RMetS
Royal Meteorological Society
Geoscience Data Journal
Open Access

EDITORIAL

Cloud Hosted Real-time Data Services for the Geosciences (CHORDS)

Branko Kerkez[1]*, Michael Daniels[2], Sara Graves[3], V. Chandrasekar[4], Ken Keiser[3], Charlie Martin[2], Michael Dye[2], Manil Maskey[3] and Frank Vernon[5]

[1]*University of Michigan, Ann Arbor, MI, USA*
[2]*National Center for Atmospheric Research, Boulder, CO, USA*
[3]*University of Alabama in Huntsville, Huntsville, AL, USA*
[4]*Colorado State University, Fort Collins, CO, USA*
[5]*University of California San Diego, San Diego, CA, USA*

**Correspondence: Branko Kerkez, University of Michigan, Ann Arbor, MI, USA, E-mail: bkerkez@umich.edu*

Geosci. Data J. ***3****: 4–8 (2016), doi: 10.1002/gdj3.36*

While modern sensing and communication technologies are enabling the observations of geophysical processes at unprecedented spatiotemporal resolutions, the development of these technologies is significantly outpacing their actual use across the geosciences. This is particularly true of real-time data systems, which are now permitting the streaming and analysis of data at the instant of their measurement. Though the use of real-time scientific data is limited, their importance is ever increasing, particularly in mission critical scenarios where informed decisions must be made rapidly.

Beyond applications tied to disaster resilience (earthquake prediction, flood forecasting, etc.), now more than ever there is potential to leverage real-time data to fundamentally change how scientific experiments are conducted. For example, in many geoscientific experiments, faulty sensors are often only detected too late, forcing experiments to be repeated. In settings where mobile sensor nodes are used, or where sampling frequencies need to be adjusted to capture events of interest, few tools are available to adaptively guide the experimental process. This often results in missed observations and wasted experimental investments, but can be remedied rapidly by enabling means to analyse and respond to streaming data.

While real-time data stand to enable a paradigm-shift in geoscientific experimentation, they rarely, if ever, form the first step in a geoscientific workflow. The vast majority of existing data platforms are inherently tuned to nonreal-time applications, where data are often stored in large databases for retrospective analysis and visualization. The few existing real-time data platforms, however, are either proprietary, feed into mission-specific tools, or are otherwise not available to broader stakeholders within the geosciences. While the complexity of these platforms presents a major barrier to the broader adoption of real-time data systems, there are also a number of technical challenges that must be addressed before the use of real-time data becomes commonplace across the geosciences.

Existing real-time data platforms

While interoperability standards such as the *Open Geospatial Consortium* (www.opengeospatial.org/standards) *(OGC) Sensor Web Enablement (SWE)* specifications (Nittel *et al.*, 2008), have created interfaces and metadata encodings to fuse real-time sensor streams into information infrastructures, a common set of tools to couple these streams with workflows and models has yet to be developed. To that end, pioneering efforts are underway by groups such as 52°North and Open Sensor Hub (opensensorhub.org) to develop tools for real-time data within the field of Geoinformatics (Reed *et al.*, 2007; Jirka *et al.*, 2012; Andres *et al.*, 2014). Beyond SWE-based initiatives, a number of other platforms have also been developed to address the emergence of real-time data within the geosciences. *UNIDATA*'s Local Data Manager (Davis and Rew, 1994) provides an event driven infrastructure to manage streaming data. While it has served the purpose of specific projects for many years, the system can be difficult even for an experienced user to install and maintain. Since LDM queues data, the system is not suited for environments in which the stability of networks cannot be assured, which may often be the case with data originating from real-world sensor networks. Its queuing process may also lead to situations where the latest real-time data are not accessible until the queue buffers are flushed, thus causing a backlog of data that prevent timely use.

Other recent real-time efforts have been undertaken through the *DataTurbine* (Tilak *et al.*, 2007) and Antelope (http://www.brtt.com) initiatives. DataTurbine is based on a ring-buffer architecture and is implemented in Java as an open-source, server-side platform for the transport and management of real-time data originating from heterogeneous sensors. While powerful, the ring-buffer architecture does not actively support real-time database operations or coupled model-sensor applications. Furthermore, local server resources can limit the size of the ring-buffer, making it possible to drop incoming data. Cloud-based functionalities and OGC standard support are yet to be implemented as features. Significant overhead exists on the part of users, as *DataTurbine* has to be individually ported to field-specific data loggers and instruments. While these examples may appear specific to one platform, they are echoed by all the other real-time data systems as well. The complexities associated with the deployment and operation of existing real-time data platforms present an overhead too large for most research groups to take on, thus significantly limiting the broader adoption of real-time data across the geosciences. The emergence of commercial data platforms under the Internet of Things (IoT) is beginning to provide easier to use alternatives, but these platforms are not directly tailored to the demands imposed by geoscientific applications (Gubbi *et al.*, 2013; Palattella *et al.*, 2013).

Challenges

A workshop was held in the summer of 2013 as part of the U.S. NSF's *EarthCube Initiative* (http://earthcube.org), entitled "*Integrating Real-time Data into the EarthCube Framework*." The *EarthCube* programme seeks to build a common framework for the analysis, aggregation, and coupling of geoscientific data and models. The primary consensus of the workshop (https://www.eol.ucar.edu/news-and-events/workshops/earthcube-realtime-data-workshop), as provided by over 75 participants spanning a broad set of geoscientific disciplines, revealed that while *EarthCube* will provide an unprecedented framework for disseminating historical data sources, the use of real-time data raises an additional set of complex challenges, which must be addressed explicitly. Furthermore, it was agreed that these challenges are not being addressed by existing real-time data tools.

Complexity of deployment is perhaps the biggest barrier to the adoption of real-time data. A key aspect of managing *in situ* and dynamic sensor data in real-time is providing efficient discovery, access and processing of sensor observations. Ideally, scientists should not have to be concerned with heterogeneous formats, sensors and sources of data. Rather, easy-to-use systems must be developed to permit scientists to focus on analysis and experimentation rather than complex system maintenance. To that end, a number of core challenges should be addressed to facilitate the adoption of real-time data:

- Continued community discussion is required to build consensus around features and the real-world uses of real-time data platforms.
- Installation and configuration of these systems should be seamless and as easy as possible. This may be accomplished by cloud-hosted infrastructure that features preconfigured instances of the platform, thus reducing the need for complex, local user maintenance.
- Real-time data systems should provide standard interoperability interfaces to sensor data to minimize the custom software required for management, visualization and analysis of different types of sensor observations. These platforms should also adhere in as much as possible to common data and metadata formats that adhere to standards (such as the OGC's *Sensor Web*).
- Platforms should also provide a system to archive, navigate and distribute nonreal-time data streams via the Internet.

A reference implementation

Presently, a working group is spearheading the use of real-time data within *EarthCube* under the *Cloud-Hosted Real-time Data Services for the Geosciences (CHORDS)* project (http://chords.earthcube.org). While the primary goal of CHORDS is to drive a community discussion around the adoption of real-time data, reference architecture is also being developed to serve as an example for future implementations of real-time data systems. A number of use cases are being evaluated within this platform to showcase the potential of real-time data towards improving scientific experiments. Examples include, but are not limited to, the analysis and visualization of measurements collected by scientific aircraft, real-time seismic sensor networks for the detection of tornadoes, GPS-based volcano monitoring, and data streaming services for a new generation of affordable 3D-printed weather stations.

One particular use case involves the coupling of real-time, distributed meteorological and hydrologic data. The use case is intended to illustrate the study of extreme events, such as flooding, where hydrologic models are forced by meteorological inputs. In such cases it is vital to couple precipitation data with local flow conditions to forecast flooding. This application couples complex raster data, time series, and metadata, which must be reconciled within the same framework. The CHORDS reference architecture (Figure 1) is explicitly developed with ease-of-use in mind, permitting even small research teams to have a turnkey path towards using real-time data. Three main layers comprise the architecture: (1) the CHORDS Portals, which are the entry and distribution points for all real-time data, (2) CHORDS Services, which provide optional, value-added features, and (3) powerful standards to interface with workflows and EarthCube building blocks.

A CHORDS portal can be launched as a preconfigured instance on a commercial cloud platform,

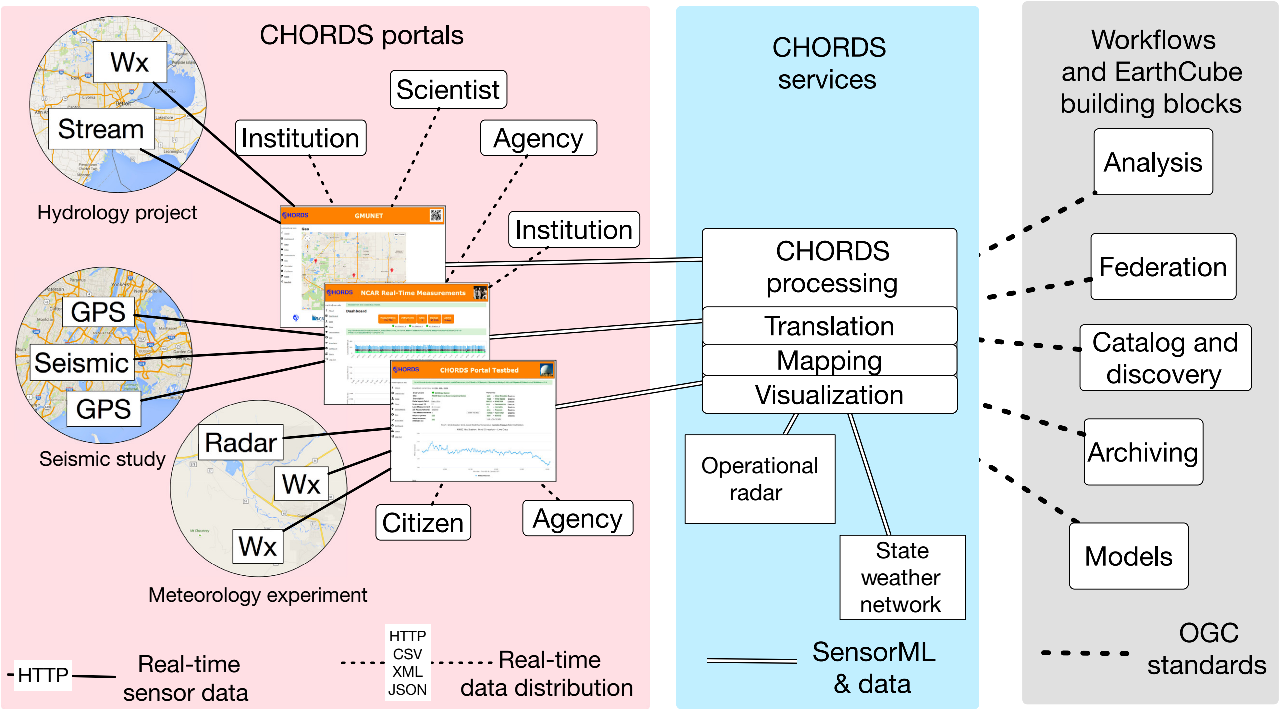


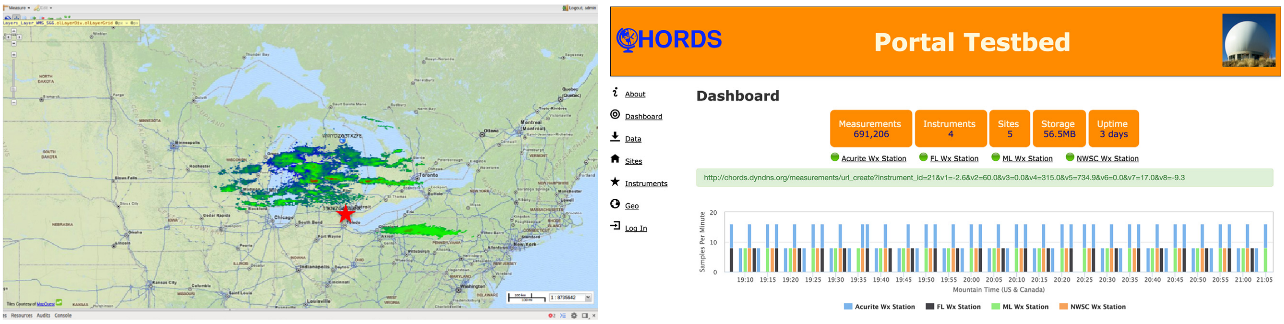
Figure 1. CHORDS architecture: sensors push real-time data to CHORDS Portals, which provide easy web-services access to the data streams. Portals can optionally interface CHORDS Services, which provide additional functionality and interoperability with popular standards and EarthCube services.

allowing users to deploy it with minimal setup overhead. Each CHORDS user owns and manages their own CHORDS instance and interfaces it with their data streams. A preconfigured web server on the instance hosts a user interface, which is used to define data streams that will be ingested by the instance. This interface is used to generate simple URL schemes, which can be loaded directly into data sources (data loggers, instruments, algorithms, etc.). A corresponding ingester is generated for each URL scheme, which translates the incoming sensor data into a common CHORDS format that is then hosted by the portal for external distribution. This permits users to keep their data sources relatively unaltered, having only to push a simple HTTP/REST post when new measurements are made. Data can be written to and read from the CHORDS portal via a set of standard encodings, such as JSON and XML. Data can even be pushed into CHORDS using simple CSV or binary formats to limit programming of field-deployed devices. The support for other popular data formats is continually expanding, with plans to incorporate formats such as netCDF in the near future.

While CHORDS portals provide a rapid way to ingest and share data from multiple real-time sensor networks, their functionality can be vastly expanded by interfacing with the CHORDS services layer. This layer is hosted by EarthCube's CHORDS team and provides a central registry of all deployed portal instances. It serves as a repository of metadata and expands the portal functionality with additional higher-level features such as visualization, mapping, and basic resampling or filtering algorithms. It even provides access to some popular real-time feeds, such as radar data or operational weather networks. A GIS framework (*GeoServer* (http://geoserver.org/)) is built into the services layer to facilitate the visualization, retrieval and discovery of data based on geographic regions of interest (Figure 2). The services layer interoperates with the larger family of evolving web-based OGC data services and standards, a feature that is continual maintained and updated by the CHORDS team to support a growing set of external services and workflows, such as those offered by EarthCube.

A major advantage of CHORDS will be that the end user can work in whatever environment is most effective for them. No specific programming languages are forced onto data producers or end users, as the only requirement is the ability to process HTTP/RESTful requests. This permits the seamless integration of CHORDS services into most existing instrumentation, models, and visualizations. Once configured, research teams can then easily incorporate a suite of algorithms into their real-time workflows. These workflows could include systems ranging from highly integrated command and control systems, data assimilation into models, field project control centres, standalone applications, web visualizations, or spreadsheets.

While the project is still in its infancy, initial use-case assessments are very favourable. CHORDS does not aim to be a one-size-fits-all solution for real-time data, nor is its present implementation an operational real-



http://chords.dyndns.org/measurements/url_create?instrument_id=21&v1=-2.6&v2=60.0&v3=0.0&v4=315.0&v5=734.9&v6=0.0&v7=17.0&v8=-9.3

Figure 2. Example use case: data from a hydrologic sensor node (red star) is coupled with radar feeds to predict local precipitation and flooding.

time data platform. A number of limitations currently exist, which will be addressed in the future based on community feedback. All of the current use cases are based on low latency requirements. The current implementation does not support photo or video data, which may be relevant to studies that require real-time image analysis. While existing systems could readily support data rates at 10–60 Hz per feed, data rates at higher magnitudes, especially for spatial data, would require further testing and improvements. Model integration has also not been tested yet, but use cases are underway to investigate how to best couple CHORDS with publically hosted modelling services. For example, work is underway to connect the real-time hydrometeorological application with hydrologic models for flood forecasting. Bi-directional communications are currently not supported, which means that CHORDS can receive data from remotely deployed instruments but not control them. More advanced OGC SWE functionalities, such as Sensor Planning Services, are also planned for implementation to enable remote tasking of a field sensors, which will enable adaptive sampling of geoscientific phenomena (Andres *et al.*, 2014). Given the infancy of the project, there are many more features that will be required to make CHORDS a fully hardened real-time data platform. This will require the need for built-in security and encryption, which will be vital in protecting field-deployed scientific assets and servers. CHORDS is also not a storage repository, data discovery or cataloguing service, as those features are expected to be addressed by existing domain-community repositories and services. Rather, its goal is to serve as a reference for community feedback, which will ultimately lead to consensus on architectures for real-time data.

Discussion and Conclusions

Given resource constraints of existing experiments, real-time data has the potential to play a pivotal role in the future discovery of geoscientific processes. This will be achieved by responding to data as soon as they are collected to detect faulty instrumentation and adaptively allocate in situ measurement resources. Furthermore, many geoscientific data streams have the potential to change how information is consumed by nonscientific stakeholders (during disaster events, for example). Given the complexity of existing platforms however, much work remains to be done on simplifying the use of real-time data platforms, so that scientists may focus on experimentation, rather than platform maintenance. Over the coming years, the CHORDS initiative will seek to carve out a vision and reference implementation of real-time data. During this process, community engagement will be the most critical mechanism towards making real-time data in the geosciences a reality.

Acknowledgements

The authors acknowledge the support of the United States National Science Foundation in supporting this work through award 1440116.

References

Andres V, Jirka S, Utech M. 2014. *OGC Best Practice: OGC Sensor Observation Service 2.0 Hydrology Profile (OGC 14-004r1)*. Wayland, MA, USA.

Davis G, Rew R. 1994. The Unidata LDM: programs and protocols for flexible processing of data products. *International Conference on Interactive Information and Processing Systems*.

Gubbi J, Buyya R, Marusic S, Palaniswami M. 2013. Internet of Things (IoT): a vision, architectural elements, and future directions. *Future Generation Computer Systems* **29**: 1645–1660. doi:10.1016/j.future.2013.01.010.

Jirka S, Bröring A, Kjeld P, Maidens J, Wytzisk A. 2012. A lightweight approach for the sensor observation service to share environmental data across Europe. *Transactions in GIS* **16**: 293–312. doi:10.1111/j.1467-9671.2012.01324.x.

Nittel S, Labrinidis A, Stefanidis A. (eds.). 2008. *GeoSensor Networks (Vol. 4540)*. Springer Berlin Heidelberg: Berlin, Heidelberg. doi:10.1007/978-3-540-79996-2.

Palattella MR, Accettura N, Vilajosana X, Watteyne T, Grieco LA, Boggia G, Dohler M. 2013. Standardized protocol stack for the internet of (important) things. *IEEE*

Communications Surveys & Tutorials **15**: 1389–1406. doi:10.1109/SURV.2012.111412.00158.

Reed C, Botts M, Davidson J. 2007. Ogc® sensor web enablement: overview and high level achhitecture. In *2007 IEEE Autotestcon*, Baltimore, MD, 2007, pp. 372–380.

Tilak S, Hubbard P, Miller M, Fountain T. 2007. The ring buffer network bus (RBNB) dataturbine streaming data middleware for environmental observing systems. In *e-Science and Grid Computing*, IEEE International Conference on, Bangalore, India, 2007, pp. 125–133.