Towards a self-serve data ecosystem

The purpose of this report is to provide a picture of what software and packaged tools within other data analytics ecosystems could aid ICTs to perform their duties throughout the research and data management tasks. In this version, an initial section provides principles and rationale for searching and choosing tools that not only satisfy the data management and analysis needs of individual teams, but also help bring consistency and harmonization to the data operated on within their workflows.

Given their widespread adoption, technologies chosen with this rationale as guidance, should ultimately lead to a sustainable, self-serve data ecosystem. Nonetheless, with the intent of allowing ICTs to pursue research however aligns best with their goals, the approach described in this report should be seen as sufficient, but not necessarily the only way ICTs could manage their information. To ensure the research portfolio is managed in a way where data outputs are reusable and discoverable, tools should be chosen not prescriptively, but in the spirit of open source, where common formats and community support create shared knowledge.

The report concludes with a list of technologies, curated and given prior assessment by the CS-DCC team to help economise these tool-choice decisions made by ICTs. This section is a living document that will be expanded during the course of the project, with high-level recommendations provided and expanded upon after feedback from ICTs.

These will come in the form of “bundles”, combinations of tool choices which interoperate well while comprehensively covering the activities performed during the data lifecycle. Given these options, ICTs will not only be able to adopt existing solutions for their team, but aide the productivity of the research portfolio on whole through consistent data and shared practices.

Key principles for choosing a self-serve data ecosystem tool.

The original proposal made by CS-DCC described a “data mesh”. To simplify discussion in this paper, the word “self-serve data ecosystem”, or simply “data ecosystem” will be used in its stead. A self-serve data ecosystem is a form of organization where producers of data are the same as its users. This leads to greater ownership over the data management lifecycle.

The result is that content of datasets is better curated and reflects more closely the shape of data required by researchers and analysts to perform their workflows. Each domain has its own bundle of datasets and acts as a product for the existing team, as well as other analysts. When datasets are exchanged between teams, the producing team becomes a resource for understanding the terms of use for this data, its content and rules for access.

Consistent metadata and documentation addresses these key concerns for negotiating data sharing, also allowing this process to become automated. Data harmonisation leverages these teams while

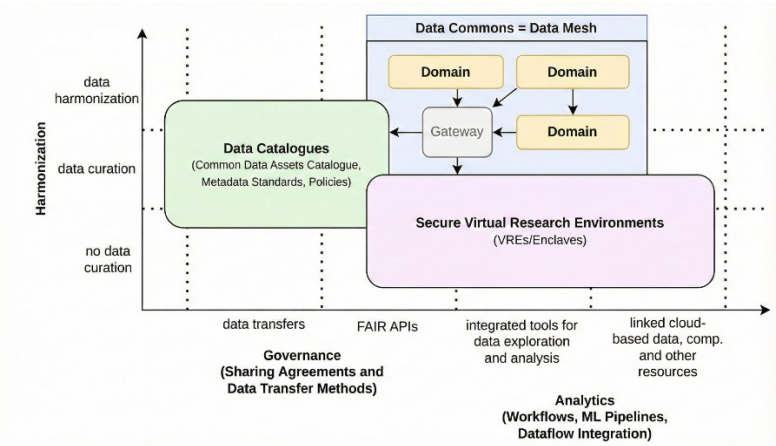
ensuring FAIRness. Within this project, this machine-operable documentation is known as a “data contract”.
To achieve this goal, these principles should drive technology choice, stated in the CS-DCC proposal:

- P1: open, inclusive RDM ecosystem encouraging partnerships from diverse domains and sectors.
- P2: using an open technology stack for the core components of the ecosystem.
- P3: data governance emphasising equitable data sharing and collective benefit sharing.

During the discovery process, the CS-DCC identified that the following features are necessary requirements for the portfolio of technologies and tools used in the ecosystem to facilitate a self-serve data management ecosystem, while satisfying the principles.

- *Interoperable standards*: ICTs develop as necessary, and use similar metadata and formats modelled through controlled vocabularies and ontologies that produce consistent annotations and APIs for datasets and resources throughout the ClimateSmart ecosystem.
- *Distributed computing*: ICTs may store their data in their own way, using their own hosting, cloud storage, filesystems, or databases, within their own control. They may bring their own analytical tools to leverage existing partnerships, or to keep compute costs within their means. They may deploy their own Agrifood-partnered data platform if desired.
- *Shared-responsibility access*. ICTs have the responsibility of deciding the access rights other groups have on their data, in a staged process designed to prevent data leakage.
- *Federated discovery*: “Wherever the data is, I may find it if I’m allowed to”. Combined with interoperability standards, it is easy for ICTs to select and trace out relevant collections of datasets that they have access to from other ICTs using descriptive metadata formats.

Nonetheless, not every team has the full expertise for every aspect of data management. CS-DCC exists to provide recommendations and training when an ICT team requires it. The technologies described later all have some connection to the CS-DCC team or their broader research community, ensuring the possibility of training and support.



Requirements and functions for tools in a self-serve data ecosystem

The next section will feature descriptions of tools assessed by the CS-DCC. They are classified by their relationship to data management tasks, as described in this section, as well as the figure above.

In the experience of experts on the CS-DCC teams, data harmonisation, and data access and privacy, are concerns which are best discussed at the beginning of a research consortium's formation, when data is expected to eventually be under FAIR access. Thus, discussion forums where these policies and reusable schema are discussed and formed. Although not considered a part of the data lifecycle directly, they augment the planning and collection phases within data management.

Task	**Description**	**Solution Forms**
Data Harmonization	Implement tools that support schema creation, annotation, and data harmonisation. This ensures that data is consistent, reusable, and interoperable.	Schema Toolchain, Registry, Dataset Annotators, Dataset Validation against schema, provenance for evolving standards.
Data Preservation	Employ solutions that prioritise long-term data storage with appropriate version control and metadata management.	Dataset Archives, either self-hosted or with wide adoption in the research community of the ICTs. Storage for both pre- and post-publication artefacts.
Data Access and Authentication	Solutions should include robust security features to control who can access data, safeguarding sensitive information.	Federated Authentication
Data Analysis	Tools should support complex data processing and visualisation needs, enabling ICTs to extract insights efficiently.	Recommended Analytics Toolchain, Self-Hosted Data Analysis Platform
Data Discoverability	Data assets produced by the ICTs are browseable and accounted for in correspondence to their place in the project portfolio's research objectives, to show visible progress and provide opportunities for disseminating key results to other scientists, and policy-makers.	Central Data Portal, CMS for Content
Workflow Reuse	Tools and practices are consistent	Common Forum, Standard

	between ICTs to allow more interoperability between their research initiatives, and a synergizing of their expertise.	Deployments, Training Programmes

Drilling down into each of these elements, there are key features and tasks software that fulfil a certain stage in the life cycle needs to have. For this report, there are two focuses:

- First, data management platforms that act both as stakeholder-facing "data portals" as well as provide a toolchain that improves interoperability and security through consistency and integrity checks of datasets against standards and fine-grained access and authentication policies.
- Next, dataset archive solutions, which work as hosting solutions for medium to large-sized datasets (hundreds of gigabytes to low number of terabytes) that improve discoverability and adoption by allowing for both private and public access, has means dataset for discoverability based on the schemas used in data harmonisation.

Appendix 1 contains a faceted comparison of the data management platforms compared, emphasising the principles.

A curation of tools

Starting even before the project proposal, the CS-DCC team has investigated numerous tools and software systems which satisfy either independently or jointly, the requirements stated in the section above. Restating that the mandate is to guide, not choose on behalf of ICTs, the systems below are presented with a brief description alongside pros/cons or other rationale on how they would help ICTs achieve their goals throughout their data management tasks.

Bento
https://github.com/bento-platform

Bento is a data management and exploration platform designed to facilitate collaborative research, particularly within -omics communities. While many larger datasets in ClimateSmart are not strictly -omics data, almost every ICT research project includes -omics components, making Bento a relevant tool for projects with a genomics emphasis.

Advantages:

- **Integration with -omics communities**: Provides a collaborative platform for projects involving genomic, transcriptomic, proteomic, or similar datasets.
- **Self-hostable deployment:** Allows institutions to deploy and manage their instances, ensuring data control and compliance. Uses Docker and Docker Swarm.

- **In-house development**: Supported by active development teams, fostering continuous updates and improvements. The Bento team has a direct relationship to the Laval University team on CS-DCC, potentially allowing this project to grow around ClimateSmart's needs.

Disadvantages:

- **Development stage**: Still evolving, which may imply potential stability issues or limited features.
- **PhenoPacket**: Bento's use of the PhenoPacket schema emphasises human -omics concepts, which might limit full applicability to broader non-human or agricultural research contexts.

Tripal
https://tripal.info/

Tripal is a content management system (CMS) tailored for the curation and display of biological and agricultural data. It is widely used among agricultural research projects, ensuring familiarity within the community.

Advantages:

- **Established platform**: Trusted and adopted by various agricultural research initiatives, fostering community support and knowledge sharing.
- **Current adoption in Climate Smart:** two ICT teams are using Tripal for their data management at time of writing (Nov 2024).
- **Self-hostable**: Provides ICTs the option to host their data on their servers, using Docker containers, ensuring data privacy and compliance with local regulations.
- **Ontology and Controlled Vocabulary Driven Data Entry and Content Management:** Tripal is developed with ontologies in mind from the ground up, ensuring that datasets which contain entities or have particular metadata can be programmatically accessed through a REST API. This is a consistent interface that provides patterns for discoverability that is FAIR.
- **Agriculture-First:** Unlike other potential data access and sharing portals listed, the Tripal platform was developed with agriculture in mind. This leads to an inherent support of modelling many entities which may be familiar to our ICTs beyond simply -omics information.
- **Connections to LinkML Team:** Tripal's ontology schema is based on ChADO, a schema developed by the same institute as LinkML which the CS-DCC team has expertise with.
- **Community Support:** The Tripal agricultural community works with many different crops including wheat, lentils and more, with frequent interactions with the Tripal development team itself, allowing this platform to potentially grow with the ClimateSmart consortium's needs.

Disadvantages:

- **Analytics support**: Native analytics capabilities are limited and reliant on third-party plugins developed within the Drupal platform, which may require additional configuration or development for novel dashboards or data views.

Indoc Pilot
https://www.indocresearch.org/pilot

Indoc Pilot is a data management, research and analytics platform developed by Indoc Research, a non-profit software consultancy firm dedicated to providing bespoke deployments for Pilot to university teams and biomedical initiatives across Canada.

Advantages:

- **Proven:** The team at Indoc Research has experience customising and deploying Pilot with at least three biomedical research initiatives covering a range of domains and security constraints, having worked with the Ontario Brain Institute, EBRAINS/The EU-funded Health Data Cloud, and the Ontario Health Data Platform.
- **"Staging room" for publishing datasets:** The Indoc Pilot platform has a "green room" feature that can be configured to stop the upload and publication of datasets.
- **Mature data access and tenancy model:** Pilot provides fine-grained access controls that can be configured on a per-team or per-role basis, and can be configured to handle these data access policies between deployments. This satisfies the multi-tenancy requirement.
- **Data validation support with JSON Schema:** The green room as mentioned above features a data validation step that leverages JSON Schema encoded metadata standards to assess the integrity and compliance of datasets, ensuring their portability.
- **Open source Kubernetes deployments:** Indoc Pilot is deployable using the container orchestrator Kubernetes, making it portable to most hardware and cloud solutions in use by the majority of users today.
- **Integrated analytics and dashboard creation:** During discovery, Indoc Research demonstrated its dashboard and analytics capabilities leveraging Apache Superset, an open-source solution. It is extensible to other modules.
- **Integrated computing environments:** Pilot can also spin up virtual environments that researchers could use to perform computing tasks while having their access rights restricted to the data sets their roles sustain. This could allow for agreements to be made between teams for data sharing, but only when using these environments, keeping the data compartmentalized.
- **Existing relationship:** Indoc Research has an existing relationship with the UGuelph part of the CS-DCC team on a different but related project.

Disadvantages:

- **High development cost:** although the deployment configurations are open-source, the complexity of the Indoc Pilot platform does require consultation from Indoc to deploy and customise features adequate to their users. The provided estimate was at least 1.5 years of work for the CS-DCC project at the time of grant submission.
- **Biomedical focus:** much like other platforms considered, the -omics focus of Indoc provides limitations when handling and storing other non-tabular datatypes that may emerge during the course of the project.

- **Vendor Lock-in:** Although Indoc Pilot is “source available”, development of Indoc would still be owned by Indoc Research, including upstream contributions made by teams within the CS-DCC if extensions were to be provided to it.

DNAStack

DNAStack provides a comprehensive platform for managing, sharing, and analysing genomics data, with features optimised for searchability and data retrieval.

Advantages:

- **Faceted data search**: Enables users to filter and navigate large datasets effectively.
- **Proven**: DNAStack hosts a large number of virus and pathogenomics datasets across many existing projects outside of ClimateSmart but with a large diversity of metadata, demonstrating its maturity as a platform.

Drawbacks:

- **Subscription-based**: Access requires a paid licence, which can limit availability for teams with budget constraints.
- **Potential migration needs**: Organisations may need to migrate data should licensing terms change or end.

FRDR (Federated Research Data Repository)

FRDR is a Canadian platform provided by the Digital Alliance, specifically designed for the storage, preservation, and sharing of large-scale research datasets. It plays an essential role in supporting the research community by providing reliable, secure data archival solutions that align with data governance standards in Canada.

Advantages:

- **Supports large data**: Optimised to handle the archival and sharing of substantial research datasets, making it ideal for ICT projects with significant data volumes.
- **Canadian data residency**: Ensures that all datasets are stored within Canada, which helps meet data sovereignty and regulatory compliance requirements.
- **Federated login**: Integrated with the **CANARIE network**, allowing researchers to use existing institutional credentials for seamless and secure access.
- **Connection with data management plans:**

Drawbacks:

- **Limited API functionalities**: Currently lacks an API for querying metadata programmatically, which can impede automated data discovery and integration into workflows.

- **Potential learning curve**: New users may need training to navigate the platform's capabilities effectively, especially when managing complex data projects.

FRDR provides a critical infrastructure component for ICT teams requiring robust data preservation and sharing solutions. While it has some limitations, particularly with metadata querying, its alignment with Canadian data policies and support for large data sets make it a valuable tool for ensuring long-term data security and accessibility.

Borealis

Borealis is a data repository platform, popular among researchers in various scientific fields, including agriculture.

Advantages:

- **Adoption in agriculture**: Well-recognized in the research community for storing and managing agricultural data.
- **API availability**: Dataverse includes API calls for streamlined data interactions and integration with external tools.

Drawbacks:

- **Custom development**: May require additional customization for advanced data handling or unique workflow needs.

Recommending "Bundles"

In the long-run, interoperability is ensured when the appropriate tools and practices common to many scientific domains are similar or shared between the data producers within the ecosystem. This becomes even more likely when there is less "friction" between tools and practices. With these

While the CS-DCC team has the mandate to guide and support the usage of software and practices chosen by ICTs, it may also recommend such tools and practices in support of the data ecosystem's long-term health. To that end, several of the technologies described above form

Archival Bundle: FRDR + Borealis

This bundle provides solutions for the secure storage, preservation, and long-term management of large-scale research datasets:

- **FRDR (Federated Research Data Repository):** A Canadian platform that supports the archival and sharing of substantial datasets, ensuring data residency within Canada and meeting data sovereignty requirements. It serves as an ideal end-point for the conclusions of different projects in ClimateSmart's research portfolio.

- **Borealis:** A data repository platform based on Dataverse, popular for its use in agricultural and scientific research. It includes APIs for programmatic access, enabling integration with other data management tools. It can be used as an intermediary platform for uploading data, and the extensible capabilities of Dataverse allows ClimateSmart to add privacy management and discoverability features in subsequent project phases.

Use Case: This bundle is perfect for ICT teams seeking reliable archival solutions that comply with Canadian data governance standards. It supports large-scale data storage and ensures long-term accessibility and preservation of datasets.

Community and Portal Bundle: DNAStack + Bento or Tripal

This bundle aims to create a connected research community and an interactive data portal for data sharing and collaboration:

- **DNAStack:** A powerful data management portal with advanced search capabilities, supporting data discovery and collaborative efforts.
- **Bento:** An open-source platform that facilitates data sharing and interactions, particularly in projects involving -omics data. It offers a self-hosted deployment for more control over data.
- **Tripal:** An alternative to Bento, suitable for agricultural research. It provides a CMS for managing and displaying biological data, widely adopted in agricultural projects.

Use Case: This bundle supports ICT teams in building community-focused data portals that encourage data sharing and collaborative analysis. It enhances the visibility of research projects and promotes shared practices, helping researchers discover and leverage each other's data efficiently. The range of choices for data hosting platforms could both give a default location for browsing project datasets (DNAStack) while giving ICTs a migration path if they want to showcase their data to their specific research community.

By adopting these bundles, ICT teams can better align their practices with those common in their respective fields while ensuring seamless data integration, enhanced collaboration, and long-term sustainability. These combinations of tools are designed to help ICT teams navigate the complexities of data management, from harmonization to archiving and community-building, fostering a unified and interoperable data ecosystem.

Conclusion

The CS-DCC team's comprehensive investigation into tools and software systems has led to a curated list that supports the full spectrum of the data management lifecycle. By considering the principles of common use, community adoption, open-source sustainability, and scalability, ICT teams are empowered to make informed choices that align with their specific research goals and requirements. These tools, when chosen and implemented effectively, can foster a self-serve data ecosystem that emphasizes data harmonization, accessibility, and re-usability.

Future Recommendations and Next Steps

The success of a self-serve data ecosystem relies on continuous assessment and iterative feedback from users. As ICT teams adopt and use the tools outlined in this report, the CS-DCC will monitor their integration and collect feedback to refine recommendations further. This report is intended as a living document, with tool bundles evolving over time to accommodate technological advancements and shifting research needs.

Looking forward, the CS-DCC team recommends:

- **Continuous Training and Support:** Ensuring that ICT teams receive adequate training on tool usage and data management best practices will be essential for maximizing the benefits of the selected technologies.
- **Feedback Mechanisms:** Creating forums or channels for ICTs to provide feedback on tool effectiveness and suggest potential improvements or new tools for consideration.
- **Expansion and Updates:** Regularly updating this curation to include emerging technologies or new versions of existing tools that could provide enhanced capabilities or address current limitations.

With these strategies, ICTs will be better equipped to achieve a sustainable, scalable, and harmonized data management approach that supports impactful research outcomes. This unified vision will aid in building a strong research ecosystem where data flows seamlessly, insights are derived efficiently, and collaborative projects thrive.

Appendix 1: Faceted Data Management Portal Comparison

Name	Metadata Standard	Open source	Versioned datasets?	Domain relevant codebase?	Reusable modules	Extract, Transform, and Load (ETL)	Harmonisation	Analytic modules	Cloud-native Deployment	Sustainment model	Access Permission Controls	Project Maturity
Tripal	ChADO	Yes	No	Yes	Yes	Yes	Yes	Yes	Docker		Yes	Mature
Bento	Phenopackets	Yes	No	No	No	?	?	Yes	Docker		?	Immature
Indoc Pilot	JSON Schema	Yes*	Yes	No	Yes	Yes	Yes	Yes	Kubernetes		Yes	Mature
DNA Stack	JSON Schema	No	Yes	No	No	No	?	?	No		Yes	Mature