI,[3] Schlumberger[4] and Aquatic Informatics[5]). These types of software packages typically allow ingestion of a variety of data sources including telemetry from automated gages, perform quality assurance, and usually are coupled to integrated analysis tools. They are also able to store a broad set of other hydrological, meteorological and climate data. Most of these systems use relational database technology as the persistence mechanism, which is then attached to a series of tools, as can be seen in the abstract model of a timeseries data management system in Fig. 26.3 below. In this diagram, we map the functional elements described by WMO in Fig. 26.1 above to this abstract model. For these systems, the data output toolsets are increasingly being used to deliver data outside the enterprise using web services and open standards such as WaterML2.0 (Taylor et al. 2013).

This ability to deliver data outside the enterprise becomes very useful for integrated studies and allows time series systems to become part of a web-based data network, which is discussed further below in web-based data management and modeling section.

[3] http://www.kisters.eu/english/html/homepage.html.

[4] http://www.slb.com/services/software.aspx.

[5] http://aquaticinformatics.com.

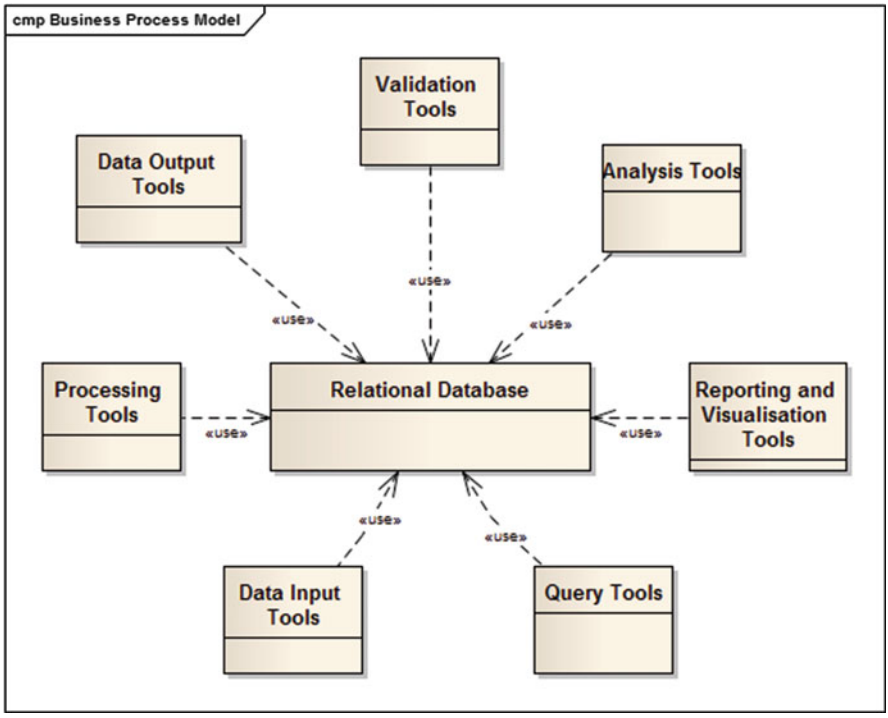

Fig. 26.3 Abstract model of a time series data management system

26.4 GIS toolsets

GIS systems are a core tool for integrated environmental modeling and are widely used (Argent 2003; Gogu et al. 2001; Whiteaker et al. 2006). GIS toolsets are used for spatial and temporal data management, spatial data-processing and analysis, and they can form a software framework for integrated modeling scenarios (Ames et al. 2012).

In Fig. 26.4 above, Argent (2003) describes how GIS systems can be used for integrated modelling application. Two workflows are described, one simply uses GIS for spatial data management (diagram on the right) and the other (on the left) describes a more integrated use of GIS toolsets. In this workflow, the GIS becomes the integration tool, where various modeling applications are created and run. For a good example of this type of workflow, see Gogu et al. (2001).

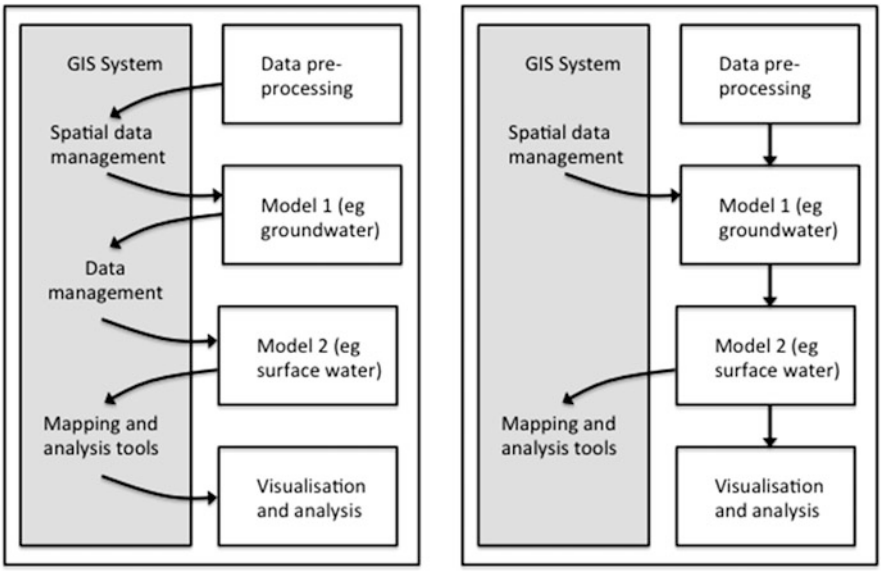

Fig. 26.4 GIS workflow for integrated modelling after Argent (2003)

26.5 Examples of GIS Data models

The widespread use of GIS systems as a data management and data integration tool has led to the development of domain specific geospatial databases, called GeoDatabases. These are optimized for the sorts of data commonly used in geospatial studies, in this case with integrated groundwater studies. These Geodatabase models (Strassberg et al 2004; Jarar Oulidi et al 2009; Chesnaux et al 2011; Yang et al 2010b) represent the features and properties of hydro-geological systems, in ways that allow storage, integration and manipulation of the spatial and time series data. In the hydrology domain, the two most widely used models are ARCHydro (Maidment 2002) for surface water studies, and ARCHydro-GW (Strassberg et al. 2004) for groundwater studies.

