GLOBAL WATER FUTURES
OBSERVATORIES

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

Global Water Futures Observatories
Data Governance Framework and Architecture

O'Hearn, S.* **, Kendall, K.*, DeBeer, C.*, & Pomeroy, J.*

*University of Saskatchewan
**Corresponding Author

GLOBAL WATER FUTURES OBSERVATORIES

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

Forward

Global Water Futures Observatories (GWFO) provides open access to a vast array of high-quality observational and experimental freshwater data, in original and value-added, post-processed form, originating from its 64 instrumented water observation sites optimally distributed in lakes, rivers, wetlands, glaciers, and drainage basins across Canada, 15 deployable measurement systems for specialized field data acquisition, and 18 water laboratories located at the partner universities where detailed water quality, biological, and other analyses take place.

GWFO is a large national program, and consequently, there are numerous forms of multidisciplinary data (hydrometeorological, turbulent fluxes, quality, isotopic, hill-slope runoff, soil infiltration, eDNA, etc.) originating from a diverse, geographically widespread distribution of sensors, laboratory instruments, and deployable systems, etc. (and here we coin the term **multi-data**). In addition, GWFO sustains a legacy of invaluable freshwater observations from its predecessor programmes (e.g., MAGS, DRI, IP3, CCRN, GWF) which collected data from many of the very same sites that GWFO now does as long as 50+ years ago, and this continuum allows the monitoring and detection of hydrological and water quality changes, and the development of an in-depth understanding of the physical and biological responses to climate change and human pressures.

This report presents GWFO's **Data Governance Framework** and its related **Data Architecture** for covering its nationally distributed data sources, data sharing and transfer protocols, data management systems, data policies, and metadata requirements. In addition to these standard aspects of data management, the Governance Framework and Architecture for GWFO data were further expanded to produce a more *progressive* data management of its holdings through open data and through the adoption of approaches and systems leading to a robust cross-referencing of data with related information from GWFO, and from the several other foundational programs that led to it, on publications, basins, observatories, sites, stations, model inventories, laboratories, principal investigators, software, other data, programme/ project associations, and much more. Thus our data management strategy lends itself to continuous improvement as our data holdings scale upwards and become increasingly integrated, strongly contextualized, and valuable. This pattern of fortification, in turn, provides an ongoing powerful incentive to retain, maintain, and continually build on a comparatively future-proof collection of readily discoverable, interrelated, high-quality data and information for use in the advancement of freshwater science.

GLOBAL WATER FUTURES OBSERVATORIES

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

GWFO Data Governance Framework

GWFO's vision is to operate a national freshwater research facility that supports critical water research to safeguard Canadian water resources in an era of rapid change. GWFO's instrumented water observing sites, supporting deployable observing systems and laboratories provide open access data. This data informs the development and testing of water prediction models, monitors changes in water sources, underpins the diagnosis of risks to water security, and helps in the design of solutions to ensure the long-term sustainability of Canadian water resources. The following components of GWFO's Data Governance Framework are designed to ensure that its data are of the highest quality possible, are properly managed, and that GWFO's vision is fully realized.

Governance Board

GWFO's **Governance Board** oversees all operations, including data management activities, and ensures that the vision and mission of GWFO are achieved, monitors activities and progress towards the objectives of the programme, and holds the facility to the highest standards of operational excellence.

Operations and Coordination

An **Operations Team** reports to the Governance Board and consists of facility leads, managers, technicians, data management staff, the Secretariat, and the Strategic Management Committee. The Operations Team coordinates interaction between all the facilities and teams, identifies gaps, opportunities, and best practices, tracks progress, and facilitates communication, knowledge mobilization, and data sharing. This helps to ensure that the various different types of data and data processes are identified, that similar data processes are standardized as much as possible, and that all sources of data are accounted for.

User Advisory Panel

GWFO has established a **User Advisory Panel** comprised of representatives from industry, government (municipal, provincial, and federal), water management groups, Indigenous communities, data management experts, and data user groups. The Panel informs the Operations Team and builds user engagement with GWFO, provides insights into how data will be employed in the science and decision support needed by users, and provides recommendations on how GWFO's data should yield real-world impacts and new opportunities for GWFO's user-base and services to grow.

