Membership Maturity

Wim Hugo
November 2018
WDS Repositories Day
IDW, Gaborone, Botswana
Responsible Citizenship

Open research output is the tax of the world of science.

Metadata is a *tax return*.

It is not always pleasant to pay, or fun to complete your return, but citizens cannot expect benefits (infrastructure, services) if they do not pay their taxes.

You can, of course, and ask a *consultant* to file your return.

And *crowdfund* your tax.
Review of the Maturity Landscape
There is a wide variety of considerations that have a bearing on creation of any Research Data Infrastructure, and by extension, an ILTER infrastructure. These can be grouped as follows:

- **Science concerns**: questions posed by science, which, in turn, finds expression in a collection of **standard variables**, which if observed at the correct temporal and spatial scale, over an adequate period of time, and using sufficiently aligned protocols, can inform analysis and hypotheses. In addition, **Open Science** promotes the reproducibility and veracity of science;

- **Governance concerns**: largely finding expression in open data considerations, by extension requiring all tax-funded research output to be openly and freely available unless some specific limitations apply, resulting in policies, best practice guidelines, and ways of working, sustainability and trust;

- **Informatics concerns**: finding expression in considerations such as reference models, information models, vocabularies, standards and specifications, and architectures.
Some More Considerations

- Accredited Trusted Repositories
  - Minimum requirements for digital object and data management, including workflow, metadata, and preservation
- Need to Integrate with International Networks and Infrastructures
  - Standard protocols,
  - Conventions, good practice, and metadata requirements
    - Data Families
- Global e-Infrastructure Guidance
  - Standardisation of services, vocabularies, and practices
  - Increasing Reliance on Persistent Identifiers
- Funder imperatives
  - Data Management Plans
    - Metadata ‘inception’
  - Proof of Deposit
    - Persistent identifier(s)
    - Link to a Trusted Repository
- Open Science
  - Reproducibility and Re-Use
  - Liberal Licenses
- Automation

Given these guidelines and constraints, we need to define and build systems and procedures that are as efficient and streamlined as possible, taking stakeholder-specific requirements into account.
A typical WDS Member needs to align with and obtain inputs from a number of global initiatives and established community consensus. This includes, *inter alia*:

- The [FAIR](https://fairdata.org) data principles;
- [GEO](https://www.geohub.org) Infrastructure, the [GEOSS Data Management Principles](https://www.geohub.org/glossary), and supporting [Implementation Guidelines](https://www.geohub.org/glossary);
- ICSU-World Data System guidance on certification as a Trusted Repository, developed in conjunction with [DSA](https://www.era-dsa.org) and governed by the recently constituted [CoreTrustSeal](https://coretrustseal.org) Board;
- Published and emerging guidance within the broader [Research Data Alliance](https://www.rda.info) collaboration, for example
  - Dynamic data citation;
  - [WDS-RDA Working Group on Data Fitness for Use](https://www.rda.info/working-groups/working-groups/working-group-95).
- Emerging practice in respect of [Essential (Standard) Variables](https://www.rda.info/working-groups/working-groups/working-group-29).

The typical stakeholder groupings are **Individual Researchers**, their associated **Institutions**, Initiatives or Projects, **Repositories** and Data Centres, and **Community Initiatives** and **Global Authorities**. For each of these groupings, it is possible to define sets of guidance and best practice that are collectively referred to as the basis of **Responsible Citizenship of the Data World**.
Generalised Use Cases: Architecture

Authenticate

Curate

Data/Services

“Publish”

Metadata

“Find”

Mediate

License

“Bind”

Cite

Access/Download

Analise/Visualise

Process

Assess/Rate

“Find”

Discovery
Generic Use Cases: Any e-Infrastructure

• Registration and Authentication
• Publication of Data (“Deposit”)
  – Ingest, accession, and archiving (original state)
  – Confirmation of rights, licenses
  – Minimum metadata (more on the next slide)
  – Standardise, format, and publish as a service
    • Publish as a data paper with extended abstract
  – Persistent Identifier linked to metadata
• Discovery
• Application
  – Query/Subset/Download
  – Inclusion into online processes
  – Visualisation and Exploration
  – Inclusion into VREs - notebooks, etc.
• Rating and Annotation
• Curation
  – Follow up on annotation, feedback, comments
  – Format preservation and fixity checks
1. **Citation Metadata**: information related to the owner of the work, the title, and publication details - the minimum required to publish a citable reference to the work.

2. **Discovery Metadata**: additional information, largely related to the coverage of the work. These metadata include spatial, temporal, and topic coverage (by way of keywords and linked concepts), and/or a narrative (abstract) from which coverage can be understood. The former are usually machine-readable, while the latter is not.

3. **(Re)Usability Metadata**: additional information covering aspects such as a license, protocols, provenance and lineage, a data dictionary, ratings and quality measures, and other information that may be of value to the end user.

4. **Administrative Metadata**: information on curation, workflow states, hosting institution and conditions, fixity, and similar.

1. **Contextual Metadata**: *Site-level and protocol/essential variable context.*
   - Example: DEIMS
State of Variables and Metadata

• “Raw Data” - data for which no quality assurance, correction, or other form of preparation has been done, but discovery metadata may be available;

• “Publication-ready Data” - for which appropriate quality assurance has been done, and citation, administrative metadata is additionally available;

• “Analysis-ready Data” - for which usability metadata is available, and community conventions in respect of vocabulary (semantic), and structural (schematic) interoperability conventions have been applied. Typically, portfolios of Essential Variables fall into this category.

• “Indicator-ready Data” - for which in addition to the above, syntactic (service) interoperability is also available and data can be included in trans-disciplinary indicator combinations. In such cases, it is very useful if semantic annotation is applied to contextualise the application of one or more Analysis-Ready data sources in the derivation of an Indicator.
The main focus of efforts such as FAIR and GEO is to define the ‘**Performance Parameters**’ under which data (and by extension any other research output) will be optimally produced and re-used.

- Parameters: Usability, Accessibility, Ownership, Stewardship
- Superimposed on these are ‘**Concerns**’
Scope of Design Considerations for RDI

Development of appropriate portfolios of Essential or Standard Variables that allow integration of data across temporal and spatial scales, across biomes, and across regions.

- Existing and emerging portfolios to be taken into account (EBVs, Ocean Variables, Climate Variables, Fluxes, ...)
- Elements of scientific relevance, protocol, semantics and vocabulary
- Define appropriate ‘Data Families’ and associated interoperability standards
- Consider mapping to global indicator portfolios (SDG, Sendai, Aichi, ...)
- Formalise ‘State of Readiness’
- Critical for transdisciplinary research
• “Raw Data” - data for which no quality assurance, correction, or other form of preparation has been done, but discovery metadata may be available;

• “Publication-ready Data” - for which appropriate quality assurance has been done, and citation, administrative metadata is additionally available;

• “Analysis-ready Data” - for which usability metadata is available, and community conventions in respect of vocabulary (semantic), and structural (schematic) interoperability conventions have been applied. Typically, portfolios of Essential Variables fall into this category.

• “Indicator-ready Data” - for which in addition to the above, syntactic (service) interoperability is also available and data can be included in trans-disciplinary indicator combinations. In such cases, it is very useful if semantic annotation is applied to contextualise the application of one or more Analysis-Ready data sources in the derivation of an Indicator.
Scope of Applicable Variables

<table>
<thead>
<tr>
<th>Essential Climate Variables</th>
<th>Essential Biodiversity Variables</th>
<th>Essential Ocean Variables</th>
<th>Anthropic Factors</th>
<th>Other Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above-ground biomass</td>
<td>Genetic Composition</td>
<td>Ocean Surface Heat Flux</td>
<td>Land use/land use change</td>
<td>Topography</td>
</tr>
<tr>
<td>Albedo</td>
<td>Species Composition</td>
<td>Sea Level</td>
<td>Human population</td>
<td>Surface roughness</td>
</tr>
<tr>
<td>Fire</td>
<td>Species Populations</td>
<td>Sea Surface Temperature</td>
<td>Economic development</td>
<td>Crop yield</td>
</tr>
<tr>
<td>FAPAR</td>
<td>Species Traits</td>
<td>Sea State</td>
<td>Livestock population</td>
<td>Ground/soil heat flux</td>
</tr>
<tr>
<td>Glaciers</td>
<td>Community Composition</td>
<td>Sea Surface Salinity</td>
<td>Ecosystem/agricultural management</td>
<td>Soil type</td>
</tr>
<tr>
<td>Groundwater</td>
<td></td>
<td>Subsurface Currents</td>
<td></td>
<td>Soil quality/health</td>
</tr>
<tr>
<td>Ice sheets and ice shelves</td>
<td></td>
<td>Subsurface Salinity</td>
<td></td>
<td>Dissolved organic/inorganic carbon (terrestrial)</td>
</tr>
<tr>
<td>Lakes</td>
<td></td>
<td>Subsurface Temperature</td>
<td></td>
<td>Atmospheric/Planetar y Boundary Layer</td>
</tr>
<tr>
<td>Land surface temperature</td>
<td></td>
<td>Surface Stress</td>
<td></td>
<td>Atmospheric nitrogen deposition</td>
</tr>
<tr>
<td>Latent and sensible heat fluxes</td>
<td></td>
<td>Inorganic Carbon</td>
<td></td>
<td>Infiltration (hydrology)</td>
</tr>
<tr>
<td>Leaf Area Index</td>
<td></td>
<td>Nitrous Oxide</td>
<td></td>
<td>Runoff (hydrology)</td>
</tr>
<tr>
<td>Permafrost</td>
<td></td>
<td>Nutrients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>River Discharge</td>
<td></td>
<td>Ocean Color</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snow</td>
<td></td>
<td>Oxygen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil Carbon</td>
<td></td>
<td>Transient Tracers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil Moisture</td>
<td></td>
<td>Marine Habitat Properties</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precursors</td>
<td></td>
<td>Plankton</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(supporting the Aerosol and Ozone ECVs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud properties</td>
<td>Anthropic GHG emissions</td>
<td>Below-ground biomass</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ozone</td>
<td>Anthropogenic water use</td>
<td>Natural GHG flux</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precursors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Trust

Trust is at the centre of the process of science, and increasingly important in grant-funded research. Extensions beyond data to publications, software, vocabularies, ...

- Network-level membership of ICSU World Data System
  - Accreditation of network members
  - Individual certification of data centres and repositories using CoreTrustSeal
- Dataset-level quality assurance and metrics in collaboration with GEO, RDA
- Potential future implications of BlockChain technology for trust and provenance
- Critical for automation
Open Science is driven largely by a desire to make science more useful and to improve **validity**

- From RDI perspective, policy is not important, licenses are
- Embargo periods and common sense
- Creative Commons Licenses preferable
  - Machine-readable exceptions to CC Licenses
- Legal Interoperability Framework needed
- Again, critical for automation
Identity is increasingly important for the actors and systems in Research Data Infrastructure, and underpins necessary constraints on the Semantic Web.

- Promotion and use of Persistent Identifiers
- Federated, single sign-on - EduGain/ ORCID
- Agreement and guidance on vocabularies
- Registries of important elements
  - Especially Sites
- Citation of dynamic data sets
- Potential BlockChain implications for Identity and Federation
ICSU-WDS Knowledge Network: the Fabric of Science

- TDRs (WDS, DSA, DataCite*)
- Scholarly Publications (CrossRef)
- Samples and Events
- Coverage (Temporal, Spatial, Topic)
- Data Citations (DataCite)
- Use, Caveats, Lineage, Methods
- Licenses (CoDATA, Creative Commons)
- People (ORCID)
- RDI Outputs/Online Resources
- Projects
- Initiatives
- Networks
- Institutions (?)
- Funders (Fundref, ...)

