19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Maintaining synchronized water datasets

R. Power[1] and G. Walker[1]

[1]*Commonwealth Scientific and Industrial Research Organisation*
Email: robert.power@csiro.au

Abstract: The CSIRO and the Bureau of Meteorology (the Bureau) are involved in numerous collaborative projects to transform the way Australia manages its water resources. One of these, the Water Data Transfer Standards project, is concerned with the next generation of encodings for the transfer of water information. Water observation data is a key element of a water resources information system. A standard encoding for data transfer is desirable for a number of reasons: to support integration and update into databases; to define standard data delivery interfaces; and to develop applications with a standard data binding or import/export mechanism.

Water data is currently being delivered from various agencies for integration into the Bureau's national water database: the Australian Water Resources Information System (AWRIS). This process of data delivery and integration is termed Data Ingestion. The current solution uses the File Transfer Protocol (FTP) where the data provided is in various file formats ranging from Excel spreadsheets, text files, binary formats and eXtensible Markup Language (XML).

AWRIS is a centralized data warehouse replicating the data holdings of the water information systems distributed throughout the country and is kept consistent through regular incremental updates from the data providers. The contributing sources of water information are heterogeneous in many ways. For example the water information management systems used are varied, different data models are employed, the terms to describe water related features are not consistent and the labels to identify features are not nationally unique. Maintaining AWRIS as a consistent amalgamated copy of the contributing databases is difficult in practice due to these differences in the participating systems.

The Water Data Transfer Format (WDTF) is an XML data transfer format developed by CSIRO and the Bureau to provide a uniform and consistent description of water information facilitating the process of Data Ingestion. The aim is, over time, for all data to be delivered to the Bureau as WDTF.

Central to the success of AWRIS is the ability for it to maintain a timely consistent replicated copy of the water data recorded at the source agency. The process of keeping AWRIS up to date through regular updates can over time result in discrepancies from the data source. Synchronization covers the process of updating AWRIS as well as identifying and rectifying any drift.

Approaches to solving the synchronization problem have been investigated by the Bureau and CSIRO. These proposed synchronization solutions are presented and discussed. The approaches are yet to be realised since they require supporting functionality scheduled to be developed over time as part of the Bureau's current data management work plan.

Keywords: *Service-Oriented Architecture, Web Technologies, Water data, Data Ingestion.*

1. INTRODUCTION

The synchronization task is a mechanism for the data in AWRIS to be made consistent with the water data managed at the source agency. The agency may provide changes to the Bureau as they occur to synchronize the two systems. No system is fool proof however, so it is expected that a process is also required to the Bureau's data with the source data managed elsewhere to discover anomalies. When differences are found, the Bureau data needs to be made consistent with the source.

These tasks are complicated due to the differences between the ARWIS and source datasets. Data integrated into AWRIS during ingestion undergoes a number of processing steps such as validation, translation, authorisation, verification and conversion as described in Chan (2010). This ensures the incoming data is standardised by transforming it into a common data format, translating it to use standard units of measure (heights expressed as Australian Height Datum, standard date formats, megalitres), feature identifiers mapped to nationally unique values (state based site identifiers mapped to Bureau site identifiers) and so on.

The synchronization task must deal with these data transformation issues to be able to compare data.

The rest of the paper is organised as follows. Section 2 provides background information on the role of the Bureau in maintaining a national database of water information, the function of the data providers, the role of data ingestion, and the use of standards. Section 3 describes the task of maintaining a central data warehouse as a synchronized replicated copy of the various component data repositories. Detailed methods of performing the synchronization task are described in Section 4. The paper concludes with a discussion and outlines future work.

2. BACKGROUND

2.1. Australian Water Resources Information System

The Australian Water Resources Information System (AWRIS) is a centralized data warehouse of water observation data maintained as a replicated copy of information already managed by various Australian agencies. This information is integrated and standardised through regular updates from various sources describing groundwater, river flows and quality, storage volumes, water use, restrictions, entitlements and trades. It will be used to produce regular reports of water assessments and forecasts and the data will be made available to the Australian water informatics community and the general public.

AWRIS is being developed in phases which will see incremental functionality integrated with each release. The first release, AWRIS Phase 1a in 2010, was focused on supporting data for the Water Storage[1] product. The website portal provides access to up-to-date standardised water storage levels and volumes for around 250 lakes, reservoirs and weirs Australia wide. This information can be viewed by state, city or drainage division and includes historical data for comparison purposes. Subsequent phases will focus on including more water information and supporting tools into AWRIS.

