EPiC Series in Engineering

Volume 3, 2018, Pages 2019–2027

HIC 2018. 13th International Conference on Hydroinformatics

Federating and harmonising disparate soil moisture data sources

Matt Stenson[1], Ashley Sommer[1], Ross Searle[1] and David Freebairn[2]

[1] Commonwealth Scientific and Industrial Research Organisation, Brisbane 4102, Australia
[2] University of Southern Queensland, Toowoomba 4350, Australia

Corresponding author: matthew.stenson@csiro.au

Abstract. As with many industries, digital disruption will play a major role in shaping agriculture over the coming years as decisions become increasingly data driven. A significant proportion of this data will come from on-farm sensors that are becoming easier to source and deploy. While access to sensors is becoming increasingly cost effective, accessing and integrating the data they provide is still a major issue for many, due to the use of different standards for describing and sharing the data. The *Soil sensing - new technology for tracking soil water availability, managing risk and improving management decisions* project has developed a distributed system that addresses the technical challenge of federating disparate data sources through the use of a software mediation layer and a semantically enabled metadata harvest, search and discovery tool. These web services, the O&M Translator and the Data Brokering Layer, allow a unified and federated view of the data, enabling integrated search and discovery and provide access through a SOS compliant API, delivering the data to client using the O&M data model and a TimeseriesML representation. The resulting Data Stream Integrator is already being tested in applications such as SoilWaterApp.

Keywords: Observations & Measurements, SOS, Soil moisture

1 Introduction

Data are an integral part of the decision-making process for land managers. In recent years agricultural decision making has moved from being human observation driven, experience and intuition based, to empirical data driven, and measured using increasingly cheap and easy to deploy on-farm sensors. These sensors have improved in both their connectivity and in the number physical processes they observe.

G. La Loggia, G. Freni, V. Puleo and M. De Marchis (eds.), HIC 2018 (EPiC Series in Engineering, vol. 3), pp. 2019–2027

These data are useful in visual analysis of past and present conditions, especially when several observations are used simultaneously, and when used in models to assess physical indicators or process not easily observed, or to make future predictions on which current decisions rely.

While the cost to access and if needed combine data should be decreasing in both time and financial cost, several factors are in various combinations working against the natural trend of greater volumes of data equals greater availability and use. These are;

- Data are hard to discover. There is no Google for data so there is a great reliance on the collectors and custodians of the data publishing and adequately describing the data,
- Data are often persisted and shared with no formal agreed or reusable data model that can be widely understood and validated, especially by machines,
- Data are often made available in non-standard encodings which diminishes reuse of existing tools,
- Data are not well described, so it is often difficult to understand, trust and repurpose, and
- Data can be locked away by the companies that supply or operate the sensors, or by the land owners that use the sensors, with a belief that openly sharing their data may dis-benefit them.

This is fundamentally a socio-technical data supply chain challenge. The *Soil sensing - new technology for tracking soil water availability, managing risk and improving management decisions* project jointly funded through *CSIRO and the Australian Government's National Landcare Program* is developing tools and technologies to help improve the ability to discover, access, understand and use soil moisture data across disparate data providers. It has done this through the development of loosely coupled services into a Data Streams Integrator system (DSI). Core to the DSI system was extending the Data Brokering Layer (DBL) concept developed as part of the eReefs project [1] as the point of truth for search and discovery, while also tackling specific issues of presenting a consistent view of all the data sources it knows about to the users, realised as standard Sensor Observation Service (SOS) [2] web calls. It does this by harvesting and caching in the DBL crucial metadata about the data sources, and upon a data request, a custom software layer, the O&M Translator (Figure 1), converts the SOS calls to the native calls of the host service and retrieves the requested data stream. It then maps the resulting data to the Observations and Measurements (O&M) [3] data model and returns the data to the requesting user in TimeseriesML [4] encodings Figure 1.

2 The Data Stream Integrator

The Data Streams Integrator (DSI) system developed by the project involves a number of loosely coupled services designed to address different parts of the data supply chain challenge Figure 1. A more detailed explanation can be found in Sommer et al [5]

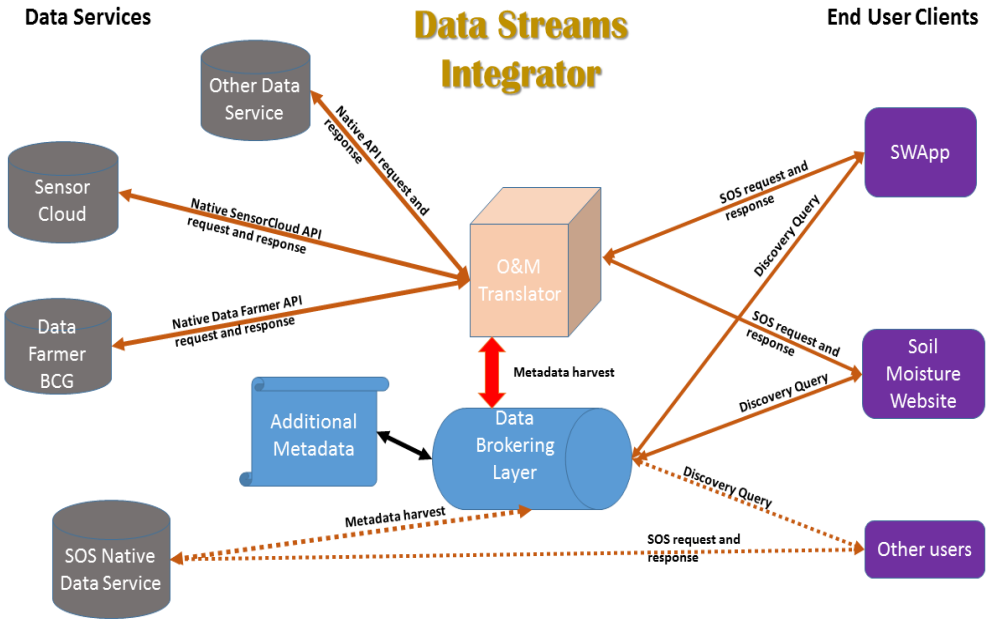

Figure 1 The DSI system showing currently supported data sources on the left, discoverable through the DBL and accessible via API endpoints through the O&M Translator. Data requests are translated on the fly in the O&M Translator to the Observations and Measurements (O&M) data model with TimeseriesML encoding and delivered to the various requesting apps on the right. The DBL harvests and caches metadata from the services it knows about either natively or through the O&M Translator to aid in search and discovery. Services that natively support SOS calls, or raster layers through THREDDS (not represented here) can also be searched and discovered and deliver data directly to the user clients.

2.1 Search and Discovery; the Data Brokering Layer

The various data streams must be discoverable and searchable through a central point of truth. This required that metadata about the data service be harvested, cached and exposed in a standardised way. This was challenging as there proved to be a great variety in the detail, format and access to metadata. To perform this task the DBL concept was adopted and adapted to act as both the metadata harvester and as the point of truth for search and discovery. Once discovered, the DBL provides calling

information for the client to interact with the desired API directly, either native data services, or those provided by the O&M Translator.

Figure 2 A selection of the various portals a single landholder may have to visit to view relevant data about their property.

2.2 O&M Translation

In order to facilitate seamless search, discovery and access of the various disparate data services, each service needs to conform to a standardised data model, and be delivered in a standardised encoding. The Observations and Measurements international standard was chosen due to its proven track record, strength of adoption, governance, and core support within the SOS web interface standard. The TimeseriesML encoding of O&M was chosen once again due to strong support, but it would be possible to support other encodings if needed. Ideally each data service

known to the DBL would provide SOS compatible endpoints, or at least data modelled in O&M, but all the data sources so far added have used bespoke data models, with custom encodings and various levels of metadata. This is in a way understandable given the maturity of agreed standards when many of the services were developed, but also perhaps reflects the somewhat siloed nature of data capture within the agricultural industry, and a general lack of incentives to adopt standards as often it is the data suppliers who must absorb the cost, while the data users see the benefits. Consequently, when a user discovers data they wish to access and makes a data request using standard SOS calls, the O&M Translator converts the SOS request to the source data service's native API and then maps the returned data to the O&M data model and encodes it into a TimeseriesML representation before returning it to the requesting client.

2.3 Data warehousing

While ideally all data federated through the DSI would be retrieved through services at the point of request, or through a collection of data with some form of governance, i.e. probe providers, cooperatives, or government agencies, the reality is that in some cases, the owner of the probe has no clear place for the data to be stored, curated and shared. In these cases, the project has made available an instance of the CSIRO developed Sensor Cloud (now Senapps) [6] web based data storage and discovery system. The Senapps system uses custom adapters to either pull data directly from probes, scrape data from websites and FTP sites, or allow data suppliers to push data in simple plain text formats. In this way data that would otherwise not have been available to the system is able to be warehoused, discovered and shared in a standardised, scalable way.

3 Application of DSI

3.1 Demonstrator Portal

One of the most useful and immediate things that can be done with the data streams accessible via the DSI is to provide a web based interface for potential users to search and visualise the disparate data streams rather than accessing each stream individually through proprietary web interfaces as shown in Figure 2. From here the users can either take away new insight, download a selection of the data in a file based format or decide on a stream of interest and include the API in their consuming application. While portals are no longer particularly new, the integration of so many data streams, accessible and usable in a standardised format via one API presents a powerful capability and one that is neatly hidden from the users. Figure 3.

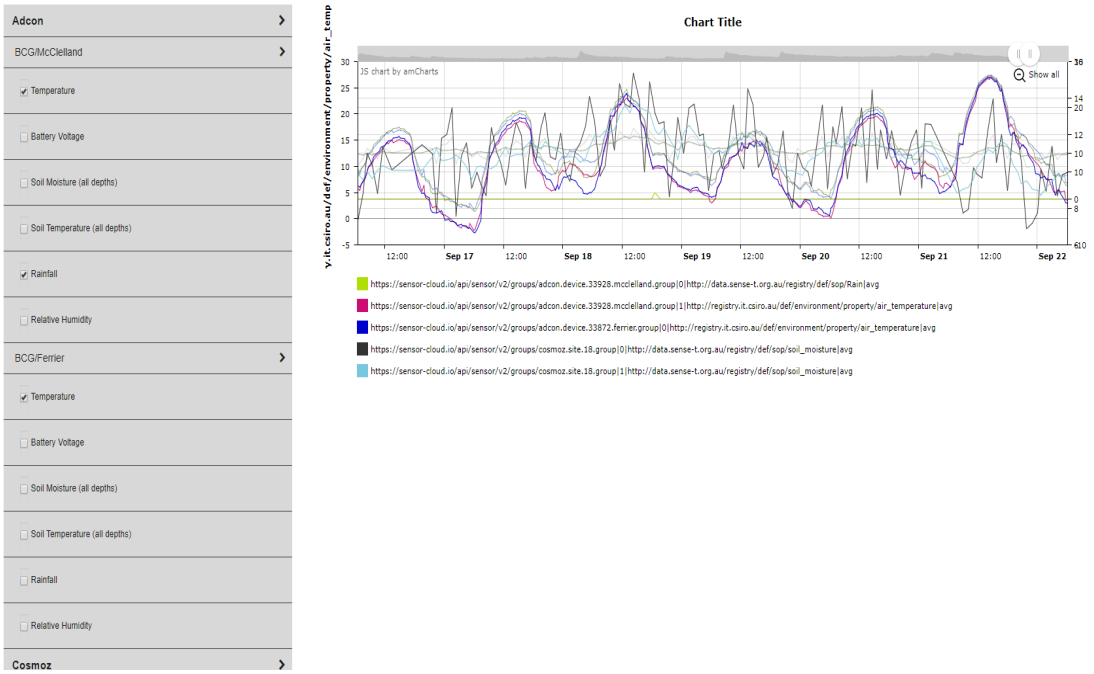

Figure 3 A sample data portal developed for the project showing all the federated data streams available for discovery and display via the one API.

3.2 SoilWaterApp

SoilWaterApp (SWApp) [7] Figure 4 developed by Grains Research and Development Corporation (GRDC) and University of Southern Queensland (USQ) is a low cost, grower friendly application for iPhone and iPad that estimates soil moisture for a specific location and aims to assist growers to make more informed decisions. To run a simulation SWApp requires the user to estimate a starting soil moisture. In the absence of other data these were often not much more than educated guesses. However, with the connection of SWApp to the DSI, users were/are able to search and select the most appropriate soil moisture estimate from the range of sources available within the DSI. Likewise, many other data layers that are currently individually accessed or even cached by SWApp will be available through the single SOS based API provided by the DSI. This will reduce the developer's workload by reducing the number of APIs they need to understand, write adapters for and maintain, while also potentially increasing the available data stream choices available to SWApp users.

Figure 4 SoilWaterApp is an example of a soil moisture data consumer that prior to the DSI was required to develop and maintain links to multiple disparate data sources in order to provide a rich diversity of data for the applications users. With the DSI only one set of APIs' is required for most data discovery and access.

4 Discussion

4.1 Valuing Data

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value." (Clive Humby circa 2006). Data may not necessarily have to be broken down for it to have value, but it needs be discoverable, accessible and understandable, so that its value can be realised through (re)use. The greatest value can be generated from data when it is integrated with other data to form a richer picture, this is often a high cost, manual process. In the absence of an accessible data integration mechanism, each app developer must bear the costs of data access (incentivising potential data providers through paying for data) and integration of the data themselves. Typically, very few apps generate enough value and thus able to generate enough revenue to provide a

financially viable business model or return on investment for the app developers [8]. This market failure can be overcome by investment in common platforms that lessen data integration costs for each app developer.

4.2 Complication of technology; picking winners

To a large degree technology has been a great enabler of change. When developing a system like the DSI no regrets choices can be made around high level architectural approaches but the technology stack used to implement the architectural design will always need to be selected at a point in time; you have to pick winners. The same goes for standards. The inevitable consequence of this is that technology moves on and evolves, and standards are updated and replaced. With this in mind, the real outcomes of this project are not software, services or tooling, it's the concept that federating disparate data streams unlocks unrealised value in data and from a community perspective is worth the effort.

4.3 Social and institutional concerns

A key success factor for this project has been the willing participation of data generators, data custodians, data publishers and the data users. All four actors are vital to realising the potential of shared data but without the data generators and custodians of shared resources agreeing to share their data, a project of this nature would not be feasible.

Scaling up the community of data providers and engaging with organisations other than government or public good oriented organisations with a mandate to govern, curate and share data assets, is challenging. Attempts to achieve data sharing within and between private sector agricultural businesses and in particular farmers (many of whom are small businesses), presents some unique challenges. These relate to potential disbenefits of sharing even apparently innocuous data such soil moisture, that may negatively impact land valuation. To build systems that realise value for a community of users, effort needs to be invested in developing and nurturing data sharing arrangements that do not disbenefit data providers. Developing appropriate institutional mechanisms (such as data governance decision making and rules) that create a trusted space to share data is a key requirement [9] as this creates a trusted environment for data sharing. System technical design (access permission, security and functionality) will be informed by these institutional arrangements that address the real concerns of data providers and other participants.

5 Conclusions

The combination of the DBL and O&M translation layer has shown to be extremely effective in providing integrated data discovery and access, and has been tested in the SoilWaterApp soil water estimation software.

The major challenge still faced by the project is in authentication and access. With so many disparate data providers and each of their users potentially having different access arrangements and needs, management of who can access what data streams becomes quickly unworkable. The ultimate solution may not be technical, but rather social by demonstrating to data owners the extra value that can be realised through the use of open access arrangements, and making it safe for them to do so.

Acknowledgments

This project is jointly funded through CSIRO and the Australian Government's National Landcare Program

Reference

[1] J. Yu, B. Leighton, N. Car, S. Seaton, J. Hodge, The eReefs data brokering layer for hydrological and environmental data, Journal of Hydroinformatics, 18 (2016) 152-167.
[2] O.G. Consortium, Sensor Observation Service Interface Standard, Open Geospatial Consortium, 2012.
[3] O.G. Consortium, Abstract Specification, Geographic information - Observations and Measurements, Open Geospatial Consortium, 2013.
[4] D. Lowe, P. Taylor, J. Tomkins, S. Cox, F. Guillaud, P. Hershberg, J. Lindsey, A. Ritchie, M. Utech, A cross-domain standard for representing timeseries data. , 36th Hydrology and Water Resources Symposium: The art and science of water, Engineers Australia, 2015, pp. p955.
[5] A. Sommer, M.P. Stenson, R. Searle, A Technical Breakdown of a Time-Series Agricultural Data Federation system, Submitted - 13th International Conference of Hydroinformatics (HIC2018)Palermo, Italy, 2018.
[6] M. Coombe, P. Neumeyer, J. Pasanen, C. Peters, C. Sharman, P. Taylor, Senaps: A platform for integrating Time-Series with Modelling Systems., MODSIM 2017, 22th International Congress on Modelling and SimulationHobart, Australia, 2017.
[7] D.M. Freebairn, B. Robinson, D. McClymont, S. Raine, E. Schmidt, V. Skowronski, J. Eberhard, SoilWaterApp - monitoring soil water made easy, Proceedings of the 18th Australian Society of Agronomy Conference, 2017.
[8] T. Sanderson, A. Reeson, P. Box, Cultivating Trust: Towards an Australian Agricultural Data Market, 2017.
[9] P. Box, T. Sanderson, P. Wilson, National Soil Data Project - recommendation for a farmers' data market, 2017.