ARCHydro is a geographical data model for hydrological systems designed to support a cartographic representation of hydrological features. It is designed to provides a unified model for geospatial and time series data in support of integrated hydrological modeling and analysis (Strassberg et al. 2004). It allows different aspects of the water-resource systems, such as a drainage system, hydro-network and channel system, to be linked to time series flow observations and managed within the GIS system.

ARCHydroGW provides a data model for hydrogeologic units, boreholes and other aspects of groundwater systems that can be used for integrated modelling.

There are many studies which have successfully used these types of models (Whiteaker et al. 2006) in conjunction with GIS toolsets.

One issue that arises concerns unique identifiers in these types of systems (called HydroID in ARCHydro-GW), which identify features in the geospatial databases. Usually these identifiers have local scope, meaning that they are assigned to be unique within a GeoDatabase, and are most usually non-unique when combining or integrating databases. As a result, it becomes difficult to automatically merge databases when conducting integrated studies, requiring significant effort to match or differentiate hydro-geological features based other information.

Another issue concerns the assignment of a fixed geometry to a feature type. For example, a borehole might be represented by a point, in one particular GeoDatabase, and by a line in another GeoDatabase. Thus integrating the different representations between GeoDatabases becomes problematic. This has led to the development of the Hy-Features (Atkinson et al. 2012) conceptual model, in which the features are defined independently of representation. The difference may seem to be esoteric, but defining features in this way allows for easier integration of data for a particular feature type, and greatly eases integrated studies.

26.6 Metadata Requirements

For the integrated modeler, the discovery of data suitable for modeling studies always depends on the availability of suitable metadata and an ability to search across it. Most organizations with data management programs will have metadata standards or profiles defined. Examples include the Australian and New Zealand Land Information Council (ANZLIC) in Australia, and the Federal Geographic Data Committee (FGDC) in the US. In general, there is a significant international adoption of the ISO/TC211[6] standards, and many of the emerging national metadata standards are now using ISO as a core, with profiles or extensions as required. Because of this standardization, many tools are appearing which support these standards and leverage them to allow federated searching capabilities. Examples of these include GeoNetwork (http://geonetwork-opensource.org), GI-Cat (http://essi-lab.eu/do/view/GIcat), and Esri Geoportal (http://www.esri.com/software/arcgis/geoportal). In all of these examples, the tools support a number of different metadata profiles and have the ability to harvest metadata records from other catalogs. This federated search ability distributes the responsibility and burden for the generation and management of metadata to data providers, and then allows federated catalogs to be easily assembled and queried by users.

[6] http://www.isotc211.org/.

26.7 Conceptual Models

In hydrological modeling the need for a scientific conceptual model is well known (Refsgaard et al. 2010). Though related, scientific conceptual models are distinguished from information conceptual models (discussed in semantics below). Information conceptual models consist of theoretical knowledge (consistent with the scientific conceptual model), such as feature types and scientific theories, whereas scientific conceptual models are essentially re-constructions of a physical area and consist of representations of actual features. Scientific conceptual models provide a description of the agreed understanding of the system under study. Refsgaard et al. (2010) argue for a scientific conceptual model repository to help combine knowledge effectively. We argue that defining both scientific and information conceptual models, and having them discoverable and readily available, is a key requirement for integrated studies.

26.8 Web-Based Data Management and Modeling

Integrated studies by their very nature have significant data management and integration challenges. When coupled with the rapidly growing data holdings (for example, in national agencies), an environment is created where discovery access and use of data becomes increasingly difficult. As a result, an interest in interoperability has grown, and practitioners are increasingly looking to the web for help in data management and modeling, such that web-based data access and management is now common place (Granell et al. 2009; Frehner and Brändli 2006). Much of the recent advances in this area have been precipitated by the more than a decade's interest in Spatial Data Infrastructures (SDI; Masser 2010), which has directly led to the development of pan-national standards such as INSPIRE in Europe (http://inspire.jrc.ec.europa.eu), and the construction of associated data networks, including those for hydrology and hydrogeology. In this model of data management, organizations are responsible for management of data and making it discoverable, accessible and available by way of a data network. This approach has significant benefits for integrated studies.

In the next section, we discuss challenges and approaches to building and coupling groundwater data networks, and describe several examples: one example from Canada, two from the US, a unified Canada-US example, and a US example from academia.

26.9 Groundwater Data Networks

Groundwater data networks are becoming an important source of data for groundwater studies, due to the increased breadth and depth of their data holdings (Refsgaard et al. 2010). In data networks, autonomous data sources are federated into a composite entity, which behaves as a unified single enterprise. For example, regional groundwater monitoring networks, water well databases, aquifer maps, and other relevant data, are being variously integrated into larger networks in Australia, Canada, and the US (Booth et al. 2011; Brodaric et al. 2011; Dahlhaus et al. 2012). Such networks are typically arranged in some form of distributed architecture, which dynamically retrieves data from original sources, thus ensuring access to current data. They also typically enable users to query and obtain data via a unified common view, shielding users from the heterogeneity of the original sources. In this way, more data, and more data types, are more readily accessed by those studying groundwater, including modelers.

26.10 Challenges: Data Interoperability in Groundwater Data Networks

Data access is a key issue faced by all groundwater data users, including modelers, particularly those carrying out integrated studies using multiple data sources. Barriers to data access involve data availability, fragmentation, and heterogeneity: i.e. not all data are available online, and groundwater data are divided unevenly amongst multiple providers, such that the structure and content of the data is quite heterogeneous. This leads to problems in its usage, because data are hard to find, and once found are difficult to exploit due to the immense work required to re-format the data into a common usable structure. Figure 26.5 illustrates an example of heterogeneity in the lithology descriptions of water well databases from two adjacent Canadian provinces: note the differences in language (French/English), structure (one field/many fields), and content (sand/fine and medium sand).

Overcoming the data access barrier thus requires a solution to the alignment of multiple heterogeneous and distributed data sources, i.e. to the data interoperability problem. Spatial Data Infrastructures (SDI) are a leading approach to this problem, and they are actively being adopted by various water data networks, including those for groundwater. Solutions to data interoperability typically require alignment of the data at five levels: systems, syntax, structure, semantics and pragmatics (Brodaric 2007). Ideally, SDI standards are used at each level, and in the water domain these are being developed in coordination with the Open Geospatial Consortium (OGC), the International Organization for Standardization (ISO), and professional bodies such as the World Meteorological Organization (WMO) (Zaslavsky et al. 2011):

	cle_noseq integer	epaisseur double precis	matprim character var	fiss_prim character var	mat_sec character var	fiss_sec character var
1	1	1.5	SABL/BLO	INCO		INCO
2	1	3.4	SABL/BLO	INCO		INCO
3	1	3.4	ARGL/GRA	INCO		INCO
4	1	2.7	SABL/GRA	INCO		INCO
5	1	0.3	TERR	INCO		INCO

	materialcolor character var	material_1 character var	material_2 character var	material_3 character var	topdepth real	bottomdepth real
1		Topsoil			0	0.3048
2	Black	Muck			0.3048	0.9144
3		Medium Sand			0.9144	1.524
4		Fine Sand	Silt	Clay	1.524	7.3152

Fig. 26.5 Heterogeneous water well data from the Canadian Groundwater Information Network (www.gw-info.net)

- **The systems level** involves the deployment of standard web interfaces to the data, typically web services such as WFS (Web Feature Service), SOS (Sensor Observation Service), and WMS (Web Map Service), which transmit features (e.g. wells), observations (e.g. groundwater levels), and map images, respectively (Boring et al. 2012; De La Beaujardière 2006; Panagiotis 2005).
- **The syntax level** involves the use of standard data languages, such as GML (Geographical MarkUp Language; Portele 2007), which can be used to encode data.
- **The structure level** includes standard data schema, such as OGC Observations and Measurements (O&M), WaterML2 (WML2), and GroundwaterML (GWML), which are built with GML and constitute a common structure for observations, water time series, and groundwater features, respectively (Boisvert and Brodaric 2012; Cox 2011; Taylor et al. 2013). Standard schemas are typically diagrammed using well-constrained methods, such as UML, and can be expressed in a variety of formats, such as XML.
- **The semantics level** refers to the use of standard concepts and related terms. The terms are typically organized in vocabularies or codelists, and the concepts are typically organized in computational ontologies. Both can be applied to (1) data content, such as common rock type terms and their definitions, and (2) data structure, such as a commonly defined lithology field containing rock type terms. However, they can also refer to scientific knowledge in general, distinct from data, that is to the components of a scientific conceptual model. This includes definitions for the types of entities in the model, and expressions of underlying theories that drive the model.
- **The pragmatics level** includes standard tools and methods, so that data are collected and processed using common scientific protocols.

As an example, the heterogeneous rock type descriptions from Fig. 26.5 can be resolved via transformations of the data at each level: a query in a web browser, for example wells possessing certain rock types, is translated into requests to WFS web services layered over each database (systems); the web services return water well records, by transforming the structure of the databases into standard GWML (syntax, schema), which uses one field to hold rock types, and the content of this field is populated with the rock types in the logs transformed into a standard English vocabulary (semantics). Community agreed protocols are used to determine how rock type terms correlate between the source data and the standard vocabulary (pragmatics). Finally, the results from each web service are integrated, producing a single unified GWML file that is returned to the modeler.

Note that data networks can vary according to where the transformations occur, for example locally at the source, or centrally, and some networks utilize a hybrid strategy that includes local transformations for some network nodes and centralized transformations for the remainder. Likewise, the degree of data centralization can also vary, as evident by the rise of hybrid approaches that use frequently updated central data caches as access points for some, but not all, of the data in a network. Lastly, the location of catalogs can also be centralized, distributed or hybrid; catalogs contain metadata that enable data to be found in the network and that facilitate data transformations, for example by serving local and standard vocabularies and ontologies. However, regardless of the architectural placement of these items within a network, data interoperability cannot be fully achieved without alignment at each of the five levels.

26.11 Examples

This section presents five examples. Example 26.1 is the Canadian Groundwater Information Network and Example 26.2 the US National GroundWater Monitoring Network. These are presented as examples of the trend towards large scale national groundwater data networks. Example 26.3 details an emergent North American Groundwater Data Network and discusses how individual networks, if constructed the right way, can be federated into a single federated groundwater data network. Example 26.4 is that of an academic surface water hydrological data network. Lastly, Example 26.5 discusses the use of integrated hydrological data provided from data networks in a national water assessment system. These five examples illustrate approaches that variously utilize hybrid methods for the placement of data, transformations, and related data catalogs.

Example 26.1: Canadian Groundwater Information Network

The Canadian Groundwater Information Network (GIN; Brodaric et al. 2011) is a national federation of groundwater data sources managed by Canadian provinces and some federal departments. At present, it contains water well records for most of Canada, monitoring records (groundwater levels) for some selected provinces, and some key regional aquifer and geology maps. As shown in Fig. 26.6, GIN is an example of an architecture in which a centralized approach is used for data transformation and catalogs, and a hybrid approach is used for data placement, that is it is a mix of centralized data caches and distributed data sources such that some data are obtained from the centralized caches and others directly from the distributed data sources.

GIN consists of three tiers. The bottom tier comprises provincial and federal data sources, exposed online ideally via standard web services and data exchange formats, or occasionally via bulk file downloads in non-standard local formats. The top tier consists of potentially many distributed web portals that provide various user interfaces to the data – included among these is the GIN portal itself (www.gw-info.net). The middle tier connects the top and bottom tiers, in that it (1) carries out the necessary transformations between these tiers, and (2) houses the data caches and catalogs required by the transformations. The data caches and

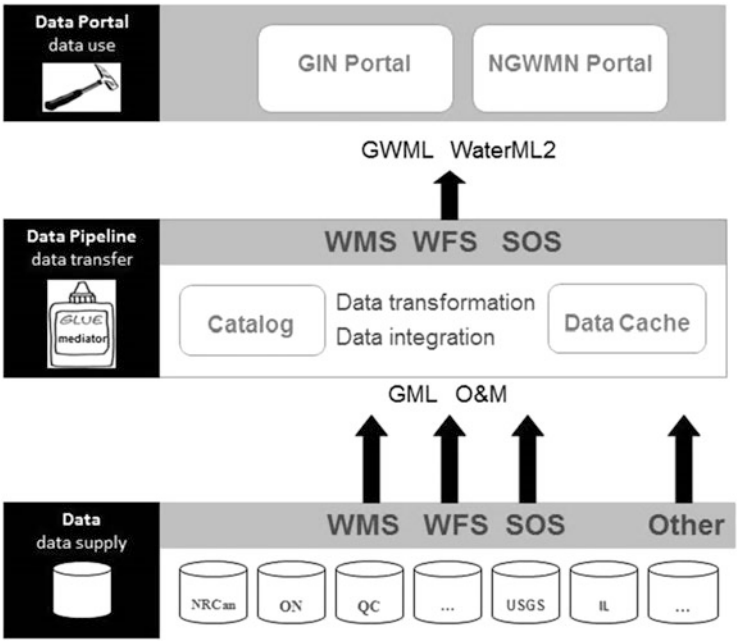

Fig. 26.6 Architecture for GIN and NGWMN – local data sources in the lowest tier, central data caches, catalogs, and transformations in the middle tier, and distributed web portals in the upper tier

catalogs are updated from local sources either dynamically online via the web services, or manually via file download. The transformations occur in both directions as the middle tier transforms requests from the portals to the local requirements of individual web services or data caches, and conversely transforms the retrieved data to a community standard, either GWML or WaterML2, as required. It also integrates the standardized data, retrieved from potentially multiple sources, into a single unified result, and returns this result to the requester in a choice of several possible file formats such as GML, KML, shape file, ESRI GeoDatabase, or PDF. Significantly, the middle tier is presented online as three web services (WFS, WMS, SOS), which effectively serve as a central data pipeline. Requests for data can thus be made in two ways: through a web portal which issues requests to the data pipeline; or the web portal can by bypassed completely and requests can be sent directly to the data pipeline, for example from an online modeling application.

The GIN architecture has proven to be efficient and effective, returning moderate amounts of data relatively quickly (e.g. hundreds of wells in several seconds), which is adequate for typical usage. Retrieval of large data amounts is enabled via bulk download of pre-packaged files.

Example 26.2: US National GroundWater Monitoring Network

The US National Groundwater Monitoring Network (NGWMN; ACWI, 2013) is a recently initiated national federation of US groundwater data. In collaboration with groundwater agencies from US states, the NGWMN links federal and state data in a virtual environment, providing a single online entry point to groundwater data holdings across the nation. NGWMN data include water-well records, water level and water-quality measurements, and references to related aquifers where possible. The NGWMN architecture is very similar to GIN's (Fig. 26.6), utilizing a three-tier portal-pipeline-data architecture, as well as centralized data transformations and catalogs. However, NGWMN differs from GIN in the extent of its data cache, as NGWM caches all data to improve speed of online usage: a data request to NGWMN will thus always retrieve data from its central cache and never directly from the original data sources. The middle tier pipeline implements the same standards as GIN, i.e. GWML, WaterML2, WFS, SOS, and WMS, and also similarly the harvester that populates the cache from local data sources uses these as well as other local standards to ensure that barriers to participation are low. At present NGWMN has completed a pilot stage and adoption continues, incorporating data from more than 20 states and enabling access to this data via an online portal (http://cida.usgs.gov/ngwmn).

Example 26.3: An Emergent North American Groundwater Data Network

Coupling of the Canadian and US groundwater data networks is highly desirable, due to the potential for high impact on cross-border groundwater studies. Encouragingly, the coupling of technologies is relatively straightforward, due to the implementation of compatible architectures, and the adherence to common standards across the bottom three interoperability levels (i.e. systems, syntax, and schema), which ensure the use of common web services and related schema. Note that discrepancies at the remaining levels (semantics, pragmatics), which involve differences between vocabularies largely caused by variations in data collection procedures, are managed through data transformations. This is feasible because each network exposes a single data pipeline, which is treated as just another data source by the consuming network. For example, NGWMN is consumed by GIN as if it were another provincial data source, one that requires mapping of vocabularies only, with that mapping taking into account procedural differences.

The coupling of the GIN and NGWMN networks has been tested in two pilot studies carried out in the course of standards development activities at the OGC. In the Groundwater Interoperability Experiment (GWIE; Brodaric and Booth 2010), water level time-series and associated wells across the US-Canada border were found, viewed and downloaded. The Climatology-Hydrology Information Sharing Project (CHISP; Brodaric et al. 2013) was more ambitious, as it involved both surface water and groundwater monitoring gauges, and addressed both water quantity and quality concerns. CHISP enabled cross-border flood risk determination and alerting through dynamic monitoring of gauges upstream from a point of interest, and it also dynamically estimated nutrient loads for any one of the mutually managed Great Lakes.

