Global Registries Initiative: Conceptual Overview

1. What

The aim of the Global Registry Initiative (GRI) is to enable a global cross-discipline view of research collections by establishing an international register of interoperable research collection registries. This register with its accompanying framework of interoperability agreements is designed to enable the global aggregation or federation of the descriptions of collections of research materials held within these registries. This aggregated view is intended to facilitate the discovery and reuse of research data collections on a global scale.

The global registry initiative is restricted to the enabling framework and infrastructure for the aggregation and federation of research collection descriptions held in participating collection registries. GRI will not itself aggregate any records (some of its members certainly will). Third parties create discovery portals, for example, to view or query the aggregated record set of a particular collection type.

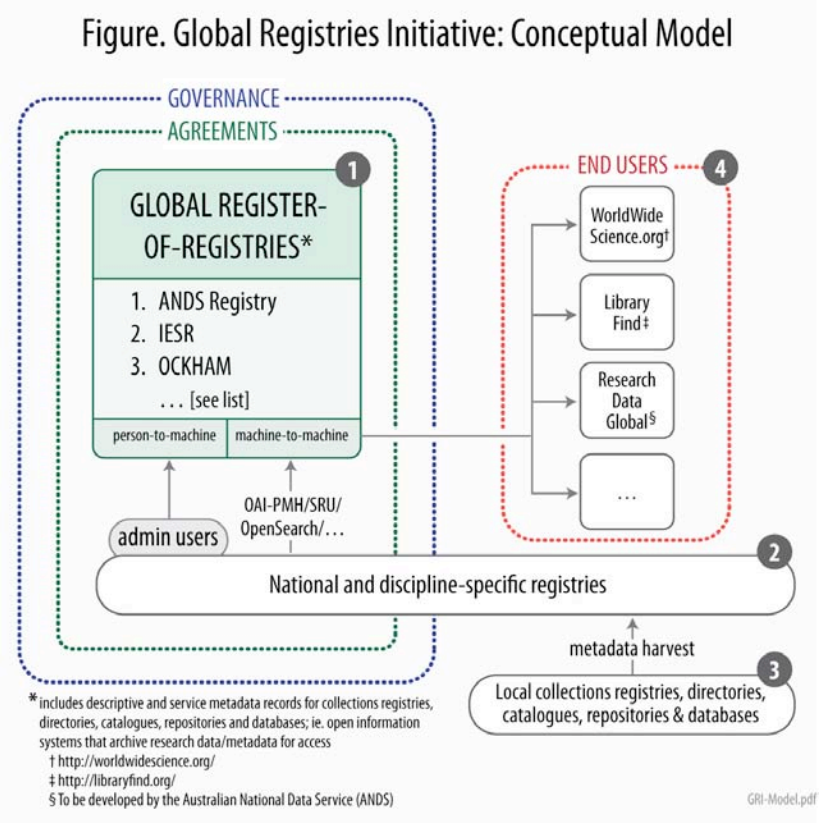
The global registries initiative will be comprised of:

1. a set of interoperability agreements around record harvesting protocols, query interfaces, and shared collection description informatics
2. an authoritative machine actionable list (register) of GRI compliant registries with interface details
3. a governance system for maintaining both the above

The objective is to provide a robust aggregation and federation platform enabled by some simple but fundamental agreements and maintained by appropriate governance.

2. Conceptual model

Figure. Global Registries Initiative: Conceptual Model



The conceptual model above illustrates a federated network of registries that operates on a 'hub-and-spoke' network topology[1]. The 'hub' is a reference register of participating GRI collection registries with all the information required to federate or harvest from them collection descriptions. The 'spokes' are the registries et. al. that are capable of operating as OAI-PMH data providers or federated search targets (e.g. Ockham, IESR, and ANDS).

The functional elements of the conceptual model are:

1. **Register-of-registries**: A simple registry instance that lists in machine readable format the registries that conform to the agreements established by the GRI; it is designed solely for the use of registry administrators
2. **International, national and discipline-specific collection registries**: Registries, and other appropriate information systems, that gather collection and services metadata from research collections or downstream metadata aggregators. In this model these registries are responsible for gathering collection descriptions from their given 'jurisdiction', carrying out the agreed upon level quality assurance for metadata, and exposing this information via GRI-compliant interfaces and information models. In this model, these registries are typically—though not necessarily—implemented by peak research bodies and groups (e.g. ANDS in Australia and JISC in the United Kingdom)
3. **Local collections registries, directories, catalogues, repositories and databases**: Registries, and other appropriate information systems, that hold data collections or aggregate collection descriptions. Ideally they will supply collection descriptions to a collection registry via some automated harvest or 'get' procedure, but equally these descriptions may be manually drawn up by either party. In this model, repositories and registries in this category are assumed to be not fully GRI-compliant (in information models or transfer protocols); rather, a collection registry from category #2 above undertakes to gather information about them. In this model, this category of systems is typically an institutional, or discipline-owned, data facility, repository or portal.
4. **End-user data access points and discovery portals**: These are independent information systems that access machine-to-machine services information from the register-of-registries (#1 above) to provide new 'products'. In this model, we assume that an independent discovery service would be systems like LibraryFind, WorldWideScience.org, or Research Data Global[2]. They may be more focused discipline specific portals that would use the GRI platform to power a 'See also' type search function to enhance local search experience. These systems are—at least theoretically—third-party down-stream systems totally independent of the GRI and its agreements and governance. Nevertheless the anticipated scenarios and use cases of this category of systems drive the information model agreements of GRI. Thus, the GRI is a federation and aggregation platform designed to empower this category of system.

