University of Manitoba

Centre for Earth Observation Science

Canadian Watershed Information Network Roadmap

Document Control

Version	Author(s)	Type	Date Modified	Comments
1.0	C. Herbert	Working Copy	2017 May 05	
1.1	C. Herbert	Working Copy	2020 Dec. 08	Edited by L. Candlish
1.2	C.Herbert	Working Copy	2021 May 20	
1.3	C.Herbert	Published Version	2021 Nov 20	

Licence

Herbert, C., Candlish L. (2021). Canadian Watershed Information Network Roadmap. Center for Earth Observation Science. University of Manitoba, Winnipeg, MB.

TABLE OF CONTENTS

FIGURES

IMAGES

TABLES

Executive Summary

The Canadian Watershed Information Network (CanWIN) is a Canadian spatial data infrastructure (SDI) system hosted at the University of Manitoba and managed by the Centre for Earth Observation Science within the Faculty of Environment, Earth and Resources. We support research and education and inform management, policy and evidence-based decision making within the Nelson River Watershed and into the Arctic via Hudson Bay. By creating an interoperable infrastructure, CanWIN facilitates the discoverability and accessibility of water and climate-related data across the freshwater-marine spectrum.

The Hudson Bay Drainage Basin is comprised of the Foxe Basin Watershed, Ungava Bay Watershed, Hudson Bay Seaboard Watershed and the Nelson River Watershed. These watersheds span four Canadian provinces and Inuit Nunangat, as well as four U.S. States and, are critical to Canada due to the inclusion of (i) the Prairies, known as the Nation's breadbasket, (ii) Lake Winnipeg as an iconic symbol of human impacts on natural ecosystems, (iii) the Nelson River as the crucial freshwater-marine interface which links Canada's freshwater and Arctic ecosystems; (iv) sub-Arctic and Arctic environments that are predicted to be extremely sensitive to future climate and land-use changes, and (v) the Port of Churchill as a gateway to Canada's North.

CanWIN partners with freshwater and arctic data centres across Canada and globally. It is a founding member of the Canadian Consortium for Arctic Data Interoperability (CCADI) a pan-Canadian collaboration of arctic data centres and domain experts such as Amundsen Science who are developing an integrated Canadian arctic research data infrastructure system (ARDI). CanWIN partnerships with freshwater research and data centres include the Lake Winnipeg Research Consortium and the International Institute for Sustainable development – Experimental Lakes Area (IISD-ELA) where we collaborate on standardizing and sharing freshwater research and community collected datasets. CanWin is a featured case study in the Gordon Foundations recommendations on Elevating Community-Based Monitoring in Canada.

Through these initiatives, CanWIN provides Canadian researchers with a collaborative web-based platform to connect multiple climate and water-related data repositories and datasets nationwide. This integrated network will centralize and facilitate the synthesis of terabytes of digital data that are collected each year, and connect with large scale monitoring initiatives aimed at the characterization of hydrological, biogeochemical, geomorphological and ecological dynamics. Data associated with such dynamics have typically been collected and analyzed in isolation, thus preventing truly comprehensive research and environmental management to be achieved across spatial and temporal scales. By harmonizing data and providing interactive visual, statistical and analytical platforms, CanWIN will enable users to gain scientific and operational insights not previously possible, transforming our ability to address critical scientific questions.

VISION

Mobilizing science for evidence-based decision making. We enable scientists to ask new research questions by giving them the ability to analyze complex, multi-themed watershed issues across broad spatial and temporal extents.

Mandate

To lower the latency for real-time decision making and support evidence-based management, policy and decision-making in the Hudson Bay Drainage Basin.

Mission

- create an online open access data repository, providing datasets integrated temporally and spatially;
- To communicate key research findings in plain language to reach the broadest possible audience;
- To the greatest degree possible, through the online data repository, provide open access to research data and reports in non-proprietary formats; and
- Use ethical data sharing methods (FAIR and CARE) to address unique key stakeholder needs and privacy concerns for information while finding ways to share and integrate Indigenous Knowledge and Western Science

CHALLENGES

Over the last 20 years, environmental and climate change, including priority concerns such as water security, has moved to the forefront of scientific and societal concerns. Projections of climate change, identification of emerging issues, and the development of evidence-based adaptation and approaches are urgent. Addressing Arctic and freshwater change at an ecosystem scale requires access to large sets of heterogeneous data and necessitates cooperation across disciplinary, cultural, and political boundaries. This comprehensive approach requires a focused research data infrastructure (RDI) to accelerate sharing, discovery, visualization, and analysis of diverse data, including potentially sensitive Indigenous Knowledge.

Canada has made significant investments in research infrastructure, including field stations and research vessels, to support the collection of data. These systems, combined with satellite and other geographical information allow for data collection at a broad geographic scale. However, most data remains disconnected, disassociated, not publicly available or difficult to discover, making them invisible to researchers and other potential users (Soranno et al. 2015). In addition, many groups generate Arctic and Freshwater site-level environmental data, each with different needs and approaches to collection, analysis, access, and sharing (Koivurova et al. 2010, Whiteman et al. 2013). Much of this data is costly and labour intensive to collect and not easily replicated (Fogal et al., 2013). Traditional Indigenous Knowledge (TK) and Inuit Qaujimajatuqangit (IQ) also present unique challenges; some may be proprietary or sensitive requiring special considerations for access and availability. In addition, almost no data is delivered or communicated back to the originating Indigenous or Inuit organization in an understandable context and format, inhibiting evidence-based decision making for these groups.

Site-level data, critical for answering ecosystem-scale research questions which are needed to address climate change and watershed management issues, is much more useful if it can be combined with the complementary spatial data available at national and regional scales (Soranno et al. 2015). Data harmonization (combining and integrating datasets at various geographic and temporal scales) has also been shown to increase the value of initial data collection cost by between 5 to 200 fold (Soranno et al. 2015) and ensures consistent data provenance in the long term.

BACKGROUND

The Canadian Watershed Information Network (CANWIN) (formerly the Lake Winnipeg Basin Information Network), is a web-based open-access data and information repository created by Environment Canada as part of the Lake Winnipeg Basin Initiative, and under Canada's Action Plan on clean water. It was created to address key water quality issues within the Lake Winnipeg Basin. In 2012 management of the network transferred to the University of Manitoba in the Clayton H. Riddell Faculty of Earth, Environment and Resources within the Centre for Earth Observation Science (CEOS).

CanWin is funded through projects within the Centre for Earth Observation Science (CEOS) at the University of Manitoba. CEOS (umanitoba.ca/ceos) is a Type 1 Research Centre within the Clayton H. Riddell Faculty of Environment, Earth, and Resources.

CanWIN has expanded its mandate to include arctic research data and is a founding member of the CCADI (www.ccadi.ca). The Consortium (Table 1) represents a critical mass of top academic researchers from six institutions, which are home to many of Canada's Arctic scholars, as well as Inuit research organizations, federal agencies, and the non-profit sector. All partners have reputations for excellence as collectively demonstrated by hundreds of grants and publications across a spectrum of Arctic marine, terrestrial, social sciences, geospatial and computing sciences. The Consortium has decades of experience working with complex (e.g. TK, IQ, geospatial, genomics) and dynamic (e.g., sensor data, community-based monitoring) data, and have the common goal of providing ethically open, accessible, and comprehensive digital Arctic resources to the broadest possible audience (**Error! Reference source not found.**). The Consortium has the expertise to work across research domains to address common data challenges facing these users.