* Including re3data, DataBib

Scholix Initiative: CrossRef/ DataCite cross-linking
Interoperability

Interoperability is directly linked to readiness of data and variables, automation and machine-readability, and application outside discipline-specific silos.

- Development of essential variables for semantic interoperability
- Agreement on metadata and data schema, services based ‘Data Family’ approach
- Brokering Framework - extending Scholix
- Loosely coupled architecture required ...
- Automation must always be in mind
Practical Steps
We need to recognise that in a network, there are stakeholder groupings that have differentiated motivations and responsibilities, tied together in an Evidence Value Chain:

<table>
<thead>
<tr>
<th>Stakeholder Group</th>
<th>Typical Motivation/ Incentive</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Researcher</td>
<td>Grant conditions, citation</td>
<td>Pay your taxes!</td>
</tr>
<tr>
<td>Institution</td>
<td>Visibility, societal benefit</td>
<td>Enabling policies, funding</td>
</tr>
<tr>
<td>Repository</td>
<td>Usage statistics, certification</td>
<td>Trust, curation</td>
</tr>
<tr>
<td>Network</td>
<td>Scale effects, societal benefit</td>
<td>Standardisation, guidance</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>Optimisation, continuity</td>
<td>IRL, knowledge retention</td>
</tr>
<tr>
<td>Society</td>
<td>Improved Decisions</td>
<td>Feedback and Needs</td>
</tr>
</tbody>
</table>

Evidence Value Chain

Publication-Ready >>> Analysis-Ready >>> Indicator-Ready >>> Decision-Ready
Roadmap Process Overview

1. Governance, Strategy and Policy
2. Structure, Policy, Architectures
3. Infrastructures, Systems, Workflows, and Procedures
4. Implementations, Components, Services, Tools
5. Applications Objects, Outputs, Artifacts

Increasing Specialisation
For each level of detail ...

- Develop a GRDI “Body of Knowledge”
- Web-based resource supplemented by review paper(s)
- Concept of “Candidates”
  - Strategies and Governance Models
  - Policies and Licensing
  - Architecture, Standards and Specifications Portfolios
  - Implementation Guidelines
- Perspectives
  - Individual Researchers
  - Institutions
  - Repositories and Data Centres
  - Initiatives and Networks
  - Communities
- Endorsement by initiatives, networks, institutions, ...
## Typical Outputs

### Annexure A: General Considerations

<table>
<thead>
<tr>
<th>Concern/Aspect</th>
<th>Guidance: ★ Best practice implementation for WDS Membership</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★ Recommended Focus</td>
<td></td>
</tr>
<tr>
<td>For Individual Researchers</td>
<td>For Repositories, Institutions and Projects</td>
<td>For the Community/ WDS</td>
</tr>
<tr>
<td><strong>Open Science: Open Access, Policy, and Licenses</strong></td>
<td>Select open licenses, preferably the most open suitable Creative Commons license possible (CC0, otherwise CC-BY or, if necessary CC-BY-NC – avoid the ND, no-derivatives and SA, share-alike options), for all published data sets unless one of a specific set of exceptions apply.</td>
<td>★ Develop data policies in support of open access and open science, and standardise on Creative Commons licenses for all but specific exceptions. Ensure that all datasets are published with a machine-readable license.</td>
</tr>
<tr>
<td><strong>Identity: Federated Identity</strong></td>
<td>Make use of globally available resources in this regard, such as EduRoam</td>
<td>★ If available, use EduRoam as a basis for service and system authentication, and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Additional guidance to ensure sustainable growth, access and discovery of repositories;
- There is sufficient funding for training and development. 
- The repository has a designated, full-time staff and management with training and development.
- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.

- Established procedures for access and usage of the repository are clearly communicated.

- Regular reporting on performance metrics is provided to relevant stakeholders.

- The repository has a designated, full-time staff and management with training and development.

- Relevant KPIs (e.g., Designated Repository Committee membership, data access, etc.) are integrated into the repository’s governance.

- Regular performance reviews and feedback mechanisms are in place.
Engineering “Bodies of Knowledge”