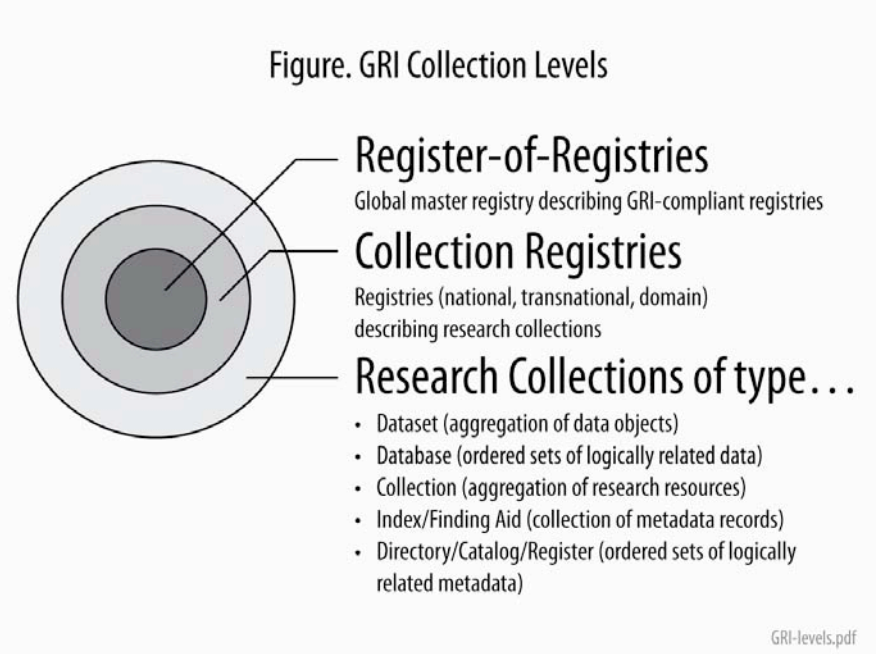
Collection types and definitions

All elements of this model are 'collections' or 'aggregations' of one form or the other (see figure below). For example, the register-of-registries (#1 above) holds collection and service metadata records for GRI-compliant registries. These GRI-compliant registries hold metadata records about 'research collections'; however, these are classified according to specific 'collection types'. The names and definitions of these collection types are to be determined. The diagram below use indicative names and definitions.

---

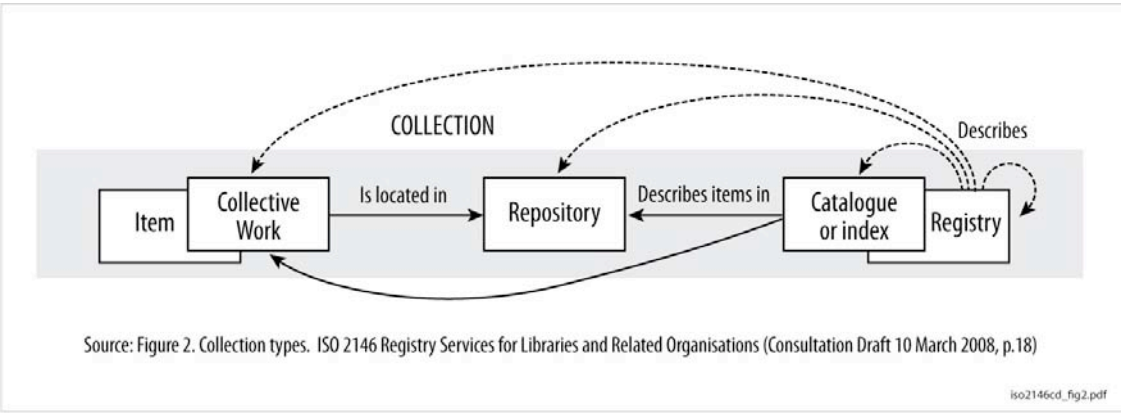
[1] Note that the 'federated' hub-and-spoke network typology does not preclude other approaches to federating systems.