Table 1. Partners, Canadian Consortium for Arctic Data Interoperability

Universities	Inuit Organizations	Federal Agencies	Not-for-Profit / Private Sector
• **University of Calgary (UC)** • **Université Laval (UL)** • **University of Manitoba (UM)** • **University of Ottawa (UO)**	• Inuit Tapiriit Kanatami (ITK) • Inuvialuit Regional Corporation (IRC)	• Polar Knowledge Canada (POLAR) • Natural Resources Canada (NRCan)	• Polar View • Cybera, Inc. • Amundsen Science

• **University of Waterloo (UW)**			

How CanWIN Can Help

CanWIN enables universal access to research data in the Nelson River Watershed, Hudson Bay Drainage Basin, sub-Arctic and Arctic regions, accelerating and generating new research and knowledge directions. By enabling multi-disciplinary data harmonization we will empower and inform users, decision-makers, community members and researchers around a host of issues, from watershed management at a community scale to remote monitoring for water quality.

> *The True Value of data can only be realized by combining disparate datasets to address broad scale questions (Soranno et al., 2015)*

CanWIN supports the FAIR and CARE principles for data sharing. The CARE principles complement the FAIR principles and encourage the sharing of open data in a way that considers both people and purpose in open data advocacy. We support these principles through:

Ethically Open access: Making data open broadly without restriction, on an equal basis, with exemptions for ethical reasons;

Findable: Making data discoverable. By using web-enabled standards (e.g. schema.org, Datacite) we make your data widely discoverable and accessible including to sites like Google dataset search and the Canadian Federated Research Data Repository;

Accessible: Building into existing UM infrastructure to safeguard against corruption and loss;

Interoperable: ensuring your data can be used and understood in the context of other data both by humans and machines. We do this by using controlled vocabularies, cross-walking common metadata standards and working closely with other initiatives (e.g. Open Geospatial Consortium, CCADI, Research Data Alliance); and

Reproducible: ensuring your data are as reusable as possible so you and others understand how the data were collected and whom to contact for more information. We do this by working with users to provide as complete a metadata record as possible and using common metadata standards like ISO-19115, FGDC and Datacite.

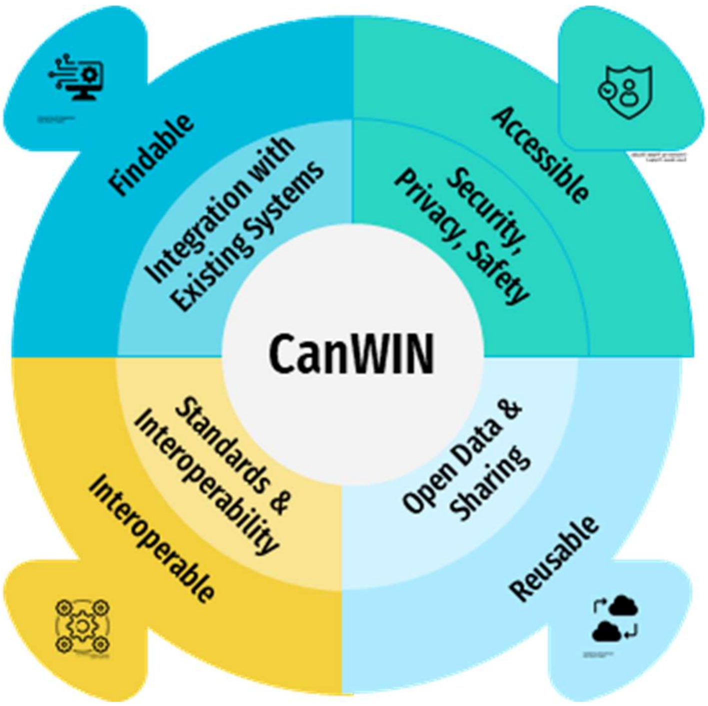

Findable
Integration with Existing Systems
Accessible
Security, Privacy, Safety
CanWIN
Standards & Interoperability
Interoperable
Open Data & Sharing
Reusable

Figure 1. CanWIN FAIR principles wheel

The CanWIN strategy is supported by five key components that make up a robust research data infrastructure (RDI). Infrastructure as a Service (IaaS)

1. Infrastructure as a Service (IaaS)
2. Data as a Service (DaaS)
3. Information as a Service (InaaS)
4. Software as a Service (SaaS)
5. Community as a Service (CaaS)

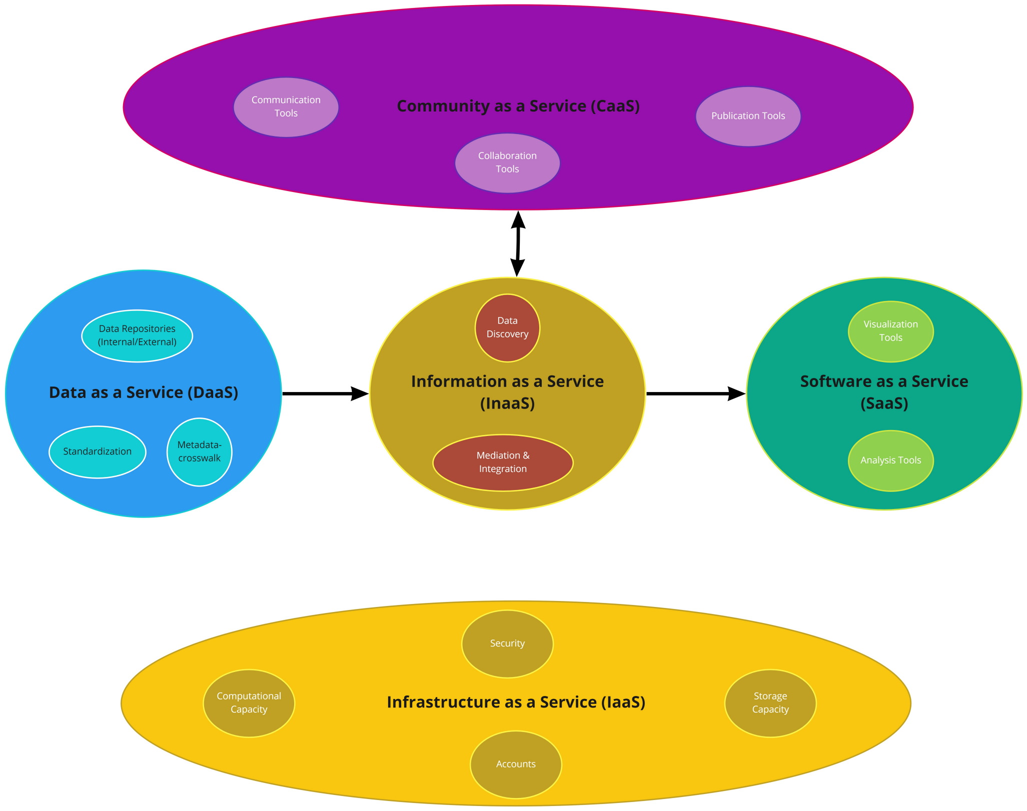


Figure 2. CanWIN Services

Figure 2 is a visualization of the five key components for CanWIN. Some of these components are already functioning, while others are on the roadmap for development as capacity and funding become available.

Table 2. CanWIN Services and Systems Roadmap

Service	Component	Tools	Technology Stack	Purpose	Status	Timeline
Data as a Service (DaaS)	Data Repositories	Geoserver, CKAN, ERDDAP	Application	• Share data in a variety of formats • Expose geospatial data for users • Endpoints for interoperability	In Progress	2021
	Standardization	FME Server, Google docs, CCADI	Application	Interoperability, Data Curation and Standardization, metadata cross-walk	Operational	2018
Information as a Service (InaaS)	Mediation & Integration	CKAN	Application	• CRIS (Current Research Information System) Mediator and Integration platform. • Data & Systems Interoperability	Operational	2019
	Data Discovery	CKAN	Application	Support the use and integration of open data and data products with community-based monitoring and research datasets	In Progress	2021
	Data Discovery	GeoServer/Geonode	Application	• Geospatial Mediator and Integration platform. • Data & Systems Interoperability	In progress	2021

Software as a Service (SaaS)	Visualization Tools	GeoServer/Geonode/other	Application	Interoperability and geospatial data sharing	In Progress	December 2021 – Feb 2022
	Analysis Tools	R Server	Application	Development of accessible data curation and analysis platform for research use – Future Work	Future Work	
Infrastructure as a Service (IaaS)	Computational Capacity	CKAN, Geonode, R Server	Infrastructure	Supported by UM Computer Services	Operational	
	Storage Capacity	PostgreSQL (CKAN)	Application	Mediation (metadata and data aggregators) and visualization layers	Operational	
	Account Services	ERDDAP	Application	Mediation (metadata and data aggregators) and visualization layers	Operational	
	Account Services	CKAN	Application	Redesign and upgrade of platform for interoperability	In Progress	
	Storage Capacity	Data Preservation and Rescue	Application	Datasets are curated into either active sharing platforms (CKAN, Geonode/server, Alfresco), internal active project or archived (GIT, Alfresco)	Operational	
Community as a Service	Communication Tools	Alfresco Document Server	Application	Improve community use of collaborative sharing tools for document, code and data sharing	Operational	

	Collaboration Tools	Mspace	Application	Document sharing and metrics	Operational	
	Publishing Tools	GitLab Server	Application	Version control, data curation	Operational	
	Publishing Tools	CanWIN Main Website	Wordpress Web Server	Communication of CEOS Research programs, platforms and facilities – to be integrated into new CEOS research website	Operational	
	Collaboration Tools	LabCollector	Application	Supports MBGL lab management	Operational	2016
	Communication Tools	CanWIN/CEOS research site	Webserver	Communication of CEOS Research programs, platforms and facilities	Depends on above	

Data as a Service (DaaS)

1. Through its collaboration with the CCADI, the CanWIN has been developing a robust data infrastructure ecosystem, to ensure we can provide data based on the FAIR and CARE principles. We provide data as a service by implementing strategies such as metadata cross-walking of standards (e.g. schema.org, DCAT, DataCite 4.2, ISO 19139, FGDC) to facilitate interoperability between multiple repositories.
2.

Services for Data and Systems Interoperability

As part of the CCADI, CanWIN cyberinfrastructure will support researchers and others throughout the data life cycle. Interoperable data infrastructure that supports the processes (white ovals) and product generation (grey ovals) that are part of the data lifecycle are integral to support research and decision making in any system (Figure 3). IO enables datasets to be combined, and services to interact, without repetitive manual intervention, enhancing the value of the data beyond its original purpose (Euskirchen 2014). An IO system must enable data access that can support many different users. By collaborating with key groups in freshwater and arctic data management, CanWIN ensures that it can provide data and metadata that is machine and human readable and accessible.

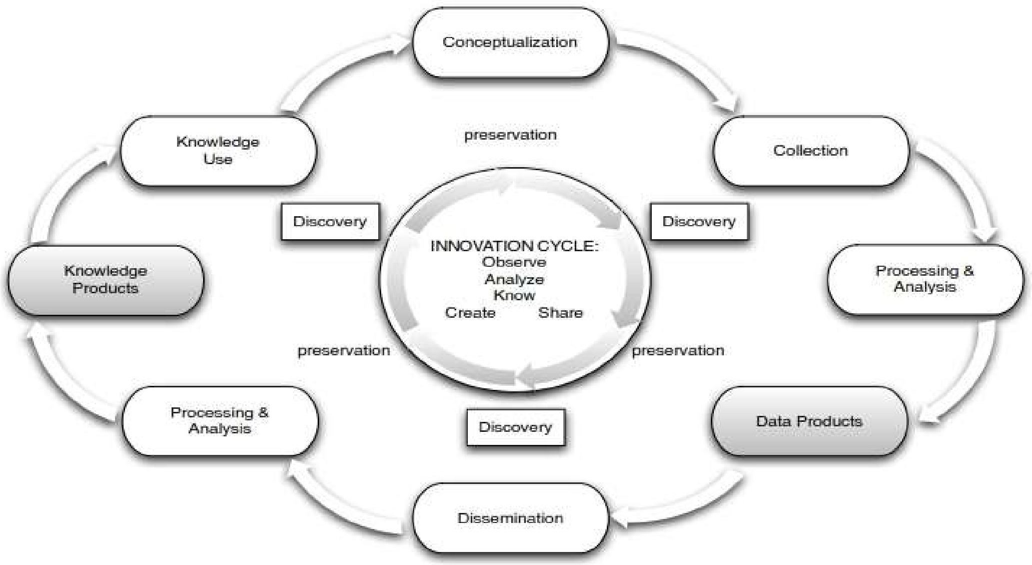


Figure 3. CCADI cyberinfrastructure. Diagram courtesy of Peter Pulsifer

Establishment of Standards and Specifications

The promotion and facilitation of international collaboration towards free, ethically open, sustained, and timely access to data through useful, usable, and interoperable systems are key principles in CanWIN.

The Joint Declaration of Data Citation Principles (JDDCP) laid out a set of principles that stressed data should be considered "legitimate, citable products of research" (Fenner et al. 2016). This can only be achieved by establishing and adopting national and international standards to support data exchange and interpretation.

Through its use of open-source software such as CKAN (www.ckan.org) which powers our data hub, CanWIN supports interoperability, in part by employing the same metadata standards adopted by the Governments of Canada, the U.S., U.K., Australia and multiple provincial, federal and environmental organizations.

Standardizing metadata terms (cross walking) between partners and their various metadata schemas permits interoperability of multi-disciplinary data and facilitates FAIR sharing. Adaption of the Federation of Earth Science Information Partners (https://wiki.esipfed.org/Schema.org_Cluster) schema.org profile allows CanWIN data to be broadly discoverable by sites like Google dataset search and the Federated Research Data Repository (https://www.frdr-dfdr.ca/repo/).