2.2. Data Providers

The following is a general overview of the water data collection and processing performed by various state agencies in Australia. This is representative of the water data management operations performed and the task of regularly supplying data updates to the Bureau.

An agency has a number of environmental monitoring sites continuously measuring, for example, surface water quality and quantity. This information is sent to the agency in near real-time, for example every 15 minutes, and loaded directly into their water information management system. This information is referred to as provisional continuous time series data as it is 'raw' data from the instrument. A scheduled task is run on a regular basis, for example every hour, that exports all modified data in the database which is then sent to the Bureau's FTP server. This data flow is depicted in Figure 1.

The provisional data will be curated by the agency as part of standard quality control and assurance procedures. This process will result in changes to the provisional data which are managed as a new modified dataset. Other data changes can occur, for example by applying a new rating curve to recalculate derived data. All such changes are included in the export sent to the Bureau. When the provisional 'raw' data straight

[1] http://www.bom.gov.au/water/waterstorage/index.shtml

from the instrument has been checked by a person to be 'sensible', the data is deemed to be *validated*, also referred to as *archive* data. Validated data is also regularly sent to the Bureau as a distinct dataset, for example once a day, and is on average more than a month older than the corresponding provisional data on which it is derived.

Figure 1: Example data supply to the Bureau.

Some agencies, for example the Murray-Darling Basin Authority (MDBA) and the NSW State Water Corporation, collect water data from state based authorities in a similar fashion as the Bureau. This data is then provided to the Bureau for ingest resulting in the same data being supplied from multiple agencies.

2.3. Data Ingestion

The process of data delivery and integration into AWRIS is termed *Data Ingestion*. Water data is delivered in various time frames from hourly, daily, weekly, monthly or annual. The current solution uses the File Transfer Protocol (FTP) where the data provided is in numerous file formats ranging from Excel spreadsheets, text files, binary formats and eXtensible Markup Language (XML).

A schematic of the data ingestion process is shown Figure 2. There are currently over 260 data providers supplying water information in various file formats to the Bureau's FTP server. Each provider has their own secure FTP directory to place the incoming data which must be packaged as a ZIP file. The file name indicates the type of data being supplied (provisional continuous time series data, validated archive data, groundwater levels and so on), the location the data is for (the site) and the file format (XML, CSV, Excel among others). A record is kept of what data has arrived from who and when. A scheduled task is run every hour to check for newly arrived data on the Bureau's FTP server. The files are un-zipped and copied to a staging area ready for processing in the data ingestion pipeline as described in Chan (2010).

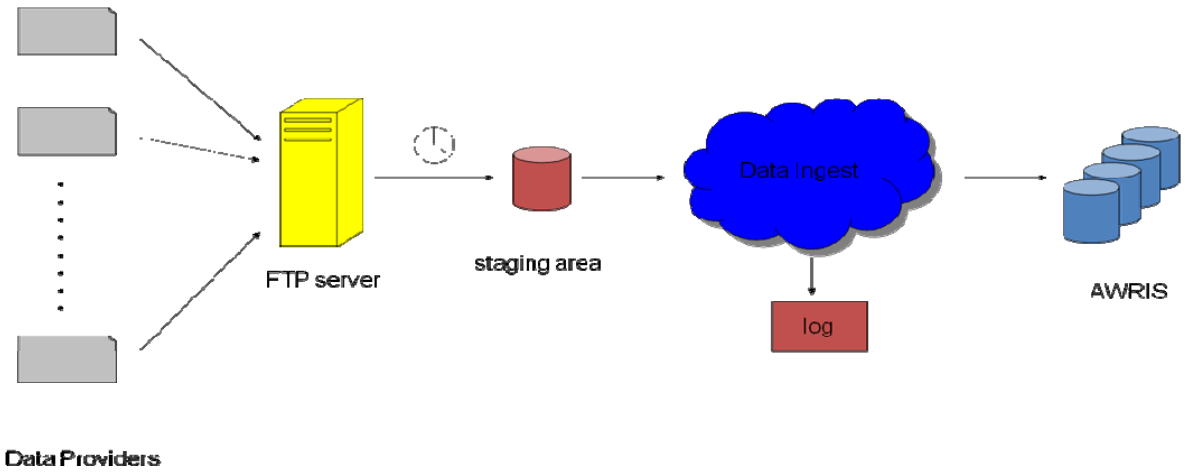

Figure 2: Data Ingestion.

The process of data ingestion and the role of AWRIS are demonstrated by the Water Storage application. The source data is managed by various water agencies around Australia who regularly supply data in various

formats and frequencies to the Bureau. The data needs to be standardised to express values for 'volume' and 'level' in a consistent way (which they aren't around the country), the units need to be the same (megalitres) and the determination of 'daily' values interpolated from data supplied over different time intervals.

2.4. Data Exchange Formats

In the discussion above, the information sent to the Bureau may be in any format. Water data is required to be delivered to the Bureau from data providers by an Act of the Australian Parliament, Section 126 of the Commonwealth Water Act 2007[2]. This Act requires the Bureau to collect and manage Australia's water information, however is does not detail how the information is to be made available to the Bureau.

In 2008 the Bureau started receiving water data from the providers. There was no agreed water data exchange standard available so the data was supplied in a multitude of formats. For this reason the Bureau engaged with CSIRO to develop a standard for the exchange of water information. This was the beginning of the Water Data Transfer Format (WDTF) (Walker et al., 2009).

WDTF is an XML data transfer format that provides a standard description of water information. Its primary purpose is to support the delivery of water data to the Bureau for ingestion. Having a single format simplifies the process of data ingestion and the aim is, over time, for all data to be delivered to the Bureau as WDTF.

3. PROBLEM DESCRIPTION

3.1. Overview

The background information presented above goes some way to outline the complexity of the water data management issues associated with data ingestion. There are numerous details only touched on or omitted, for example: establishing the initial copy of the provider's data in AWRIS (termed onboarding); harmonising different data models (choosing consistent terms for features, using standard units, and so on); identifying updates to send to the Bureau; the data life cycle (retiring old monitoring sites and including new ones); and where is the point of truth?

The problem under consideration in this paper is maintaining AWRIS as a central data warehouse being a replicated copy of the various provider databases and kept consistent through regular data updates. This process of incremental update can over time result in discrepancies between the data source and the AWRIS copy. A mechanism is needed to allow the data in AWRIS to be compared with the source data managed elsewhere to understand how to apply updates and to discover anomalies.

3.2. Incremental updates

WDTF allows specification of both the data and what to do with it when it gets to AWRIS. It does this through transactions. A transaction is an operation over:

a) The domain of an object.

   Where an object represents a time varying data set such as a time series the transaction specifies the temporal subset of the data set which is being delivered. In theory this could also be applied to spatially varying data sets, but there is currently no demand for it.

b) A group of objects bound by a domain.

   For example all the water course extraction permits issues in a time span. The domain does not need to be temporal. For example all the water quality observations pertaining to a particular sample can have the sample as the domain

c) Individual objects not bound by a domain.

   For example a monitoring station (though if the history of a station is of interest it becomes a time varying object). In WDTF these are specified without transactions as they simply replace the existing instance with the new data.

The transactions allow objects to be created, deleted and updated. Updates include new domain bound entries, such as new time value pairs in a time series, as well as updates to values already sent. It is not

[2] http://www.comlaw.gov.au/Series/A00137

uncommon for a quality controlled data set to be edited and the edits sent on to the Bureau. Updating data can have flow on effects for derived products, such as reports. It is important to only update where there are new data or where there are changes. A process is needed to compare the updated transaction with existing data so only changed values are updated.

Some of the varying data is accompanied by invariant data to provide context. For example with time series observation data, metadata accompanies the new time values pairs. In this case the metadata must be considered part of the update. It will replace the equivalent fixed metadata in AWRIS. Object referred to in transactions and identified by identifiers.

3.3. Data system drift

Though updates are used to synchronize provider and AWRIS data systems it must always be expected that something will be missed and the data systems will drift apart. A process is needed to detect a rectify this data system drift.

4. SYNCHRONIZATION

The synchronize task requires two steps: identify inconsistent data and fix it. This rudimentary approach is concisely stated in Algorithm 1 where data from the provider is compared on Bureau systems. The data would be in WDTF. The difficulty is that the Bureau data has been modified during the ingest process and cannot be easily compared as stated in step 4: either the Bureau extract needs to be converted back to its original form or the provider extract converted into a Bureau compatible version.

Algorithm 1: Simple Synchronize

1: $W_p \leftarrow$ data extract from provider
2: Send W_p to the Bureau
3: $W_b \leftarrow$ corresponding data extract from the Bureau
4: **if** $W_p \neq W_b$ **then**
5: Replace W_b with W_p at the Bureau

The data from the provider can be defined by a transaction. The corresponding data from the Bureau (step 3) can be retrieved by applying the same transaction specification to AWRIS.

There are a number of ways to approach the task of comparing the data extracts:

- Transform the Bureau data back to its original provider specific format and content. This ‘un-ingest’ process would allow the Bureau extract to be compared with the provider extract.
- Transform the provider data to the expected Bureau format and content so it can be compared with the Bureau extract. This would be an easier option than above since the software is already available as part of the existing data ingest system.
- Ingest the provider data extract to a temporary ‘scratch’ database location which is specifically reserved for the synchronization process. The advantage is that the existing ingest infrastructure can be used to process the data extract, modified to send the result to the ‘scratch’ location. The result will be a copy of the provider data in a format that can be directly compared with the Bureau's replicated version of the data. The comparison could proceed within the database or extracts made and compared.
- Maintain a pre-ingest version of the data supplied from each provider. This information would be used as the Bureau data extract, W_b in Algorithm 1, which should be in the same content format at W_p and therefore be comparable.
- Have a service translate WDTF files into the *uniform and consistent* version so they can be compared.

The ‘un-ingest’ method described above would be the most difficult to implement and maintain. Each of the other solutions are feasible, although the effort required tailoring the existing data ingest software for use in these solutions in unknown.

WDTF is based on a conceptual model for describing water information. The XML syntax used to represent the data allows the same concepts to be described in different ways: while the semantics remains the same, the syntax can be different. For example, the same water features can be identified using different labels and data values can be expressed using different units of measure, such as megalitres and gigalitres.

A *uniform and consistent* WDTF refers a standard representation of water information where all parties agree to use the same units of measure and same identifiers for features and so on. This allows the WDTF content to be directly comparable.

For two such *uniform and consistent* WDTF files to be directly comparable using text based comparison tools such as `diff` (Linux) or `comp` (Windows) a further agreement needs to be made: the serialization of the XML. This refers to the layout of an XML document so that different unimportant details are removed. For example the use of whitespace, the order of elements and attributes, the use of default values, line breaks characters, and so on. Such an agreed representation is defined as the 'canonical' form for the XML (Boyer 2001). Note that software is freely available to convert an XML document into a canonical representation. For example xmlwf (linux) or the Xerces XML parser.

Two XML documents are identical when their canonical versions are the same. This property can be used for the synchronization task as shown in **Error! Reference source not found.**. Note that the `UniformConsistent` service generates *uniform and consistent* WDTF and the `Canonical` service generates the 'canonical' form for the XML. Also, step 6 may not be necessary since the data in AWRIS should already be *uniform and consistent*.

Algorithm 2: Synchronize using Canonical WDTF

1: $W_p \leftarrow$ data extract from provider
2: $U_p \leftarrow$ UniformConsistent(W_p)
3: $C_p \leftarrow$ Canonical(U_p)
4: send C_p to the Bureau
5: $W_b \leftarrow$ corresponding data extract from the Bureau
6: $U_b \leftarrow$ UniformConsistent(W_b)
7: $C_b \leftarrow$ Canonical(U_b)
8: **if** $C_p \neq C_b$ **then**
9: replace W_b with C_p at the Bureau

Canonical XML is used in XML Signatures when generating a digest for an XML document (Bartel *et al* 2008). An XML signature is the electronic analogy of a human signature: it uniquely identifies the signing entity, ensures the signed object originated from the signing entity, and that it hasn't been altered since it was signed. The W3C Recommendation describes how XML documents can have a digital signature created for them. This process makes use of a cryptographically strong hash, termed a message digest, on the document to be signed. The digest is a hash value generated from the document content, similar to generating an `md5sum` or `cksum` value for a file. The XML document must be in a canonical form so that when the digest is re-calculated and compared with the original value, only changes to the XML content are identified.

The use of digest values can be applied to the comparison of canonical WDTF files as shown in Algorithm 3.

Algorithm 3: Synchronize using digest values

1: $W_p \leftarrow$ data extract from provider
2: $U_p \leftarrow$ UniformConsistent(W_p)
3: $C_p \leftarrow$ Canonical(U_p)
4: $D_p \leftarrow$ Digest(C_p)
5: send D_p to the Bureau
6: $W_b \leftarrow$ corresponding data extract from the Bureau
7: $U_b \leftarrow$ UniformConsistent(W_b)
8: $C_b \leftarrow$ Canonical(U_b)
9: $D_b \leftarrow$ Digest(C_b)
10: **if** $D_p \neq D_b$ **then**
11: get W_p from provider
12: replace W_b with W_p at the Bureau

When the incremental updates arrive at the Bureau, AWRIS needs to determine which part of the data has changed. Sometimes updates represent a large block of domain bound data where only a small part has changed. For example a provider may send time value pair blocks of 3000 entries minimum but only two

edits in that block may have changed. Understanding exactly what has changed will minimise the impact on derived products such as reports. A digest solution would allow splitting of a transaction and comparing the digest parts with AWRIS. This would provide a quick way to search for the differences. The digest approach comes into its own when detecting data system drift. The Bureau could randomly sample parts of AWRIS, digest them and send them to providers to compare digests of the same transaction in their own systems. If a discrepancy is found the digest mechanism can be used to send related portions to the provider to search for the extent of the discrepancy.

These synchronization ideas need to be explored in close consultation with the Bureau since the implementation options require the development of a *uniform and consistent* WDTF and supporting services by utilizing the existing data ingest software.

5. CONCLUSIONS

Water information in Australia is currently managed by numerous agencies that have a legislative requirement to supply it to the Bureau. This information is collected into the AWRIS data warehouse and is kept up to date through regular data updates. Maintaining AWRIS as a consistent amalgamated copy of the contributing databases is difficult in practice and a mechanism is needed efficiently determine when the AWRIS version of water data is different to that recorded in the original provider system.

The solutions proposed rely on the use of WTDF for data representation and transactions and borrows ideas from XML signatures, notably the use of canonical XML document and message digests as a means of summarising the document content. The digest approach holds promise for minimising the impact of updates on AWRIS and detecting data system drift. To implement the digest solution, future work in defining a *uniform and consistent* version of WDTF is required. There are other issues to consider also:

- How regularly should the synchronize process be performed?
- How much data is checked with each synchronize task? This is the synchronization granularity and could range from all data managed by a provider down to a single site.
- How easy is it to identify Bureau data that has been supplied from a particular provider within AWRIS?
- How easy is it to identify the 'corresponding data extract from Bureau', the W_b in the above Algorithms?
- Is synchronization an automated process or one performed by a human operator?
- What is the best way to do the 'replace' operation? Use the normal ingest processes?
- Should full provider exports be periodically applied at the Bureau to ensure consistency 'check points'?

ACKNOWLEDGMENTS

The WDTS Project is part of a five-year water information research and development alliance between the CSIRO Water for a Healthy Country Flagship and the Bureau of Meteorology. Thanks to CSIRO colleague Michael Kearney for reviewing earlier versions of this work and the Bureau's Paul Sheahan for supporting it.

REFERENCES

Bartel, M., J. Boyer, B. Fox, B. LaMacchia, and E. Simon (2008). XML Signature Syntax and Processing (Second Edition). W3C Recommendation. June. http://www.w3.org/TR/xmldsig-core/

Boyer, J. (2001). Canonical XML Version 1.0. W3C Recommendation. March. http://www.w3.org/TR/xml-c14n.

Chan, J. (2010). AWRIS Solution Architecture Phase1a. Version 1.0. Bureau of Meteorology. June.

Power, R. (2011). Web Services to support Data Ingestion: Technology Review. CSIRO Water for a Healthy Country National Research Flagship, Canberra, Australia, 88p. June. Report number EP114323.

Singh, R. (2007). OGC Canadian Geospatial Data Infrastructure Summary Report. OGC Discussion Paper. 08-000. December. http://portal.opengeospatial.org/files/?artifact_id=26608

Walker, G., P. Taylor, S. Cox, and P. Sheahan (2009). Water Data Transfer Format (WDTF): Guiding principles, technical challenges and the future. In R.D. Braddock Anderssen, R.S. and L.T.H. Newham, editors, Proceedings of the 18th World IMACS/MODSIM Congress, pages 2377–2383, Cairns, Australia, July. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation. ISBN 978-0-9758400-7-8. http://www.mssanz.org.au/modsim09/J4/walker_g.pdf.