[2] http://services.ands.org.au/global/orca/rda/



The ISO 2146 (Registry Services for Libraries and Related Organisations) standard, for example, defines four broad types of collections:

- **collective work**: compiled content created as separate and independent works and assembled into a collective whole for distribution and use
- **repository**: collection of physical or digital objects compiled for information and documentation purposes and/or for storage and safekeeping
- **catalogue or index**: collection of resource descriptions describing the content of one or more repositories or collective works.
- **registry**: collection of registry objects compiled to support the business of a given community



The figure above illustrates the relationship between the four listed collection types in ISI2146. Items and collective works, such as datasets, are located in repositories. Catalogues or indexes describe the content of collective works and repositories. Registries describe collective works, repositories, catalogues or indexes and other registries.

The DCMI Collection Description Type Vocabulary defines three sub-types of catalogue or index: analytic finding aid, hierarchic finding aid and indexing finding aid.

3. How

The GRI will be established in three phases. Note that each phase is path dependent and failure to achieve goals may invalidate the initiative.

Phase 1

- Establish a working GRI register of registries with initial three entries
- Develop pilot discovery service(s) (Research Data Global, LibraryFind);
- Develop and document management policies and data exchange protocols and formats; and
- Develop a plan for the long-term governance and sustainability of GRI and a strategy to achieve it by phase 3.
- Develop a communications plan

Phase 2

- Extend the number of data providers to the global registry instance;
- Develop Web applications that showcase the use of GRI;
- Action the long-term governance and sustainability strategy and plan; and
- Action the communications plan;

Phase 3

- Transfer governance and IT assets to a international governing body (see section 6)

4. When

Phase 1: Q1 2010

Phase 2: Q2-4 2010

Phase 3: Q1-4 2011

5. Proposals

ANDS will fund and resource the 'hub' GRI infrastructure for Phase 1 & 2

6. Who

ANDS, IESR and OCKHAM are the core sponsoring groups of GRI.

As previously established, the core group will need to develop a strategy and plan to extend its membership. The aim here will be to approach the leading science governance organisations, such as the International Council for Science (ICSU) and the International Council for Scientific and Technical Information (ICSTI) to find an organisational 'home' for GRI.

7. Why

Better science

The normative conventions of research practice require that the research community should have free and open access to research data in order that they can assess the validity of the claims made by other researchers or research groups. Thus, not only is it essential that research data is publically available for scrutiny, but it must also must conform to the standards expected of archival 'registries' if scientific fraud and abuse are to be avoided. The development of registries, such as GRI, are therefore essential to the practice of good science.

Better discovery

There is currently no central information directory (or system-of-systems) for research data collections, nor is there a central registry of their machine-to-machine services related to them. The GRI would enhance scientific collaboration and information sharing by aggregating this information

from the pool of authoritative international and national registries, repositories and so on, and making it available to the public.

Better collaboration

A feature of the GRI is that our target metadata protocols bind machine-to-machine service information to a collection record, thus making collection information and resoucres available to third-party standalone and web software applications. This feature has yet to be fully utilised; however, given the popularity of Web 2.0 software, the GRI could support a wide range of innovative third-party standalone and web software applications that are oriented towards research use.