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

User Engagement

GWFO maintains individual points of contact with representative users of its extensive data resources through its User Advisory Panel and through the constant monitoring of feedback from its online **User Satisfaction Form** (covering issues of quality and ease of access to sites, facilities, systems, and data). Thus active engagement with GWFO's user base is closely maintained and monitored at all times. This ensures that GWFO will continue to improve and enjoy its excellent reputation for consistently delivering high quality services and data.

Data Policy

The GWFO **Data Policy** (https://gwfo.ca/data/data-policy.php) mandates open access (under a CC_BY 4.0 license) to all GWFO "operational data" collected as part of the routine functioning of the instrumented sites, deployable systems, and laboratories, including, but not limited to hydrometeorological, cryospheric, water quality, water quantity, ecological, biological, toxicological, and other water-related variables. These data are fully GWFO-owned and are made available free of charge through web services, internet downloads, special requests, and through other internet and user-based access tools. All such data sources are indexed in the GWFO central data catalogue.

Decentralized Data / Central Catalogue

The multitude of water data collected by GWFO is not amenable to storage in a monolithic central data repository. Instead, out of practical necessity, it is necessary to keep these data stored in a distributed manner in long-standing storage systems, databases, and repositories specialized and suited best for the type of data they were designed to contain but track all of these data through a central catalogue (**GWFNet**). The catalogue is highly advanced and serves as a national directory on water data and on other assets and information related to the programme. As mandated by the Data Policy, metadata in the central catalogue shall offer centralized and well-organized access to all GWFO-owned data outputs by providing direct web links to download data from the actual online databases or other long-term repository systems where the data reside, as well as links for instant visualization to the extent possible.

Data Availability and Quality

An up-to-date inventory is maintained on all data sources produced through the normal operation of GWFO (as defined in the Data Policy). The inventory includes for each site, laboratory facility, and deployable system in GWFO, listings of documentation indicating **standard operating procedures** (SOPs) on how data are processed for every quality

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

level output (original form, quality assurance/quality control, gap filled, post-processed), along with statements on the suitability of the various levels of data quality for different purposes and end users. The inventory indicates data availabilities (date ranges available per quality level) for each facility/site/deployable system, along with the locations of SOP documents and this information must always be kept up to date in the central catalogue. All SOPs must include all aspects of data processing and management (including backup).

Metadata

GWFO has come (and continues) to identify commonalities across disparate data sources and related information and as data are integrated, the Data Management Team defines the appropriate data model and data architecture to facilitate those linkages. All metadata (and other important related information) are entered into the central catalogue with information on what the data set is, who has it, where it is located and how to obtain it, and, as necessary, to what geographical region it applies. Additionally, all metadata are rapidly searchable and are cross-linked to other helpful related information (e.g., on sites, laboratories, deployable systems, publications/ models/ software, thematic indexes, etc.) that provides context to the data. The central catalogue also provides a dashboard from which data sources can be located by variables intersected with geographic location or through a listing of facilities and deployable systems.

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

GWFO Data Architecture

GWFO sources data from instrumented sites, deployable systems, and laboratories operating across Canada, stored in distributed databases, storage systems, and repositories suited best for the type of data they contain, and provides access to these data through a central information hub, **GWFNet** and GWFNet's **Data Centre**. A conceptual overview of GWFO's data architecture appears in Figure 1.

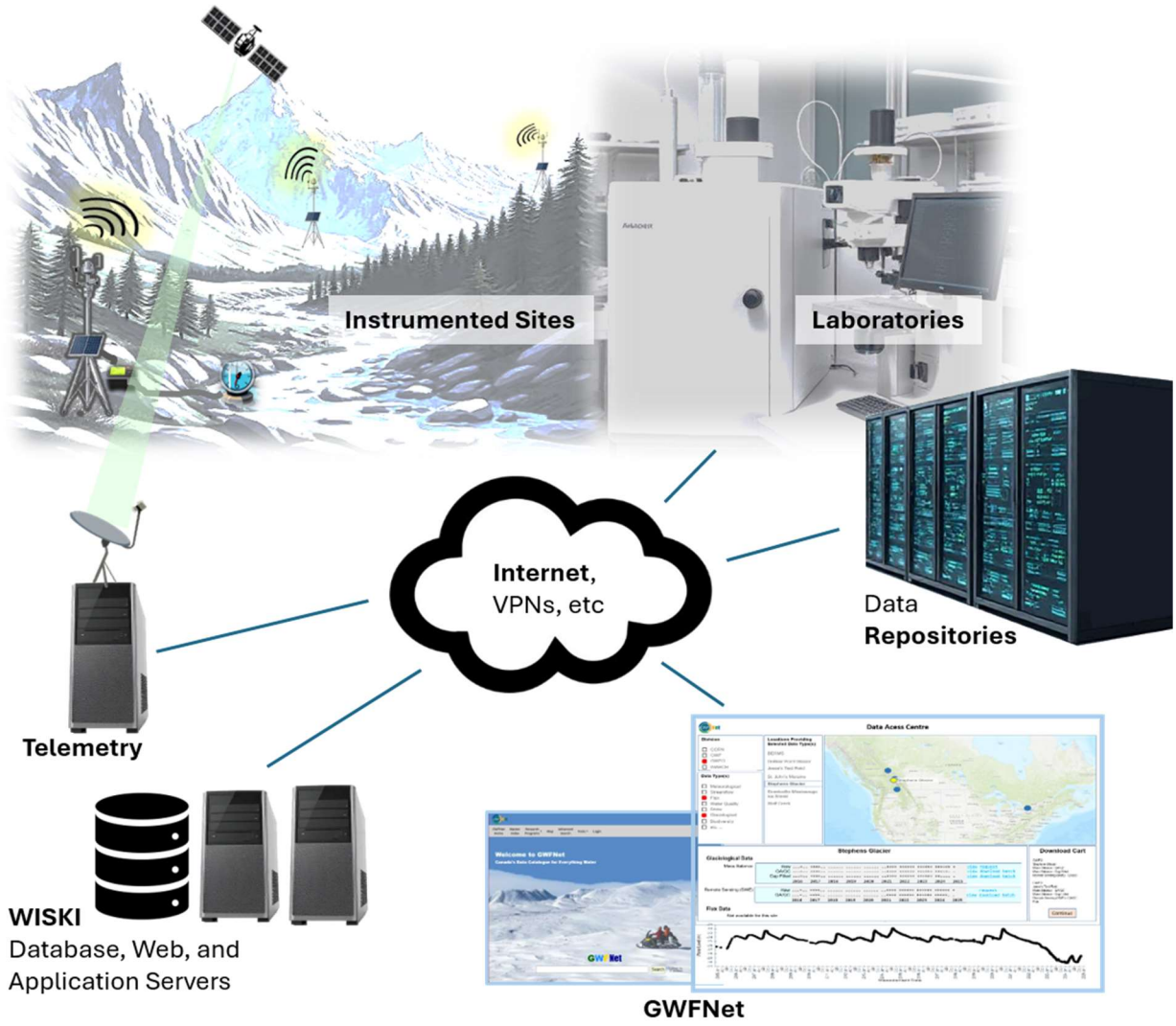

Figure 1: High-level Overview of Data Architecture

A variety of data are sourced generally, through the satellite/ Internet/ VPNs from telemetry systems at instrumented sites, buoys, and deployable systems in lakes, etc. (stored, for example, in WISKI, FLUXNET, DATASTORE.usask.ca, RAEON-Seagull GLOS

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

database) and from laboratories (stored in repositories, e.g. Federated Research Data Repository (FRDR)).

GWFNet Catalogue

GWFO distributes its data in long-standing storage systems, databases, and repositories, such as WISKI, FLUXNET, and FRDR, specialized and suited best for the type of data they contain, and GWFO's central catalogue, **GWFNet** (https://gwfnet.net), tracks the data distributed to these locations.

Allowing data to be decentralized but located through a central catalogue is the only practical way to accommodate widespread, multidisciplinary data from large national programmes. Data may be physically distributed but the catalogue provides a **central access point** for discovery, visualization, and download of every data product produced by GWFO.

In GWFNet, data records describe finalized datasets (and real-time data sources, etc.), provide links or indicate how to access them, and also provide geospatial information (e.g., bounding boxes, shape files, and geographical points of interest). All dataset records specify, among other things, title, authorship, and description, and collectively, this information has tremendous value beyond simply referencing data download locations or instructions because of other features of the catalogue, such as rapid searching, geolocation, and cross-referencing data records with other information in the catalogue.

GWFNet has a number of features which align well with the GWFO data policy (with respect to making all data discoverable, open, and accessible) and are fundamental to the success of GWFO's central platform for locating and accessing nationally distributed data. These features are thoroughly described in the "Key Features" section of the companion document, "Building a National Catalogue for Water Data: A Report on Lessons Learned". In short, they relate to providing information (i. data cataloguing, ii. publication tracking, iii. research site information, iv. geospatial/ geolocation information), linking information together (v. linked records), the capacity for large growth of the catalogue (vi. scalability), its niche area (vii. freshwater science focus), and its capacity to locate information in other data centers (viii. integration with other repositories).

Details on other important technical properties of GWFNet (records, templates, adaptive database, value-centric model, cross-referencing, and divisions) are also discussed there. However, the most relevant property to the present discussion is the "template". Templates in GWFNet are the backbone of metadata record structures and are used to specify the information content of GWFNet metadata records on datasets (inspired by ISO19115), publications, sites, facilities, models, etc., and to ensure that the information catalogued is in a standardized format. Records (based on these templates) have the

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

capacity to be cross-referenced in the catalogue to provide higher-level organization and to implicitly provide the context (e.g., with respect to basins, observatories, sites, laboratories, deployables, publications, models, etc.) within which the data were collected.

These features and properties allow data to be discovered and described through text-based and geographic-based searches. However, another important system in GWFNet is the **Data Centre**, which allows data to be found, visualized, and downloaded based on variable (e.g., temperature, wind speed, greenhouse gas concentration, turbidity, etc.); this key data locator is described in a separate subsection below.

Repositories

The most effective distributed platform integrates information in its central catalogue with *good* choices of distributed data/information repository providers wherever there is a choice. Repositories for data, information, or publications that are good choices are those which:

- ✓ have a long-standing reputation for excellence and long-term preservation (and are well-funded),
- ✓ provide unfettered access to contents/holdings through web service APIs allowing applications within the central catalog to query availabilities (e.g., data quality and timelines), to create visualizations, and to provide download access (ideally to selected subsets),
- ✓ provide persistent identifier access links (e.g., DOIs), and/or
- ✓ provide other links that otherwise yield access.

WISKI Hydrometeorological Database

GWFO collects a substantial amount of real-time hydrometeorological data brought into the long standing, highly robust **WISKI Database** from telemetry (and other automated) systems from its nationally distributed instrumented sites (the list of sites is expected to grow over the next few years). These data sources have multiple series (at different quality levels: original, QA/QC'd, gap-filled) and variables (meteorological, hydrological, groundwater, etc.) and per-site records in GWFNet catalogue all of these data streams and provides visualization and downloading through its Data Centre (discussed in the following subsection). GWFNet also cross-references the GWFO sites with other information in the catalogue, and hence, the WISKI-ingested data streams are also contextualized with that information.

A detailed view of GWFO's site telemetry/WISKI database, application server, and web server system are shown in Figure 2. This system collects data from GWFO instrumented sites, and are available on the GWFNet Data Centre through the KiWIS web service API

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

(and also from tools built into the WISKI such as web portal, designed for technicians who manage the sites). GWFO relies on a variety of communications mechanisms as the level of Internet coverage varies greatly across sites. Sites rely on communication through the GOES satellite (e.g., stations at Wolf Creek, Bologna Glacier, Peyto Moraine, Athabasca Glacier, Helen Lake, Bow Hut, Burstall Pass, etc.), through Remote LoggerNet/ cellular service (e.g., Xplore) support (e.g., stations at Fortress Mountain, Marmot Creek, St. Denis, etc.) or direct Internet/VPN connection (e.g., Brightwater, Clavet, and BERMS).

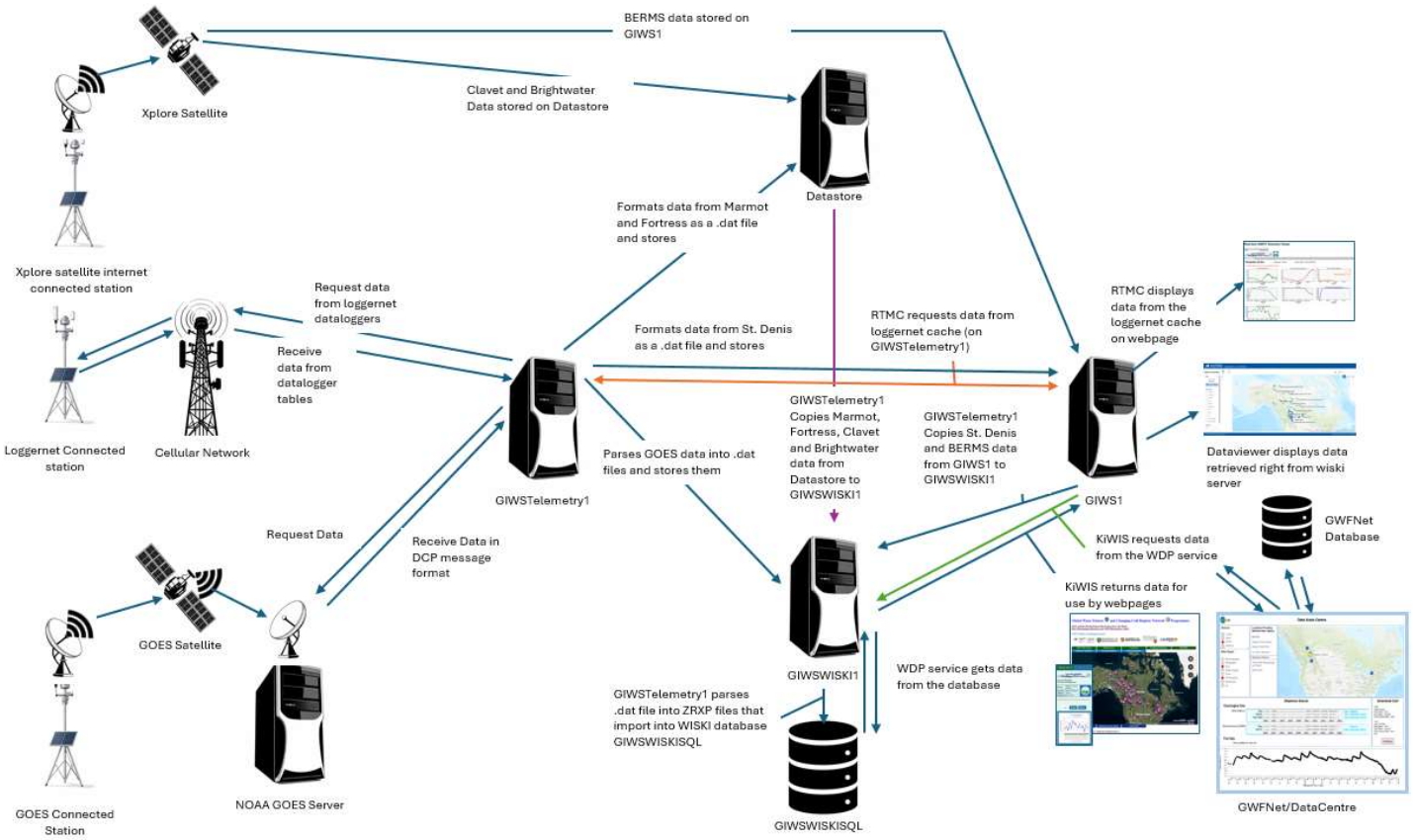

Figure 2: Detailed WISKI Data Architecture

Scheduled services move files (.dat) containing raw data from the data loggers, telemetry servers, and Datastore (University of Saskatchewan storage system) to an Application Server which, in turn, appends all new data (matched by site and by variable) into the WISKI database (stored on a database server running Microsoft SQL Server as the RDMS for WISKI).

A WISKI web server services user web service API requests for data (e.g., running the WISKI Web Portal application and GWFNet's Data Centre) through calls to KiWIS for visualization and data download.

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

GWFNet's Data Centre

GWFNet's **Data Centre** presents users with a single web page (see concept drawing in Figure 3) to access all viewable/downloadable data in the system. It amalgamates and presents per-site and per-facility data availabilities (selectable by data type, temporal extent, and/or geo point or region) and allows users to directly view and/or access data. The GWFNet Data Centre is agnostic in terms of type of data amalgamated on its selection interface. However, its interaction with web services provided by the "KiWIS" API of the WISKI system and the GLOS system for Great Lakes data from RAEON, for example, allows the Data Centre to display real-time advanced information on variables (temperature, wind, humidity, etc.), quality (e.g., original, QA/QC'd, gap-filled measurements), and timelines during which these data are available, and enables the visualization and download of selected subsets of data.

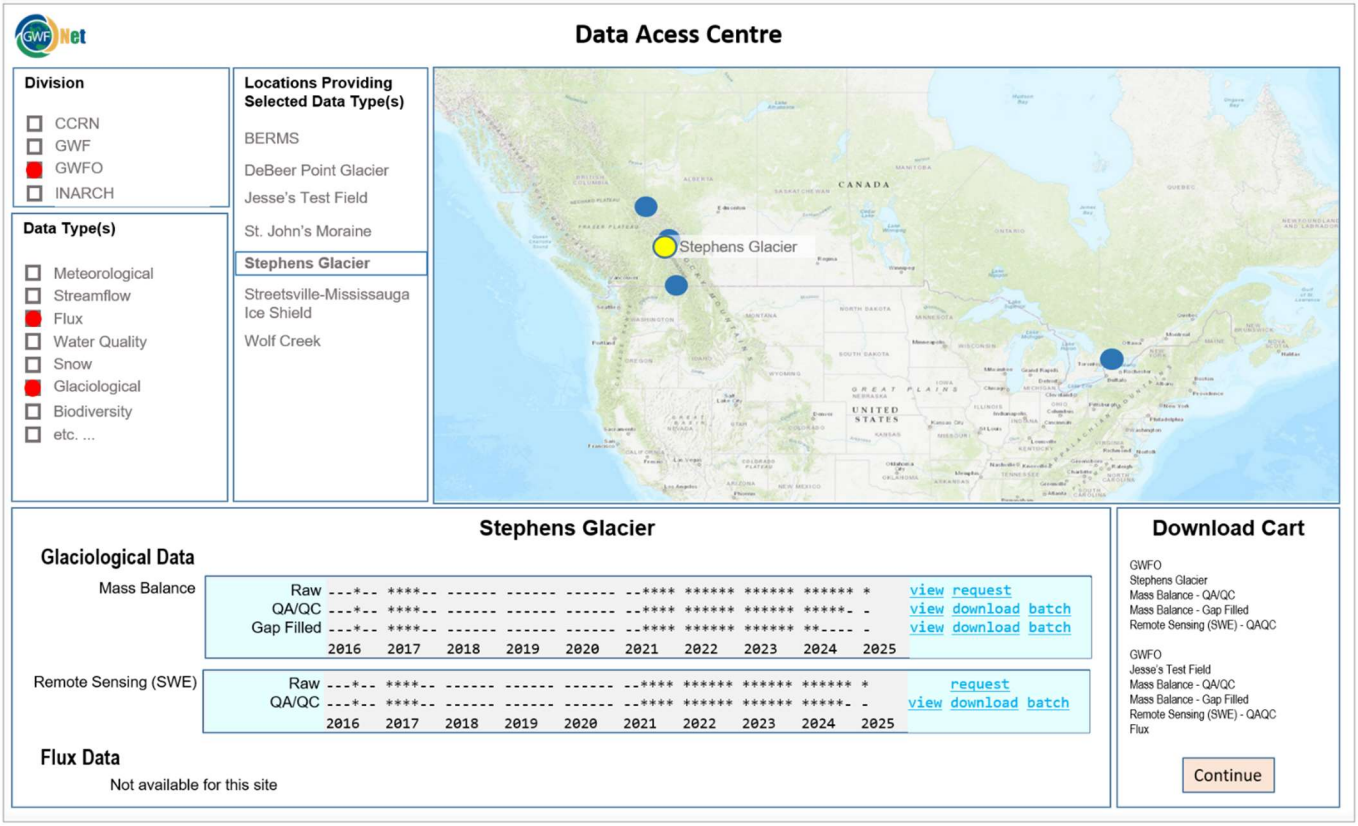

Figure 3: GWFNet Data Centre Concept

The Data Centre will display other forms of data as well (even if such data are not real-time and are manually collected or in the form of pre-packaged finalized datasets

National Hydrology Research Centre
11 Innovation Boulevard
Saskatoon SK S7N 3H5 Canada
Tel: (306) 966-1427
Email: gwfo@usask.ca

deposited in, e.g., FRDR). All information in the Data Centre is collected from records in GWFNet.

Concluding Remarks

Operating under an open data policy is instrumental for making national, widespread, multidisciplinary freshwater multi-data and its related well-integrated information readily available and is essential for realizing the full benefits of a cataloguing system like GWFNet and inline with the GWFO's vision. We have founded a distributed data architecture which is extensively scalable, and which will endure, improve, and grow.