The GWIE and CHISP studies not only demonstrated that the two groundwater data networks can be successfully coupled, they also directly led to improvements in the networks and to the identification of gaps in the standards, which are subsequently being addressed. Also significantly, they showed that key organizational mandates could be enhanced through the deployment of open standards and the resultant interoperability of the data networks. The end result is the nascent emergence of a North American groundwater data network, which is facilitating access to data for modelers and others in both countries.

Example 26.4: CUAHSI-HIS and HydroDesktop

The Consortium of Universities for the Advancement of Hydrological Science (CUAHSI) is a research collaboration of more than 100 US universities and affiliated international research organizations. Apart from its significant scientific contributions, a key achievement of CUAHSI is its hydrological information system (HIS), which enables researchers to publish, manage, and use largely surface water data online (Tarboton et al. 2011). The published data are integrated into the wider HIS data network, which links academic data with

major government data sources, such as the USGS, EPA and NOAA. HIS is by far the most de-centralized architecture examined herein, as its data holdings, transformations, catalog and portals are all distributed. Data distribution is achieved, at the moment, using custom "WaterOneFlow" web-services layered over 70 data sources. Data transformation takes place at each data source as an integral component of the web services, and is minimized as standard database structures are encouraged. For data discovery, transformation includes the semantic level, as time series parameters are mapped to a common vocabulary, enabling specific types of data to be identified within the network. However, data retrieval occurs only up to the structure level, as parameters are not mapped to a standard, but served 'as is' from the sources; moreover, data from multiple sources are not integrated into a unified file, but served individually. A central catalog tracks and publishes metadata about the data sources, which can be discovered by online tools. However, in contrast to previous data networks described herein, which are web-centric, HIS emphasizes desktop tools as primary interfaces to the data network. The cornerstone is HydroDesktop, which contains a rich suite of functions for data discovery, management, analysis and modeling. At present, plans are in place to develop HydroShare which will be an online portal that not only incorporates some key HydroDesktop functionality, but will in addition enable many types of collaborative online interactions, most notably the sharing of data and models amongst various research teams (Tarboton 2013).

Example 26.5: Australian National Water Resource Assessment System[7]

Following a period of extended drought within Australia the federal government initiated a national plan for water security, enacted as legislation through the Water Act of 2007.[8] An outcome of the Water Act was that the Australian Bureau of Meteorology (BoM) would become the custodian of national water information, and would be required to produce several new water information products, including the annual National Water Accounts and sub-annual National Water Resources Assessments. The AWRA integrated modelling system was developed to support the production of these continental-scale products and integrates three models – landscape processes (AWRA-L), groundwater (AWRA-G) and surface water routing and use (AWRA-R for rivers)

In the proto-operational version of AWRA, where possible, data fetching, pre-processing and loading of input data streams are treated as independent processes, decoupling the modelling system from the data and data management systems. In a complex modelling system such as AWRA, there are many input

[7] Note this section refers to the proto-operational development of AWRA, the final operational version my change in design, scope and implementation.

[8] http://www.environment.gov.au/topics/water/australian-government-water-leadership/water-legislation/key-features-water-act-2007.

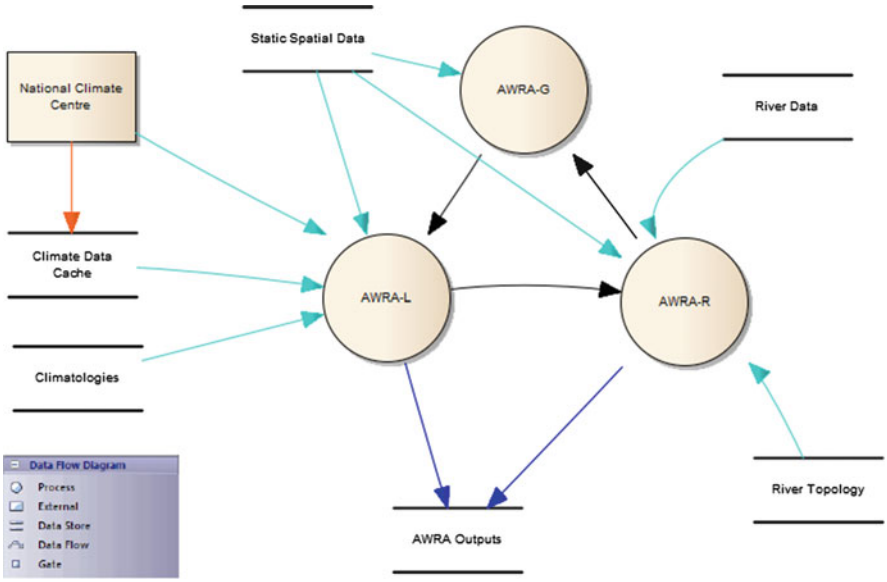

Fig. 26.7 High level representation of data flows within the AWRA system. Note the barred data sources are internal ad hoc rather than operational data sources. *Orange arrows* are ASCII grids via FTP delivery, *teal arrows* are binary files via direct transfer, *blue arrows* are NetCDF export to THREDDS server and *black arrows* are PI-XML via Delft-FEWS internal data store

data streams, some are standard products and use standardized formats and associated metadata; and they are often supported by a government mandate or service level agreement. These can be considered high trust data streams and have guaranteed availability, and are used in preference to alternatives.

In a real-time modelling system such as AWRA the data fetching is done asynchronously, to both reduce wasted time in the workflow waiting for fetch and pre-processing, and to facilitate future historic runs. The data retrieval process makes use of a local file based data store (Fig. 26.7), which it keeps up to date through both checking for new data, and updating existing data as it is re-published by the data provider following re-processing such as when updated observations become available.

While the fetching of published, operational data streams is preferable from a systems perspective, often the data are incomplete and have gaps either in space or time. In AWRA these gaps are filled through purpose developed data interpolation algorithms or by lookup default values in a post-processing step.

Figure 26.7 shows a high level view of the AWRA modelling system. The diagram shows both the flow of data into and out of the system, and internally between the three major model components. In the original design of the system many of the input data streams were hosted operationally by the Bureau, supported by its new mandate as the custodian of water information. Due to the rapid development of AWRA, and the significant technical and organizational hurdles

faced by the Bureau in streamlining the data ingestion process, none of the operational data streams, apart from climate data, are currently available for real-time use by the AWRA system. This has caused complications in the management and updating of the system, and diverted development resources. Once the data network is completed, this problem will be significantly reduced.

Ideally, work on data ingestion would have involved adhering to standards such as WaterML2 (Taylor et al. 2013) for observations, and GML (Portele 2007) for spatial data such as contributing catchments and river network topology. Instead, substantially greater work has been diverted to the collection, checking, re-purposing, re-formatting and management of input data, with all the complications of storage, deployment, duplication, broken provenance chains and a greater number of potential points where errors could be introduced. Once the data services are available through the water data network, AWRA's modular design will allow migration to these new data sources with minimal disruption.

The data sources that will benefit most from availability using a data network approach are those where identity is important such as the naming of river gauges, and those that will need to be extended in their temporal coverage such as river observations. In the current conceptual design of AWRA, the location and identity of river gauges are crucial. The location is used to identify contributing flow from the AWRA-L model and is based largely on the positioning of infrastructure within the river network, rather than by river confluences, although they may be co-located. Over time, as more river reaches are added to the model, gauges are moved or retired; or as the number of gauges used in the model are consolidated, the relationship between river reach models in AWRA-R and the contributing areas used to apportion flow from AWRA-L into those reach models will need to be updated, checked, and incorporated into the model, a time consuming and error prone task. Additionally the mix of points used to define reach models is crucial in the ingestion of observational data such as flow, extractions, diversion and storages, as the identity of those points will be used to resolve the inputs. Currently the network of points, their identities and the related observational data are compiled manually, an even more costly and error prone process than the contributing areas, as the identities are often unique to the agency tasked with monitoring them. The temporal data when collected will often be in different formats that require processing and consolidation, but more crucially the semantic definition of terms is often subtly different, requiring at least a unit conversion, and at worst a conceptual transformation.

Figure 26.8 shows the future idealised data flows into and out of the AWRA system in which the two most important data streams have been replaced by operational web services. These include the network geometry and topology, and associated contributing areas via the GeoSpatial Fabric, and the temporal observation data such as gauged river flow, storage levels and diversion via the AWRIS data warehouse. Crucially, some of greatest headaches in preparing and ingesting input data for the AWRA system will be solved using this approach. The GeoFabric will provide a resolution of identity between the spatial network, the jurisdictional agencies that collect the data, and the AWRIS data warehouse. AWRIS itself will

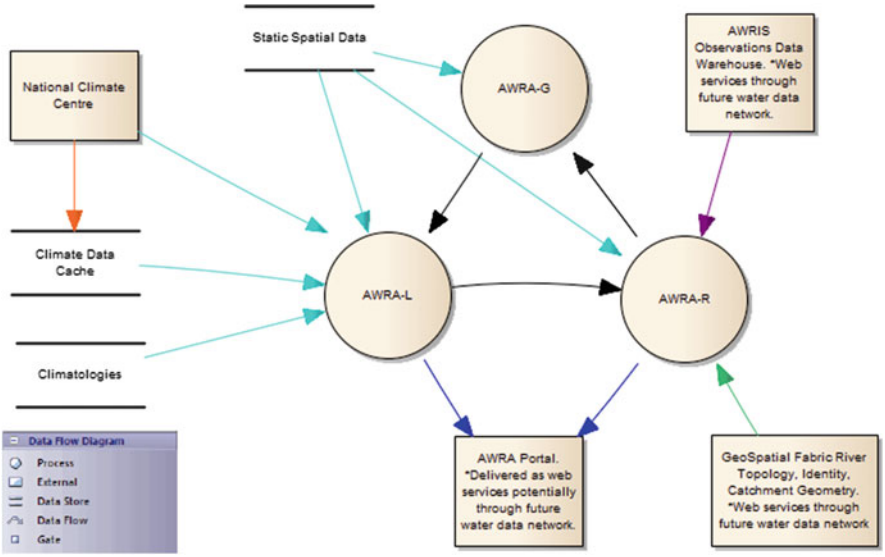

Fig. 26.8 High-level representation of future idealized data flows for the AWRA system, showing the current ad-hoc data streams replaced by operational services. Note the barred data sources are internal ad hoc, rather than operational, data sources. *Orange arrows* are ASCII grids via FTP delivery, *Teal arrows* are binary files via direct transfer, *blue arrows* are NetCDF export to THREDDS server, *green arrows* are GML via web services, *mauve arrows* are WaterML2 via web services and *black arrows* are PI-XML via Delft-FEWS internal data store

handle the ingestion, consolidation and semantic matching between the diverse sources, as well as proving a trusted data source complete with metadata, and a convenient web services interface supplying data in standardised formats such as WaterML2.

AWRA is a significant national integrated modeling application that has many data management challenges. The current system makes use of many semi-automated steps for the discovery, access, integration and use of data. We have learned that:

- Integrated modeling systems cannot be developed in isolation from the data availability and management needed to support them
- Models need to be managed and governed similarly to data
- Management of data needs to be approached from a dataset by dataset perspective
- A web-based data network would significantly ease the burden of the data management challenge for integrated modeling studies like AWRA.

26.12 Discussion of Future Trends

As noted above, it is becoming commonplace to deliver groundwater data online, typically via web services, and to incorporate such data into groundwater studies and modeling activities, also variously occurring online in workflow environments. The totality of these online resources and activities is often referred to as cyber-infrastructure. We anticipate that for integrated modelling studies the cyber-infrastructure paradigm will continue to evolve and grow, likely exponentially.

Furthermore, as cloud-computing technology is also becoming commonplace, it is likely that the processes of data storage, management and integration will occur within the "cloud" (Yang et al. 2010a). This essentially outsources the provision of the hardware side of the data management challenge, with expected gains in efficiencies, reduction of costs and potentially risks. We expect that cloud-computing technology will become an important enabler for delivery of integrated groundwater data in data networks.

Open standards (data and services) are likely to become more common-place with some good current examples being GWML, WaterML2.0 and the underlying GML and XML formats.

Finally, linked data implementations will continue to evolve and grow. Linked data is a term which refers to a set of standards and approaches for publishing and connecting data on the web (Bizer et al. 2009). Linked data is made available on the web in a standard format, usually RDF, which enables links to other datasets, or contextual data including metadata. Because linked data methods use the standard web-based linking approach of Universal Resource Identifiers (URI's), it becomes very easy to discover new data and information on the web. As a result, linked data methods are migrating from the research community and starting to become mainstream, albeit with varying levels of conformance to core linked data principles (Hogan et al. 2012). Examples are appearing in a number of countries, such as the UK location program (http://data.gov.uk/location), in which the identity of features and their corresponding properties can be easily determined.

Two related issues remain a challenge for linked data – these are particularly evident in the water domain. The first is the massive volume of data stored in legacy databases: because linked data approaches, at the moment, almost universally deploy RDF as a format, it still remains a research objective how best to layer linked data methods over non-RDF databases (Marjit et al. 2013). The second associated issue concerns granularity: what is the appropriate granule to be assigned an URI? For example, a particular measurement in a time series, the time series itself, the monitoring site, or even a specific pixel in a remote sensed image? In many of these cases the level of granularity would result in enormous and likely impractical volumes of linked entities. Thus, it becomes important to be able define a certain level of granularity, and have web-friendly mechanisms to delve deeper if required. Nonetheless, we expect that linked data approaches will continue to grow and become an integral part of data networks.

References

Advisory Committee on Water Information (2013) A national framework for ground-water monitoring in the United States, Prepared by the Subcommittee on Ground Water of the Advisory Committee on Water Information, First Release June 2009, Revised July 2013, 169 pp. http://acwi.gov/sogw/ngwmn_framework_report_july2013.pdf

Alley WM, Evenson EJ, Barber NL, Bruce BW, Dennehy KF, Freeman MC, Freeman WO, Fischer JM, Hughes WB, Kennen JG, Kiang JE, Maloney KO, Musgrove MaryLynn, Ralston B, Tessler S, Verdin JP (2013) Progress toward establishing a national assessment of water availability and use. U.S. Geological Survey Circular 1384, 34 p, available at http://pubs.usgs.gov/circ/1384

Ames DP, Horsburgh JS, Cao Y, Kadlec J, Whiteaker T, Valentine D (2012) HydroDesktop: web services-based software for hydrologic data discovery, download, visualization, and analysis. Environ Model Software 37:146–156

Argent RM (2003) An overview of model integration for environmental applications – components, frameworks and semantics. Environ Model Software 19:219–324

Atkinson R, Dornblut I, Smith D (2012) An international standard conceptual model for sharing references to hydrologic features. J Hydrol 424–425:24–36

Bizer C, Heath T, Berners-Lee T (2009) Linked data-the story so far. Int J Seman Web Infor Sys (IJSWIS) 5(3):1–22

Boisvert E, Brodaric B (2012) Ground Water Markup Language (GWML) – enabling groundwater data interoperability in spatial data infrastructures. J Hydroinf 14(1):93–107

Booth NL, Brodaric B, Lucido JM, Kuo I-L, Boisvert E, Cunningham WL (2011) Development of an interoperable groundwater data exchange network between the United States and Canada. GeoHydro 2011, Quebec, 28–31 Aug

Boring A, Stasch C, Echterhoff J (eds) (2012) OGC sensor observation service interface standard. Open Geospatial Consortium, OGC 12–006, version 2.0, 163 pp. https://portal.opengeospatial.org/files/?artifact_id=47599

Brodaric B (2007) Geo-pragmatics for the geo-spatial semantic web. Trans GIS 11(3):453–477

Brodaric B, Booth N (2010) OGC groundwater interoperability experiment final report. Open Geospatial Consortium, Groundwater Data Management, 44 pp. http://portal.opengeospatial.org/files/?artifact_id=43545&version=1

Brodaric B, Sharpe D, Boisvert E, Logan C, Russell H, Julien H, Smirnoff A, Létourneau F (2011) Groundwater information network: recent developments and future directions. Proceedings, GeoHydro 2011, Quebec, 28–31 Aug

Brodaric B, Dabolt T, Booth N, Vretanos P (2013) CHISP-1 pilot project introduces open architecture for watershed observatories. Cana Water Resour Assoc, Water News 33(1):6–12

Chesnaux R, Lambert M, Walter J, Fillastre U (2011) Building a geodatabase for mapping hydrogeological features and 3D modeling of groundwater systems: application to the Saguenay–Lac-St.-Jean region, Canada. Comput Geosci 37:1870–1882

Costello MJ (2009) Motivating online publication of data. *dx.doi.org*
Cox S (ed) (2011) Observations and measurements – XML implementaion. Open Geospatial Consortium, OGC 10025r1, version 2.0, 77 pp. http://portal.opengeospatial.org/files/?artifact_id=41510
Croke BFW et al (2006) Integrated assessment of water resources: Australian experiences. Water Resour Manag 21(1):351–373
Dahlhaus PG, MacLeod AD, and Thompson HC (2012) Federating hydrogeological data to visualise Victoria's groundwater. In: Lambert I, and Gordon AC (eds) 34th international geological congress: proceedings, 5–10 Aug 2012, Brisbane, Australian Geoscience Council, p 592
De La Beaujardière J (ed) (2006) OpenGIS web map server implementation specification. Open Geospatial Consortium, OGC 06–042, version 1.3.0. http://portal.opengeospatial.org/files/?artifact_id=14416
Frehner M, Brändli M (2006) Virtual database: spatial analysis in a web-based data management system for distributed ecological data. Environ Model Software 21(11):1544–1554
Gogu R, Carabin G, Hallet V, Peters V, Dassargues A (2001) GIS-based hydrogeological databases and groundwater modelling. Hydrogeol J 9(6):555–569
Granell C, Gould M, Manso MÁ, Bernabé MÁ (2009) Spatial data infrastructures. In: Karimi H (ed) Handbook of research on geoinformatics. Information Science Reference, Hershey, pp 36–41. doi:10.4018/978-1-59140-995-3.ch005
Hartcher M, Lemon D (2008) Data management for the Murray-Darling Basin sustainable yields project, pp 1–40. https://publications.csiro.au/rpr/download?pid=procite:51d8dcdf-203b-4cda-932e-9480ac8b2cda&dsid=DS1
Hogan A, Umbrich J, Harth A, Cyganiak R, Polleres A, Decker S (2012) An empirical survey of linked data conformance. Web Semant Sci Serv Agents World Wide Web 14:14–44
Jarar Oulidi H, Löwner R, Benaabidate L, Wächter J (2009) HydrIS: an open source GIS decision support system for groundwater management (Morocco). Geo-Spat Inf Sci 12(3):212–216. http://doi.org/10.1007/s11806-009-0048-9
Krol M, Jaeger A, Bronstert A, Güntner A (2006) Integrated modelling of climate, water, soil, agricultural and socio-economic processes: a general introduction of the methodology and some exemplary results from the semi-arid north-east of Brazil. J Hydrol 328(3):417–431
Maidment DR (ed) (2002) Arc Hydro: GIS for water resources. ESRI Press, Redlands
Marjit U, Sharma K, Sarkar A, Krishnamurthy M (2013) Publishing legacy data as linked data: a state of the art survey. Library Hi Tech 31(3)
Masser I (2010) Building European spatial data infrastructures, 2nd edn. ESRI Press, Redlands, 108 pp
Panagiotis AV (ed) (2005) Web feature service implementation specification. Open Geospatial Consortium, OGC 04–094, version 1.1.0, 117 pp. http://portal.opengeospatial.org/files/?artifact_id=8339
Portele C (2007) Geography Markup Language (GML) encoding standard. Open Geospatial Consortium, OGC 07–036, version 3.2.1, 427 pp. http://portal.opengeospatial.org/files/?artifact_id=20509
Refsgaard JC, Højberg AL, Møller I, Hansen M, Søndergaard V (2010) Groundwater modeling in integrated water resources management – visions for 2020. Ground Water 48(5):633–648
Robbins R (2012) Data management for LTER: 1980–2010. NSF, pp 1–59. Available at: http://www.nsf.gov/pubs/2012/bio12002/bio12002.pdf
Schou JS, Skop E, Jensen JD (2000) Integrated agri-environmental modelling: a cost-effectiveness analysis of two nitrogen tax instruments in the Vejle Fjord watershed, Denmark. J Environ Manage 58(3):199–212
Strassberg G, Maidment DR, Jones N (2004) Arc Hydro groundwater data model. In: Geographic information systems in water resources III, AWRA spring specialty conference, Nashville, May 2004, pp 17–19

Tarboton, D (2013) HydroShare: an online, collaborative environment for the sharing of hydrologic data and models. Proceedings: 2013 CAUHSI conference on hydroInformatics and modeling. http://www.cuahsi.org/pageFiles/DavidTarboton.pptx

Tarboton DG, Maidment D, Zaslavsky I, Ames D, Goodall J, Hooper RP, Horsburgh J, Valentine D, Whiteaker T, Schreuders K (2011) Data interoperability in the hydrologic sciences, The CUAHSI hydrologic information system. In: Proceedings of the environmental information management conference 2011, pp 132–137, http://eim.ecoinformatics.org/eim2011/eim-proceedings-2011/view

Taylor P, Cox S, Walker G, Valentine D, Sheahan P (2013) WaterML2. 0: development of an open standard for hydrological time-series data exchange. IWA Publishing http://www.iwaponline.com/jh/up/jh2013174.htm

Thomas W, Gregory A, Gager J, Kuo I-L, Wackerow A, Nelson C (2009). Data Documentation Initiative (DDI) technical specification part I: overview, pp 1–103. Retrieved from http://www.ddialliance.org/

Van Dijk A, Bacon D, Barratt D (2011) Design and development of the Australian water resources assessment system. In: Water information research and development alliance, Science symposium proceedings, Melbourne, 1–5 Aug 2011

Whiteaker T, Schneider K, Maidment D (2006) Applying the ArcGIS Hydro data model. http://www.crwr.utexas.edu/gis/gishydro01/support/schematutorial.pdf

World Meteorological Organization (WMO) (2008) Guide to hydrological practices. WMO, Geneva, pp 1–296

Yang C, Raskin R, Goodchild M, Gahegan M (2010a) Geospatial cyberinfrastructure: past, present and future. Comput Environ Urban Syst 34(4):264–277

Yang X, Steward DR, de Lange WJ, Lauwo SY, Chubb RM, Bernard EA (2010b) Data model for system conceptualization in groundwater studies. Int J Geogr Inf Sci 24(5):677–694. doi:10.1080/13658810902967389

Zaslavsky I, Williams M, Aufdenkampe A, Lehnert K, Mayorga E, Horsburgh J (2011) Data infrastructure for the Critical Zone Observatories (CZOData): an EarthCube design prototype. National Science Foundation EarthCube White Paper: Designs Category. http://earthcube.org/file/4024/download?token=HGb9YhHv