Information as a Service (InaaS)

In consultation with users, stakeholders in the Lake Winnipeg Basin and collaborators such as the CCADI, CanWIN has developed a multi-layered approach to provide Information as a Service. Services include:

a) a data curation process that is conducted by students under the supervision of a data curator. This process involves various workflows, including using a combination of cloud technologies (Google docs, FME server) to curate and develop reproducible templates for common data types, further processing of data using R and Python scripts and curating vocabularies using a data dictionary created from standardized and commonly used ontologies (e.g. BODC/NERC);

b) creation of a mediation and integration layer which facilitates data discovery through the use of metadata aggregators such as CKAN and Geonode or similar platforms. The Geonode platform is built on top GeoNetwork, allowing out of the box implementation of protocols such as OGC CSW and OAI-PMH

The CanWIN IaaS provides users with methods and tools to clean and standardize their data and cookbooks and codebooks to allow repeatable data analysis.

Including Inuit and First Nation Perspectives, Knowledge and Information

In this time of rapid climate change, TK and the underlying observations of Indigenous peoples are more important than ever. Along with the knowledge of non-Indigenous local inhabitants (e.g. citizen science), TK is increasingly documented and represented as digital data. Placing this information in context with other scientific research and providing a mechanism to share it back to the science and Indigenous communities is vital to understanding arctic and freshwater ecosystems.

Software as a Service (SaaS)

Bringing Computation to the Data

The quantity of data available, especially satellite earth observation data, means it is often impractical for users to download the needed data to their local environment. Through the CCADI, CanWIN works with the Polar Thematic Exploration Platform (https://portal.polartep.io/ssoportal/pages/login.jsf) to facilitate user access to algorithms and data remotely, without the need to download and manage large volumes of data. This will eliminate the need for individual users to be limited by varying IT infrastructures, computing power and software licensing.

The CanWIN infrastructure provides users with methods and tools to curate and standardize their data, and cookbooks and codebooks (details of the scripts/tools used to clean data) to facilitate repeatable data analysis and provenance. Figure 4 and 4 depicts figures from a Data Cookbook for combining on-board instrument data for algal productivity with ship meteorological data (AVOS). Users can download the separate AVOS data and run it through a python script to standardize column headers and data format. The two files from the incubator data can be uploaded by the user to an R interactive interface, and the cleaned data downloaded.

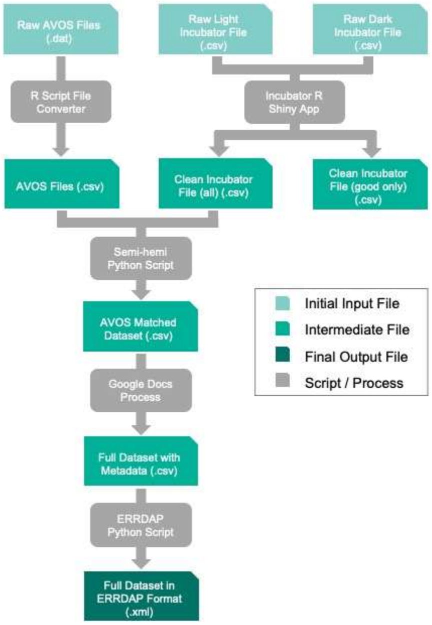
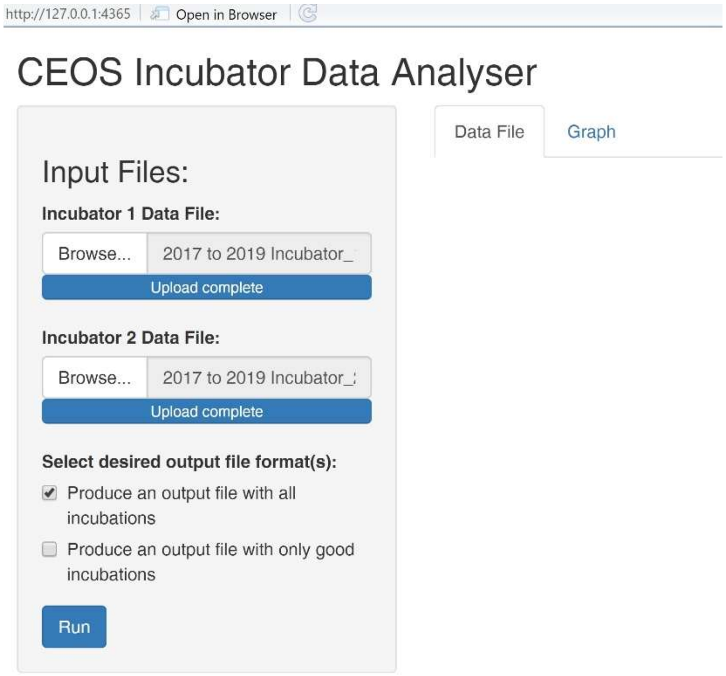

Raw AVOS Files (.dat)
Raw Light Incubator File (.csv)
Raw Dark Incubator File (.csv)
R Script File Converter
Incubator R Shiny App
AVOS Files (.csv)
Clean Incubator File (all) (.csv)
Clean Incubator File (good only) (.csv)
Semi-hemi Python Script
AVOS Matched Dataset (.csv)
Google Docs Process
Full Dataset with Metadata (.csv)
ERRDAP Python Script
Full Dataset in ERRDAP Format (.xml)
Initial Input File
Intermediate File
Final Output File
Script / Process

Figure 4. Figure from Data Cookbook, outlining standardizing and analysis of two datasets



CEOS Incubator Data Analyser

Input Files:

Incubator 1 Data File: 2017 to 2019 Incubator_

Upload complete

Incubator 2 Data File: 2017 to 2019 Incubator_

Upload complete

Select desired output file format(s):

☑ Produce an output file with all incubations

☐ Produce an output file with only good incubations

Figure 5. R UI to clean Incubator data from onboard instrumentation

Scalable Computational Power on Demand

Working with large data sets is often computationally intensive. This means modern platforms need to provide users with highly capable IT infrastructure for data processing, storage, and networking. By working within the university cyberinfrastructure and collaborating with partners such as PolarTEP, CanWIN can provide users with more computation power than a single PC can provide.

Infrastructure as a Service (IaaS)

Provisioning of computing resources, complemented by storage and networking capabilities enables cost-sharing. CanWIN allows users of all price and technological capabilities to access and process large amounts of data.

All CanWIN systems are housed within the University of Manitoba's computing infrastructure, including all key systems being placed behind the UM's security and firewall infrastructure. Servers are backed up weekly, monthly and yearly, with yearly backups housed offsite. This enables us to ensure the long term stability of our platforms and infrastructure.

Data Provenance

Data provenance is essential to produce highly curated and integrated datasets. Good data provenance provides a record that details how a dataset was produced, all changes made to the data, and any other details that may be required to analyze a dataset (Soranno et al. 2015).

Data provenance allows data harmonization, which in turn increases the value versus the collection cost of data by 5 to 200 times (Poelen, Simons, and Mungall 2014). Projects are managed using a loose agile development strategy within our GitLab server (https://cwincloud.cc.umanitoba.ca/).

CanWIN acknowledges that data has varying levels of privacy depending on the data collected and the partners involved. Traditional Knowledge information, for example, may only be available for the community involved in the collection. Data collected with industrial partners in a research program may only be available to other research partners for the duration of the project. Data associated with graduate research or publications may also have embargo periods before public release. To accommodate varying user requirements for privacy, CanWIN provides the ability to share data privately with assigned partners as well as publicly.

Services for Data Preservation and Rescue

Past observations must be continually reused and repurposed to increase current understanding and retain value. Therefore, when collecting and storing data all the necessary descriptive information (metadata) must be preserved. Preservation (Archiving) is often a forgotten component of data collection planning, leaving data managers to pursue ‘data rescue activities. These are costly and often not very effective. In constantly changing political and environmental climates, even current data are at risk and data management plans should be put in place when data collection or grant application planning begins.

CanWIN works with the Canadian Association of Research Libraries (CARL) (https://portagenetwork.ca) to provide a data management template and resources to help data providers plan their data management strategy, from pre-data collection to long term storage.

Community as a Service (CaaS)

CanWIN provides collaborative tools for users to publish, share and discuss their results. These tools allow users to share their data and software/code in a versioned, secure environment. By providing these tools CanWIN builds community among users.

CanWIN also provides data back to users through blog posts and updates on the lwbin (lwbin.cc.umanitoba.ca) website, or as “data stories” on the datahub (CKAN platform). For each dataset CanWIN hosts, where possible a “data story” is created. This story consists of the metadata and additional information that provides context for the data (creating a FAIR data object) and allows for more robust data provenance.

Community Building and Governance

Improved data sharing in the context of our global system requires community building, collaboration, and coordination of efforts. To do this, a better understanding is needed of the nature of the data community (who is doing the work, at what scale, where, what systems, etc.) and what, collectively, that community is trying to achieve. By expanding our services to include not only Inuit and Indigenous perspectives, but also citizen scientists and community-based monitoring work, we can facilitate improved communication, outreach, and coordination within communities in Canada and internationally.

CONCLUSION

There are many system-scale research problems of scientific priority where CanWIN can fill the current research gap by providing a platform to examine watershed behaviour across the Prairie to Arctic continuum. Among many that CanWIN will support include: 1) the assessment of the effects of climate forcing on water, sediment and pollutant transport dynamics in Plains and Boreal landscapes; 2) the optimization between maximum agricultural productivity and preferred downstream water quality while adapting to non-stationary climate forcing; 3) the identification of the causes, socioeconomic impacts and possible remedies to eutrophication in central Canada's great prairie lakes; 4) the combined role of freshwater in the marine system and changes in sea ice conditions; 5) the influence the freshwater-marine ecosystem has on which Inuit and others in the Arctic depend (e.g. food security, traditional activities, commercial fisheries). These issues are exacerbated in large or remote hydrosystems where hydrometric data and chemical data are either sparse or not accessible due to researchers working in silos. Addressing them requires access to large sets of heterogeneous quantitative and qualitative data which will allow the examination of the longitudinal connectivity of water, sediments and nutrients from the Rockies to Hudson Bay.

This strategy will transform our ability to address critical scientific questions and to meet unaddressed researcher, Inuit, Indigenous, government, operational service provider and private sector needs for data-driven knowledge. The RDI will facilitate data discovery and description, platform interoperability, information upload, analysis and visualization including Inuit Qaujimajatuqangit and Traditional Knowledge and information, which will lead to more efficient and effective use of data. It will increase multi-sectoral contributions to knowledge, environmental protection, heritage preservation and economic development.

REFERENCES

Euskirchen, E., et al., Snow, Permafrost, Ice Cover, and Climate Change, in Global Environmental Change. 2014, Springer. p. 199-204.

Fenner, Martin, Mercè Crosas, Jeffrey Grethe, David Kennedy, Henning Hermjakob, Philippe Rocca-Serra, Robin Berjon, Sebastian Karcher, Maryann Martone, and Timothy Clark. 2016. “A Data Citation Roadmap for Scholarly Data Repositories,” December. https://doi.org/10.1101/097196.

Poelen, Jorrit H., James D. Simons, and Chris J. Mungall. 2014. “Global Biotic Interactions: An Open Infrastructure to Share and Analyze Species-Interaction Datasets.” *Ecological Informatics* 24 (November):148–59. https://doi.org/10.1016/j.ecoinf.2014.08.005.

Soranno, Patricia A., Edward G. Bissell, Kendra S. Cheruvelil, Samuel T. Christel, Sarah M. Collins, C. Emi Fergus, Christopher T. Filstrup, et al. 2015. “Building a Multi-Scaled Geospatial Temporal Ecology Database from Disparate Data Sources: Fostering Open Science and Data Reuse.” *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0067-4.

Fogal, P.F., L.M. LeBlanc, and J.R. Drummond, The Polar Environment Atmospheric Research Laboratory (PEARL): Sounding the Atmosphere at 80° North. Arctic, 2013: p. 377-386.

Koivurova, T., Limits and possibilities of the Arctic Council in a rapidly changing scene of Arctic governance. Polar Record, 2010. 46(02): p. 146-156.

Whiteman, G., C. Hope, and P. Wadhams, *Climate science: Vast costs of Arctic change.* Nature, 2013. 499(7459): p. 401-403.

APPENDIX 1

GLOSSARY

Controlled Vocabulary: an established list of standardized terminology for use in indexing and retrieval of information. An example of a controlled vocabulary is the SeaDataNet Common Vocabulary Catalogue (https://vocab.seadatanet.org/search)

Data Harmonization: Modifying data by adjusting for differences in units, formatting, naming and other conventions and setting to a previously established standard/naming convention

Data Mart: An organized collection of data designed to provide for analysis of a specific process, subject area or question" i.e. "sample sites with chlorophyll data and total phosphorous data within a specific geographic area". Allows only data of interest to be analyzed.

Data Provenance: A record that details how a dataset was produced, all changes made to it and any other details required to analyze a dataset.

Site-level data: Data collected at small geographic or temporal scales. Often focus on relatively few sampled variables.

APPENDIX 2

CANWIN WORKFLOWS

Table 3. CanWIN Data Levels. Adapted from ODM2 Core: Processing Levels. https://github.com/ODM2/ODM2/blob/master/doc/ODM2Docs/core_processinglevels.md. Accessed 10 Nov. 2017

Level	Description
0	Raw data: unprocessed data and data products that have not undergone quality control. Depending on the data type and data transmission system, raw data may be available within seconds or minutes after real-time. Examples include real-time precipitation, streamflow, and water quality measurements
0.1	User-provided or historical data: Data that was provided to CanWIN by the user or is historical with unknown provenance. Data will not be quality controlled by CanWIN and quality may be unknown. Data will have any provided metadata applied. Data will be uploaded as-is.
1.0	First Pass QC: A first quality control pass has been performed to remove out of range and obviously erroneous values. These values are deleted from the record. e.g: Online Environment Canada streamflow data, laboratory data provided to a user.
1.1	Quality Controlled Data: Data that have passed quality assurance procedures such as Level 1.0 and have been further quality controlled by data provider before being submitted to CanWIN (e.g. Idronaut data with only downwelling (upwelling data removed) data included).
1.2	CanWIN curated data: Data that have undergone initial quality control from the data provider and have been further curated by a CanWIN data curator (e.g. cleaning script applied)
1.5	Advanced Quality Controlled Data: Data have undergone complete data provenance (i.e. standardized) in CanWIN. Metadata includes links to protocols and methods, sample collection details, incorporates CanWIN's or another standardized vocabulary, and has analytical units standardized. Note: Process still under development in CanWIN
1.6	Combined Data Product: Data that has gone through the level 1.5 data cleaning process and has additional data combined with it (ex. AVOS data combined with incubator data). The dataset was determined to provide better context for users by being combined pre-sharing through the datahub, but one or more of the datasets may be available as an individual dataset.

2	Derived Products: Derived products require scientific and technical interpretation and can include multiple collection types. E.g.: watershed average stream runoff derived from streamflow gages using an interpolation procedure.
3	Interpreted Products: These products require researcher (PI) driven analysis and interpretation and/or model-based interpretation using other data and/or strong prior assumptions. E.g.: watershed average stream runoff and flow using streamflow gauges and radarsat imagery
4	Knowledge Products: These products require researcher (PI) driven scientific interpretation and multidisciplinary data integration and include model-based interpretation using other data and/or strong prior assumptions. E.g.: watershed average nutrient runoff concentrations derived from the combination of streamflow gauges and nutrient values.

There are 3 main workflows that CanWIN manages. In workflow 1 (Image 1), data is managed at the Centre for Earth Observation Science (CEOS) from data collection via a CEOS maintained instrument to data sharing.

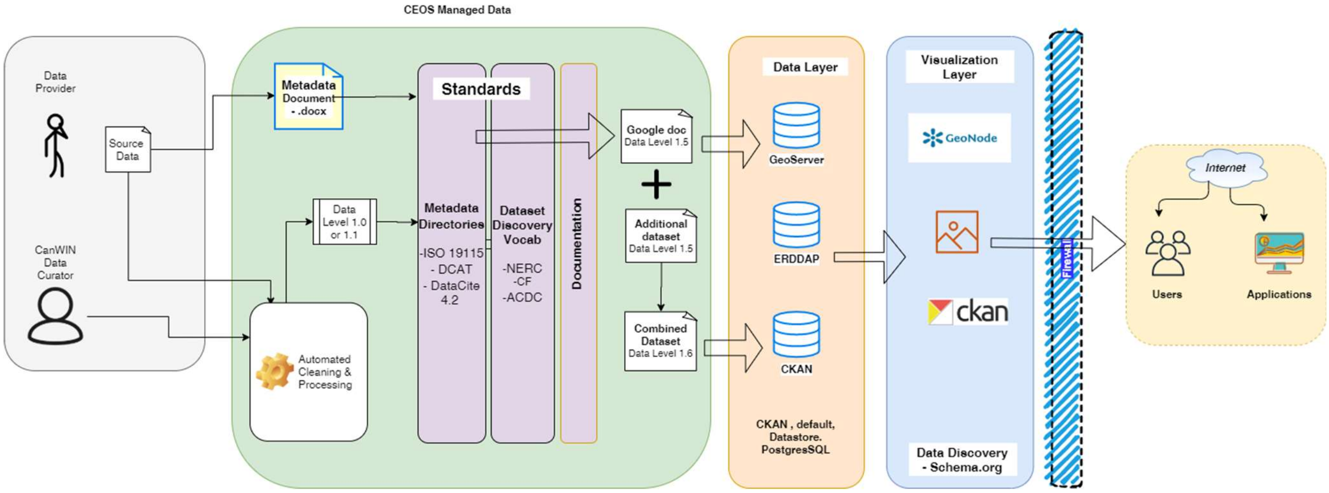


Image 1. CEOS/CanWIN instrument managed workflow

In workflow 2 (Image 2), the Data provider provides a pre-processed file for ingestion into CanWIN.

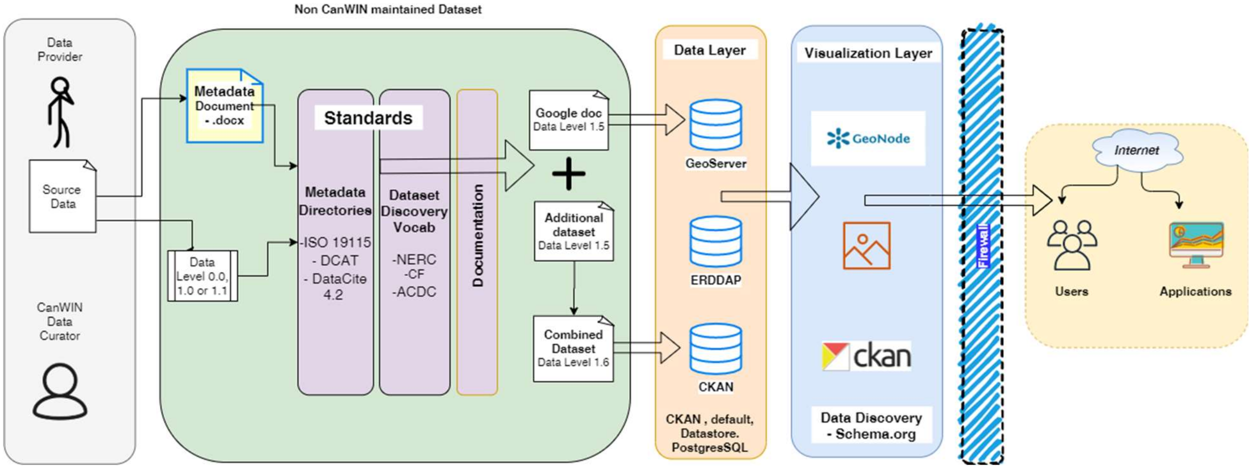


Image 2. Non CEOS/CanWIN managed dataset workflow

In workflow 3 (Image 3), data is harvested directly from a data repository

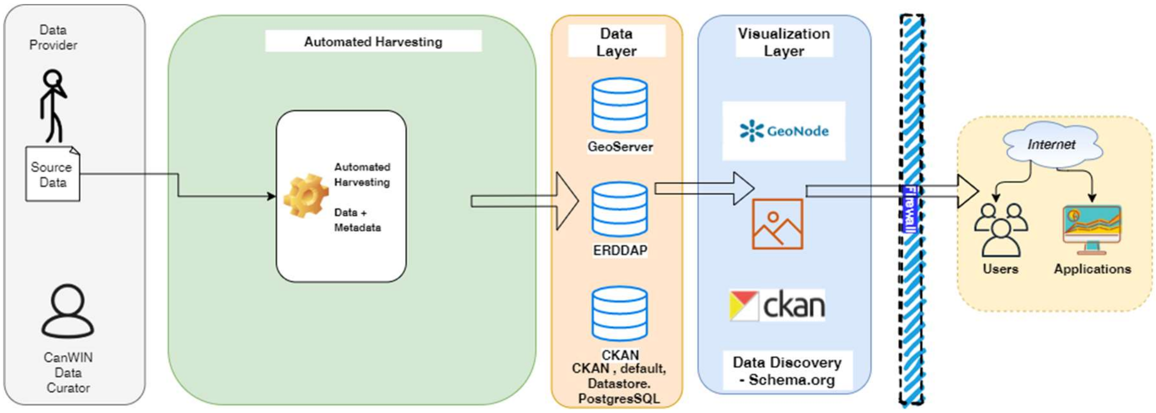


Image 3. Automatic data harvesting

Data is curated using a combination of automated scripts and a google workflow (Image 4).

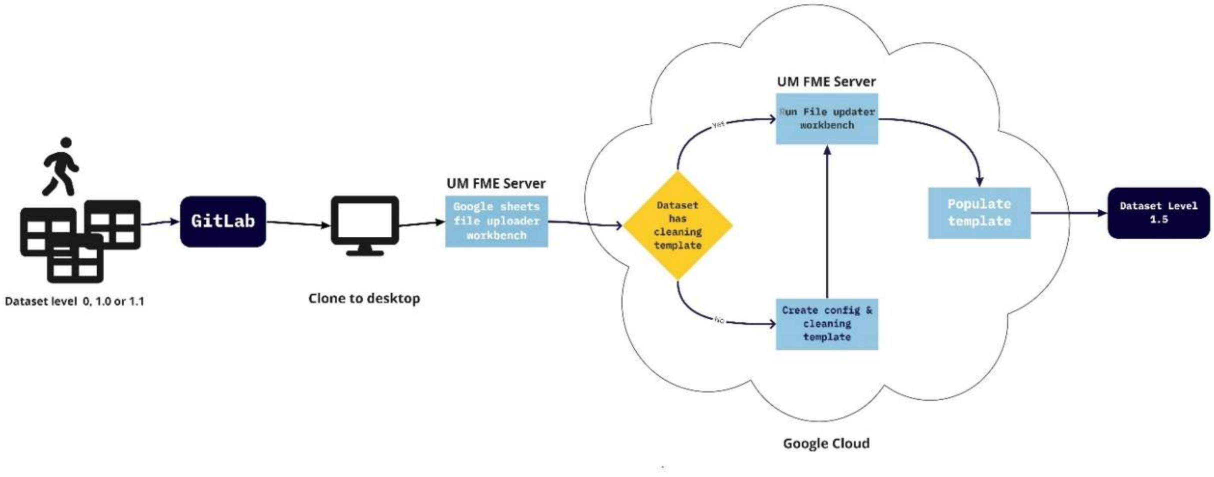

Dataset level 0, 1.0 or 1.1
GitLab
Clone to desktop
UM FME Server
Google sheets file uploader workbench
Dataset has cleaning template
Yes
No
UM FME Server
Run File updater workbench
Create config & cleaning template
Populate template
Dataset Level 1.5
Google Cloud

Image 4. Google/FME workflow

The overall workflow for CanWIN data (Image 5), including the addition of metadata standards is shown below. Sharing of data via the CKAN (Datahub) platform allows multiple information formats to be collated in one place, into a single "Data Story".

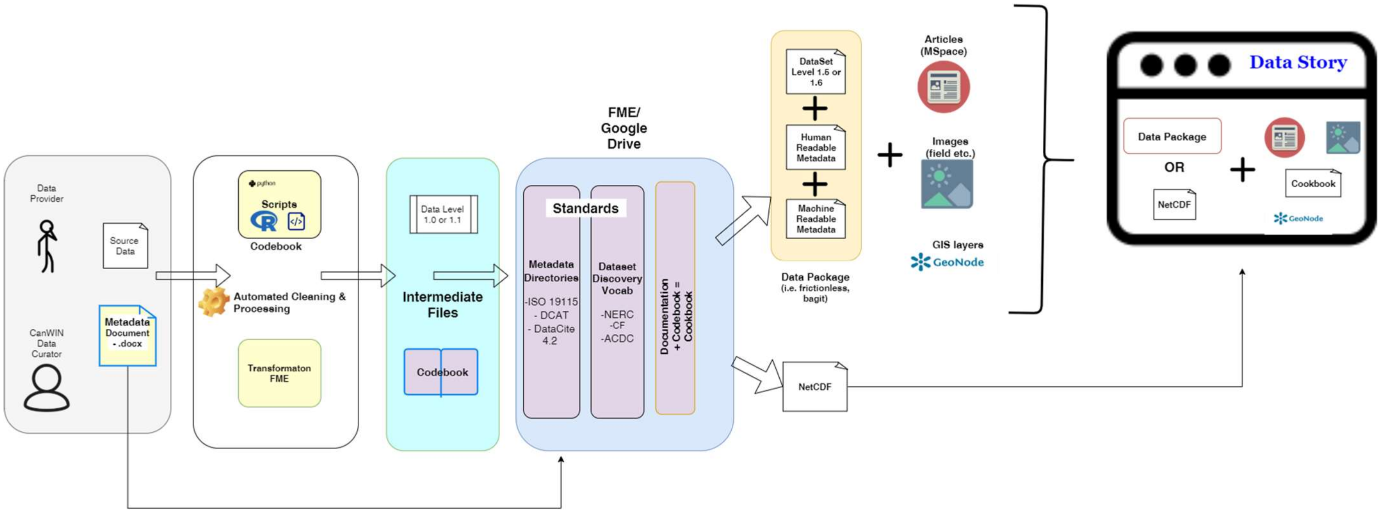


Image 5. Overview of CanWIN data sharing flow

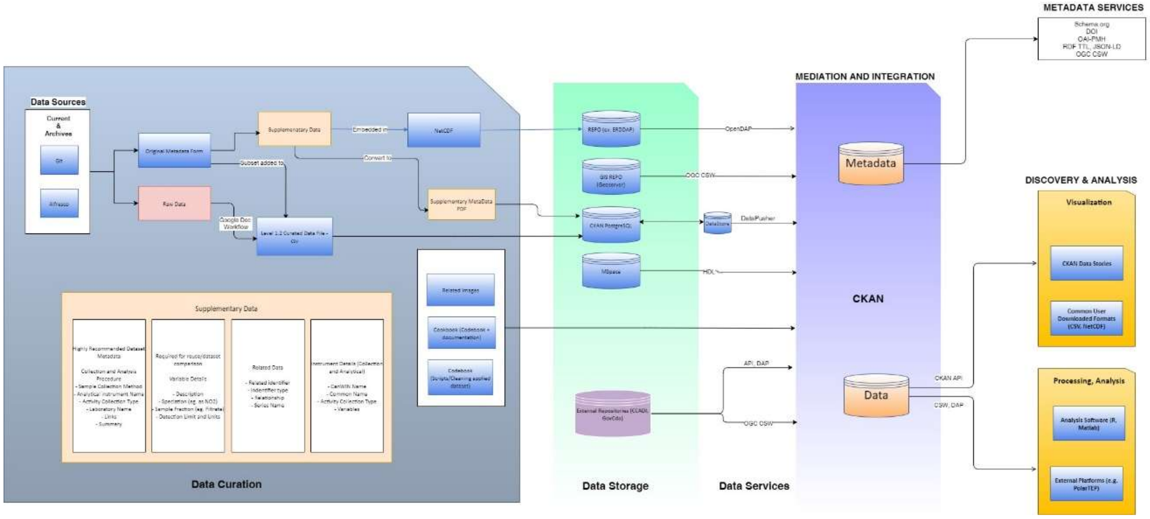


Figure 6. Overview of CanWIN Ecosystem

Figure 5 illustrates a high-level view of the CanWIN Ecosystem. Box 1 illustrates data source locations once the data is loaded into CanWIN, and the overall curation process. Box 2 illustrates the variety of data storage locations within CanWIN, Section 3 illustrates the Data Services (how systems may access the data), Box 3 shows the Mediation and Integration layer of CanWIN (how data is made discoverable to other metadata catalogues and to users and Box 4 shows the Discovery and Analysis and Visualization of datasets to users.

APPENDIX 3

CANWIN TECHNOLOGY STACK AND ARCHITECTURE

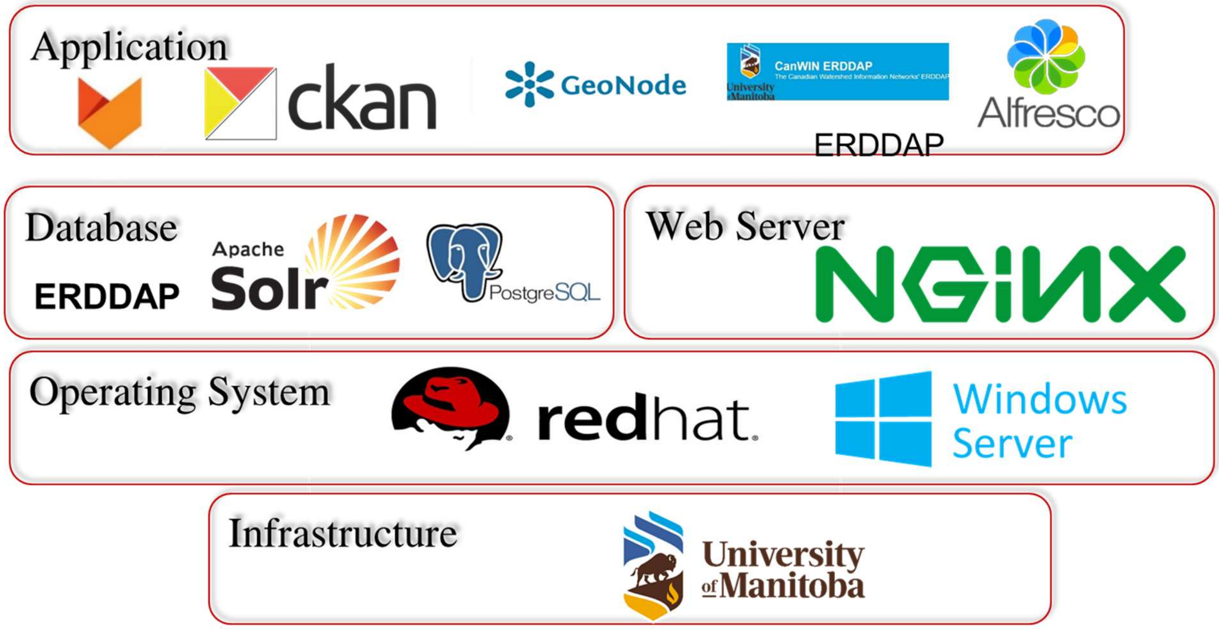


Figure 7. High-level technology stack overview

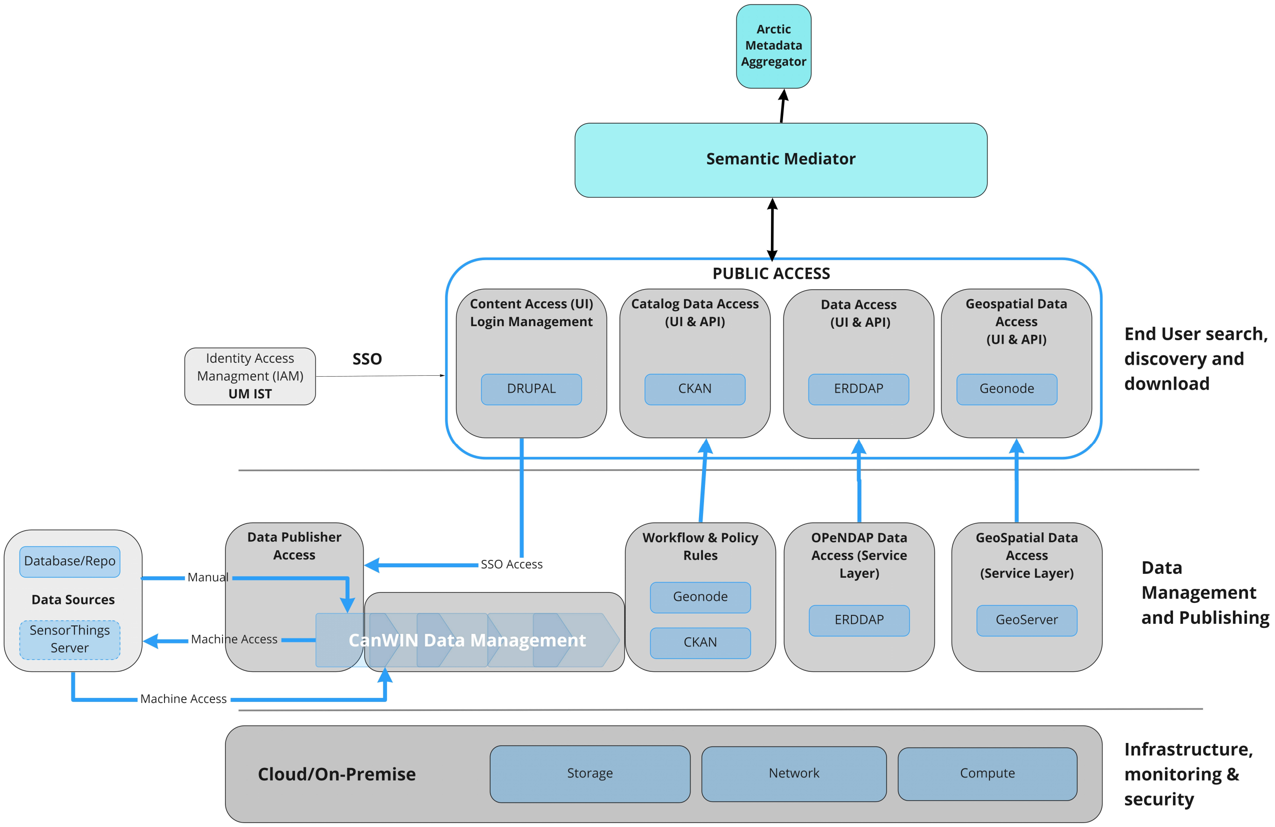

Arctic Metadata Aggregator
Semantic Mediator
PUBLIC ACCESS
Content Access (UI) Login Management
DRUPAL
Catalog Data Access (UI & API)
CKAN
Data Access (UI & API)
ERDDAP
Geospatial Data Access (UI & API)
Geonode
End User search, discovery and download
Identity Access Managment (IAM) UM IST
SSO
Database/Repo
Data Sources
SensorThings Server
Manual
Machine Access
Machine Access
Data Publisher Access
SSO Access
CanWIN Data Management
Workflow & Policy Rules
Geonode
CKAN
OPeNDAP Data Access (Service Layer)
ERDDAP
GeoSpatial Data Access (Service Layer)
GeoServer
Data Management and Publishing
Cloud/On-Premise
Storage
Network
Compute
Infrastructure, monitoring & security

Figure 8. CanWIN distributed platform and services diagram