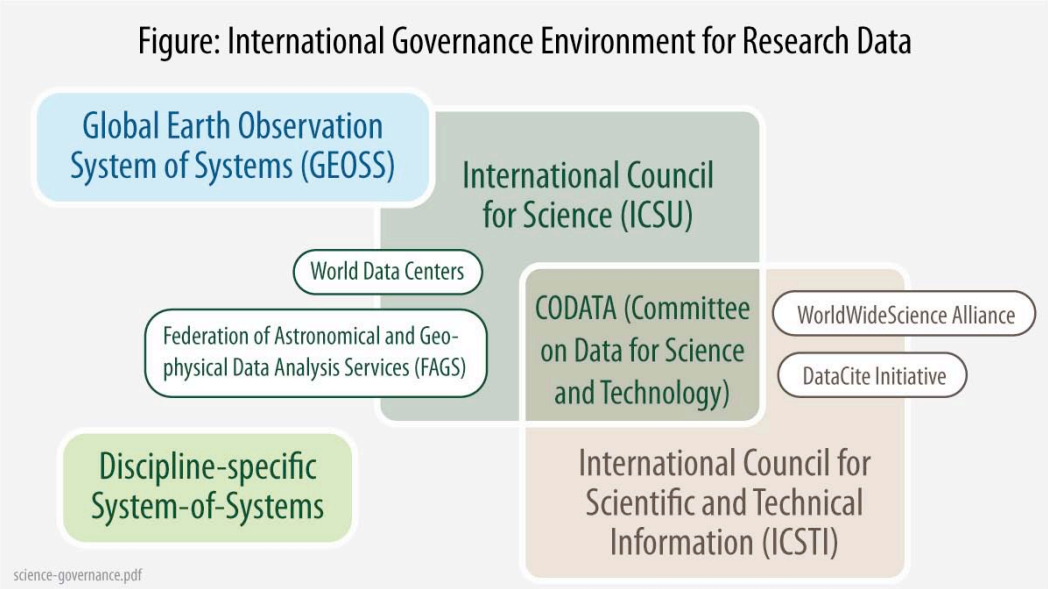
8. Issues

1. Interim infrastructure for GRI register
2. Interim infrastructure for GRI discovery portals
3. Interim governance group
4. Technical agreements
5. Process and systems for updating the register
6. Process for updating the GRI technical agreements
7. Whom should GRI approach as the governing body, and how?
8. What is the minimum acceptable and optimal data interchange formats and protocols for the GRI?
9. The GRI represents an ‘orthodox’ archival approach to the ‘authority control’ of records; however, to what extent should the GRI interoperate with Web 2.0 approaches? More specifically, how should it interact with the LinkedData initiative, or if at all?

9. Appendix A – Background on International Science Bodies

The are significant issues that need to be addressed before a home for the GRI can be found. The key issue is that the peak organisations responsible for research data and information are administratively fragmented and are often, unintentionally, working at cross purposes (see figure below) .

Figure: International Governance Environment for Research Data



The peak international governing body for science is ICSU, which includes the Committee on Data for Science and Technology (CODATA) — the key policy and operational group responsible for scientific data and information (see below).

Taken together, ICSU and CODATA are responsible for the World Data Centres (WDS) and Federation of Astronomical and Geophysical Data Analysis Services (FAGS). The World Data Centres and FAGS vary a great deal in the type and quality of data services they offer to the research community. Because some of the WDS/FAGS sub-systems are somewhat dated and are being overtaken by other approaches, there is now general agreement within ICSU/CODATA that the entire system needs renewal (see ICSU Ad hoc Strategic Committee on Information and Data 2007, 2008). Despite the enthusiasm for change voiced from within CODATA, it has a relatively small secretariat (two people) and limited financial membership funds at its disposal.

The lack of resources available to CODATA needs to be contrasted with GEOSS (Global Earth Observation System of Systems), which has a large secretariat and ample funding through the G8 'heads of government' mechanisms. Although GEOSS and CODATA are coordinating their efforts (CODATA 2007; GEO 2009), their respective capacities to build new information infrastructure systems, and operate them sustainably, are markedly different. Indeed, the scale and forward logistical momentum of GEOSS could 'crowd-out' other system-of-systems services.

Added to this mix, is the active role now being played by the International Council for Scientific and Technical Information (ICSTI). ICSTI has the advantages of a wider membership base that ICSU and appears to be more agile and responsive to the needs of it members. For example, it has established the WorldWideScience Alliance, which in turn has established a discovery portal for research information named worldwidescience.org (funded through the US Department of Energy, Office of Science and Technical Information).

Finally, to complicate matters further, there are number of systems-of-systems being developed by peak science groups and government agencies that are outside the jurisdiction of ICSU/CODATA.

These systems are typically international in scope and are designed to meet the needs of specific research community and/or address thematic global issues (e.g. climate change, national security, economic development).

Unfortunately, some of these initiatives operate in isolation and appear to have little or no peripheral awareness, nor internal mechanisms to address opportunities to collaborate if they arose. It should also be noted that these 'system-of-systems' have largely proliferated a result of national programs to build the capacity of national innovation systems around data-intensive science and technologies (or cyberinfrastructure). As a consequence of the relatively generous funding available through these programs, many of these initiatives have energetically recruited myriad partners and stakeholders, leading to very 'tangled' governance and participation structures. Such tangled structures make it very difficult to track the potential members of the GRI and assess their readiness and capacity to contribute.

This analysis suggests that the GRI will find it challenging to establish a permanent home and to get the buy-in from competing 'system-of-systems' initiatives. This 'risk' needs to be openly acknowledged and addressed in the GRI governance strategy and plan.

Bibliography

CODATA. 2007. *White paper on the GEOSS Data Sharing Principles (Review Draft)*. Paris. CODATA. http://earthobservations.org/documents/dsp/Draft%20White%20Paper%20for%20GEOSS%20Data%20Sharing%20Policies_27Sept08.pdf

GEO. 2009. *Implementation Guidelines for the GEOSS Data Sharing Principles*. Group on Earth Observations. http://earthobservations.org/documents/excom/ec16/13_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles.pdf

ICSU Ad hoc Strategic Committee on Information and Data. 2007. *Strategic Plan 2006-2011: Strengthening International Science for the Benefit of Society* Paris. International Council for Science. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/863_DD_FILE_ICSU_Strategic_Plan.pdf

———. 2008. *Final Report to the ICSU Committee on Scientific Planning and Review*. Paris. International Council for Science. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf