GLOBAL WATER FUTURES OBSERVATORIES

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

Building a National Catalogue for Water Data:
A Report on Lessons Learned

O'Hearn, S.* **, Morrison, M.*, 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

The dynamic nature of water science data—in which approaches to observation, modelling, and prediction of Earth systems are continuously evolving—shape our present data and re-shape our legacy data through a cycle of repeated reanalysis with improved or emerging technologies (e.g., UAVs with new sensors, new models, adoption of machine learning /artificial intelligence). Such repeated interactions with well-managed data—past and present—leads to improved data, new discoveries, and a sustainable process of iterative refinement of knowledge and research questions. This process inevitably results in future data which will become tomorrow's important legacy.

Global Water Futures Observatories (GWFO) is thus steadfastly committed to the long-term stewardship of its open data and, to this end, has devised a new template-based form of data catalogue, GWFNet, capable of incorporating legacy data and future data of a to-be-determined form as readily as it handles data from the present day. GWFNet's ultimate purpose and vision is to enable a variety of information seekers—from the general public to highly specialized scientists—to easily zero in on trails of information and obtain publications, datasets, real-time data sources, and other related information that delivers context to the results associated with their searches (including basins, observatories, research sites, stations, model inventories, software, equipment, principal investigators, programme/project associations, and more).

Background

Key Features

This report presents a detailed account of our experiences distilled from the construction and continuous advancement of GWFNet and lessons learned from providing its eight key features:

1. **Data cataloguing**: GWFNet provides a comprehensive catalogue of data generated at or on behalf of GWFO.
2. **Publication tracking**: GWFNet includes records of publications resulting from GWFO-supported outputs from the observatories and laboratories.
3. **Research site information**: GWFNet contains details about observatories, research sites, and stations used in GWFO-supported research.
4. **Geospatial/geolocation information**: GWFNet can include geospatial information (geolocation points and contours which, in turn, can cross-link other records) in any of its information records related to, for instance, observatory sites and stations, data

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

collection points and/or extents, and model application areas. Combined with record cross-linking, users can navigate to or from maps, and cross-linked related records to contextualize GWFO outputs not just to other records, but to locations on earth with related data/publications/sites/models.

5. **Linked information records**: GWFNet uses template-based information records that are cross-linked to provide a cohesive view of water science research.
6. **Long-term data accessibility and scalability**: GWFNet is designed to ensure ongoing public accessibility to large volumes of complex and disparate data and associated information and can be scaled beyond its current size with GWFO.
7. **Freshwater science focus**: GWFNet specifically caters to information related to freshwater sciences, aligning with GWFO strategic goals by making information readily at hand to inform development and testing of water prediction models, monitor changes in water sources, underpin the diagnosis of risks to water security, and help design solutions to ensure the long-term sustainability of Canadian water resources.
8. **Integration with other repositories**: Whilst GWFNet serves as a central catalogue, it also links to and contextualizes information from other national and international repositories and catalogues, such as FRDR, DataStream, GitHub, and to web services, such as the GWFO WISKI real-time hydrometeorological data system.

These features resulted from experience as we endeavored to implement practical solutions for the management of widespread, multidisciplinary GWFO data and other important assets of the programme. Collectively, the eight features are interdependent and align well with the GWFO data policy (with respect to making all data discoverable, open, and accessible) and are fundamental to the success of our central platform for locating and accessing nationally distributed data.

Technical Characteristic Properties

A brief discussion is presented here on a few characteristic properties of GWFNet that enable the eight features and reference will be made to them in discussing lessons learned:

- records -
  - A **record** contains information about (usually) one dataset, publication, site, laboratory facility, model, person, etc., and each record contains **visual elements**, which are nothing more than information holding elements which range from simple text fields, selection boxes, dropdown boxes, etc., to interactive map elements and charting elements.
  - Every record is assigned a persistent record identifier called a **Reference Number**, which uniquely identifies the record similar to how a digital object identifier (DOI) identifies an article, paper, or book.

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

- Records can link to any other record(s) in the system and the links appear in a "Related Items" section at the top of every record and such linked records are said to be **cross referenced**.
- **Indexes** (or **index records**) are simply records containing no visual elements, and the Related Items section *is* the index; indexes (being records) can link to other indexes (which are also just records).

- templates -
  - Every record is based a curator-provided **template** (e.g., Site, Model, Publication, Dataset (ISO-19115 inspired), Model), which is a plain text file (XML) where **visual elements** (intrinsic to GWFNet) define in sequence all of the input control types appearing on the record (text fields, selection boxes, tables, charts, maps, videos), along with *how* they appear on the web application (or any other user interface) in display/edit modes and in what sequence, and the variable names (unique to the record) associated with the values entered into them.
  - Templates can be changed underfoot to allow record information to be expanded or otherwise changed at any time without the need to alter database schemas or re-tool the website's user interface, and the unique Reference Number for the record will persist and not inadvertently change as a result of updating the underlying format of the record.

- adaptive database -
  - Automatically accepts records created from any existing or newly created template. There is no need for a pre-existing fixed database schema to be defined or updated before populating the database with new forms of records. As needs grow, new templates can be added (or removed) at any time and new records based on these templates can be added immediately and they can be cross-referenced with any other records already in the system.

- value-centric model -
  - All records are, in fact, rapidly accessed collections of **values**. Values represent atomic units of information and come from the information entered into the fields of the record, and the **variable** names under which the values are stored are defined by the *visual elements* in the templates on which the records are based. For example, a *value* of "A new understanding of extreme weather" could be stored under the *variable* "Title" for a record defined under, say, the Publication template.
  - The "gwfnetdb" database is centered around and optimized for speed in fetching values for the ultra-rapid production of information, and is by design, amenable to machine learning and artificial intelligence.
  - This model also supports **value-level controlled access** (where only certain user groups defined in "gwfnetdb" can reference certain **controlled values** of a record as defined in the record's current template, whilst still allowing the remaining,

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

noncontrolled values of the same record to be viewed by anyone. Note that *record-level* controlled access also exists.

- cross referencing -
  - allows any record to be linked (or related) to any other record through the formation of a web link from one record to the other (and from the other back to the first) with a simple drag-and-drop action. This powerful feature is used, for example, to link projects to sub-basins, model workflows and instructions to model outputs, and so forth. Cross-referencing can also be used to create indexes (any time after the fact) for any collection of records considered to be related to each other using the special Index template (which contains no visual elements).
  - The gwfnetdb database is optimized to find related links instantly

- divisions -
  - Records are generally assigned to one or more divisions defined in GWFNet (e.g., we created divisions "CCRN", "DRI", "GWF", "GWFO", "INARCH", "IP3", "MAGS", and others relating the programmes we serve or have served). Users can then filter information by the division(s) they specify causing only those records matching their division(s) to be produced when following links, performing searches, traversing "Related Links", or referencing any mapping systems, making it appear as though these were the only records in the system.
  - Records with no divisions assigned to them will appear in all divisions.
  - Divisions allow GWFNet to serve multiple interrelated water science programmes simultaneously and cross-reference information from a much broader collection of information but allow users to limit their scopes of interest to whatever divisions they have selected.

Lessons Learned

With these features and internal characteristics of GWFNet described, the lessons learned from implementing this information system can now be discussed.

Data Cataloguing

With large, national programmes such a GWFO, there exist very many forms of data originating from a diverse, geographically widespread distribution of sensors, laboratory instruments, deployable systems, and model outputs (here we coin the term **multi-data**). In addition to these, such programmes oftentimes benefit from and need to retain foundational data and metadata from past programmes that led to them (as is certainly the case with GWFO's lineage, as many pivotal data holdings have been carried over from previous programmes and enrich the current one (e.g., MAGS to DRI and IP3 to CCRN to

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

GWF to GWFO); without such commitment to data stewardship and data preservation, and if such data holdings had been discarded, significant continuity and previous effort would have been lost. Cataloguing such data holdings *centrally* by associating searchable metadata information to describe the data elements and how to access them is a fundamental first step of any data catalogue.

However, the multitude of water multi-data collected by GWFO—meteorological, flux, mass spectroscopic, terrestrial and aerial LiDAR, UAV footage, images, glacier mass balance, stable isotope concentration, turbidity—to mention just a few, are 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. In GWFNet, data records (based on the Dataset template) 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, as will be demonstrated later in this report (see cross-referencing), has tremendous value beyond simply referencing data download locations or instructions.

- Lesson Learned: *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*.

- Opportunity: Experience with GWFNet suggests that one possible role for the CWA could be to *provide a national directory on water research* using a catalogue not unlike GWFNet.

Publication Tracking

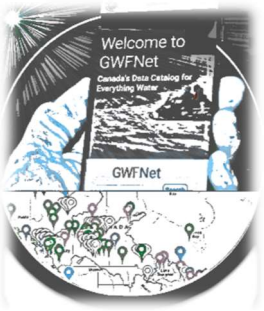
The same approach used in GWFNet for data is used for tracking publications (based on the Publication template) except that publication tracking is straightforward as most publications use document object identifiers (DOIs) and are thus, easily findable, and are generally distributed in journal databases. But in addition to journal publications, the scope of GWFNet's GWFO publication collection encompasses articles and other information collections generally not found in journal databases, including reports, policies, conferences, theses, news releases and other website information, patents, artwork, and even videos related to freshwater sciences that result from information on any of GWFO's data, research sites, laboratories, deployable systems, models, and persons; all of these are considered publications in GWFNet and the Publication template contains a selector to specify the publication type (which is especially helpful in the

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

Advanced Search function). All publications have, among other things, information fields for title, authorship and abstract. Collectively, this information, as will be demonstrated later in this report (see cross-referencing), has tremendous value beyond simply providing publication download locations (which journal publication sites already do well anyway). Coupled with GWFNet's focus on water science, GWFNet's publication collection fills an important scientific role for being a definitive Canadian information resource related to "Everything Water", as showcased on its landing page. Moreover, GWFNet's collection helps locate and present these alternative but nevertheless important sources of information on the state of the art in water science spanning decades and helps prevent these from falling into obscurity, as GWFNet's permanent *Reference Number* links help to prevent this.

- Lesson Learned: *Expanding the definition of publications beyond refereed journal publications to include additional forms of articles like reports, policies, websites related to water science helps in the discovery and preservation of important alternative information on the state of the art of water science.*

- Lesson Learned: *Alternative forms of publications (and expression) can be in the form of videos, special reports, and even artwork related to, for example, Indigenous opinions and knowledge on water science (not necessarily easily translated into the English language) and allows knowledge and opinions to be accepted from a broader range of communities of knowledge holders in ways that allow them to express themselves best.*



Research Site Information

Research site information in GWFNet generally consists of self-contained inventory style records listing major features of the site. Site records (based on the Site template) contain a special interactive "Local Map" control to present geospatial information about the site, and to classify the record. Records having "Local Map" information completed are automatically assigned a geolocation pin (and a "Legend Item") on GWFNet's "Global Map", and users can navigate back and forth between points of interest on the "Global Map" and the corresponding record and its "Local Map" (the "Local Map" may contain significant local detail, for example, navigation pins to records on stations at the site or sites within a basin, etc. not shown on and cluttering the "Global Map").

Geospatial/geolocation information

GWFNet provides a "Global Map" showing geolocation points and the classifications and subclassifications of associated records in a collapsable "Global Map Legend" for all records containing a **Map** Visual Element and whose Map Visual Element has information set.

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

Records based on any templates containing a Map Visual Element (e.g., Data template, Site template) will feature a "Local Map" for the record which can contain pins, bounding boxes, contours, labels, etc. In GWFNet's Edit mode, entering geospatial coordinates into principal location fields of a record's Map Visual Element (or by dragging a location setting pin) causes a geolocation point to be set on both the record's "Local Map" and the "Global Map" and the principal location is credited with owning the record (it might be an actual point, or, say the centroid location of a bounding box, or an assigned point somewhere in a contour, etc.). Reference to the record also appears in one or more places within the "Interactive Legend" of the "Global Map" based on the contents of the "Classification" field of the record's Local Map Visual Element (e.g., "GWFO -> Research Sites; INARCH -> Research Sites"). The pin collection on the "Global Map" and the "Interactive Legend" contents observe any division filtering set by the user, and the pins on the map and the legend items displayed appear in accordance with the divisions selected.

The "Global Map" (through the activation of either its pins or the links on the "Interactive Legend") enables discovery of records based on location (rather than on text searching for the record). At a glance, and with judicious selection of record classifications of interest in the legend, records within a particular region can be easily found geographically.

- Lesson Learned: *It is essential to provide interactive "Local Map" controls for all records (e.g., Sites, Datasets) that benefit from, or completely rely on, detailed and possibly overlapping geolocation and/or geospatial information for land-based context, and to provide a summary interactive "Global Map" with a collapsible "Global Map Legend" classifying and linking to all of these records.*

Cross Referencing

By virtue of the cross-referencing feature, GWFNet instills additional context information at a higher level than is possible through singular metadata records alone and simple cataloguing of data, publications, site information, and so on. Cross references show up as "Related Links" near the top of every record. This feature is very powerful as it enables complete overlays of high-level record organizations (such as creating indexes through the linking of related records to a blank record serving as an index) along with the formation of relationships (by linking together records which are related to each other) in the catalogue. Since any record can be linked to any other record(s), it is possible to co-associate records on data, publications, sites, models, and so forth, to indicate that they are closely associated with each other in concept, in organization (and important conference, for example), or site location (for instance, all the publications listed under some index are all related to a research site with a link to that index; and similar for all models linked to the site which operate there, or instruments that operate in a particular laboratory facility, and so on). As another example, sites may be linked to records that have detailed geospatial information associated with them. As well, the scope of cross

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

references (related items and indexes) also complies (and limits the display of related links) with any division filters set by the user, displaying cross-references only to related records within the divisions selected.

Upon incorporating machine learning/artificial intelligence modules in GWFNet, automated generation of "Inferred Related Links" listings will also appear at the top of records (in addition to the "Related Links" already present should the user choose to have them included). Such inference will promote even more cutting-edge assimilation of water science information in real time.

- Lesson Learned: *The ability to relate records of any type to any other record through cross referencing, and the ability to create indexes by cross referencing blank records with other records, turned out to be one of GWFNet's most versatile features as it allows information to be easily organized and reorganized and this feature alone sets GWFNet apart from all other information systems and catalogues.*

Long-term data accessibility and scalability

A brief discussion of few technical details of GWFNet is needed here in order to understand how GWFNet's architecture assures long-term data accessibility and scalability. Internally, GWFNet uses a database structure, known as *gwfnetdb*, optimized for producing *values* as quickly as possible. In fact, the matching of values (not records, not maps, not even variables) is what marshals the near-instantaneous presentation of information in GWFNet. Every value links to a *variable* (e.g., "Latitude", "Abstract", "Title", etc.) and to a *record*, which assembles values together into one entity. Thus GWFNet is said to have a **value-centric** database structure. Collections of values are produced first (say, from a search or from API queries for particular variables to backfill information elements of another website) and, from values, records (with maps, videos, graphs, text fields, etc.) are assembled (very quickly) from the collection. As well, values are tightly clustered together (irrespective of variables to which they are assigned) to facilitate speed and to enable the introduction of machine learning and artificial intelligence modules to GWFNet.

- Lesson Learned: *The **value-centric** approach in GWFNet has demonstrated itself to be extremely robust as it yields information quickly and lends itself to virtually unrestricted scalability as speed does not diminish detectably with size and is amenable to machine learning and artificial intelligence.*

Freshwater Science Focus

GWFNet focuses exclusively on water science and thus its information collection fills an important, well-defined scientific area, endeavoring to always be a leading Canadian

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

information resource on "Everything Water". From its inception, every aspect of its flexible design has been for accommodating widespread, multidisciplinary information on water.

- Lesson Learned: *GWFNet is committed to an unwavering, consistent focus on freshwater science and endeavors to be a preeminent indexing resource on which its users can rely on for information about water collected from decades past and for decades to come.*

- Lesson Learned: *The ability to assign divisions to records allows GWFNet to easily scale and continue to integrate information from past, present, and future programmes encapsulating their information in a well-organized way; users can leverage decades of well-integrated water knowledge from whichever divisions (representing different programmes, organizations, foundations, etc.) they wish to include.*

- Lesson Learned: *Information assimilated from multiple programmes (divisions) has benefits from a significantly greater capacity for mining knowledge from which new solutions to threats in water security found; and this benefit grows exponentially.*

Integration with Other 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.

GWFNet's Data Centre presents users with a single web page 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 Kisters WISKI system, for example, which stores our distributed hydrometeorological data, allows the Data Centre to display real-time advanced information on variables (temperature, wind,

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

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; WISKI is clearly a *good choice* because of its stability and its API.

The Federated Research Data Repository (FRDR) for finalized, static datasets provides DOIs and direct download links, and so do other repositories such as DataStream for water quality data, the Realtime Aquatic Ecosystem Observation Network (RAEON) for data on lake ecosystems—so these are good choices, and we choose to integrate our information with them.

- Lesson Learned: *In our experience, the most practical approach to making national, widespread, multidisciplinary freshwater multi-data and its related well-integrated information available to the public is to establish a central information service, like GWFNet, designed to present a functional representation of a centralized data system providing unfettered access to all data and information from the programmes it serves to the greatest extent possible.*

Conclusion: General Lessons Learned

This report shall conclude with some general lessons learned from all of GWFNet's features considered collectively.

We have continued to enhance GWFNet as we gained a better understanding of the characteristic ways in which data discovery occurs:

- through the exploration of particular research ideas, for example, "find all data and publications related to streamflow in the Rocky Mountains"
- searches and exploration of publications with their associated datasets
- collective discovery by variable subsets amalgamated from all records in the catalogue, for example, through the GWFNet Data Centre,
- exploring data related to research site(s) of interest
- exploring data linked to geographical regions using the GWFNet "Global Map"
- exploring data generated through simulations from models of interest at various geographical locations around the World.

We learned that the value of cross-linking related information together on datasets, data sources, publications (in a variety of standard and alternative forms), models, basins, observatories, research sites, stations, and indexes is even greater than first anticipated as it imbues the information referenced with a very significant intrinsic context.

Operating under an open data policy is essential for making the benefits of a cataloguing system like GWFNet possible. We have founded a distributed data architecture and

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

centralized cataloguing system in GWFNet which is extensively scalable, and which must endure and grow. The importance of government funding for architectures like GWFNet which facilitate ultra long-term data stewardship cannot be overstated.