DATA INFRASTRUCTURE IN GLOBAL SOUTH SCIENCE SYSTEMS

INSTITUTIONAL PATHWAYS AND STRATEGIC CHOICES: TECHNOLOGY PROFILE ON DATA STORAGE AND SHARING

CENTRE FOR SCIENCE FUTURES | International Science Council

To cite this report: International Science Council (May 2026). *Data Infrastructures in Global South Science Systems.*

Funding acknowledgement: This work was conducted with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada. The views expressed herein do not necessarily represent those of the IDRC or its Board of Governors.

Authors: Venkat Nadella, Chandan Nagarajappa

Reviewers: Meredith Goins, Reyna Jenkyns, Daniela Santos Oliveira

Project coordination: Felix Dijkstal, Vanessa McBride

Project chair: David Castle

Design: Mr Clinton

Cover photo: SweetBunFactory

About the International Science Council

The ISC is an international non-governmental organization with a unique global membership that brings together 250 international scientific unions and associations, national and regional scientific organizations, including science academies, research councils, regional scientific organizations, international federations and societies, and academies of young scientists and associations.

The ISC works at the global level to catalyse change by convening scientific expertise, advice and influence on issues of major importance to both science and society.

Executive summary

Data storage and sharing infrastructures underpin contemporary science, yet their institutional configuration varies substantially across science systems. In the Global South, this variation is not shaped by the availability of technology but by institutional mandates, fiscal capacity, governance authority and regulatory environments. The modal practice remains informal: local hard drives, commercial cloud, where budgets allow, and data shared by email or external media. A study that catalogued this baseline would largely reproduce well-known patterns. This profile instead examines the institutional conditions under which some science systems move beyond ad-hoc practices toward more structured infrastructure – and what that transition requires.

Drawing on primary interviews, questionnaire responses and desk research across eight case studies spanning Africa, Asia and Latin America, the profile identifies four recurring institutional pathways through which science systems configure data storage and sharing infrastructure. Each pathway is defined not by geography or domain, but by the primary constraint it addresses and the institutional configuration that emerges in response. The pathways are not a hierarchy or a developmental sequence; they are distinct strategic configurations, each carrying characteristic trade-offs.

Cross-cutting analysis identifies enabling conditions that recur across pathways: a mandate or policy anchor that legitimizes sustained investment; institutional champions embedded within governance structures; adopted standards that enables interoperability without requiring centralization; and funding continuity that outlasts project cycles. These conditions are more decisive than technological sophistication alone.

The profile examines forward-looking considerations for science-system planning – including artificial intelligence-driven data management, federated learning, evolving data sovereignty regimes and cloud-institutional sustainability – before concluding with implications for four stakeholder audiences: science-system leaders, national research funders, international partners and regional coordination bodies.

Introduction

Data infrastructure as a foundational science capability

Data storage and sharing infrastructures have become foundational to contemporary scientific practice. Across disciplines – from genomics and climate science to astronomy and agriculture – research increasingly depends on the generation, preservation, analysis and reuse of large and heterogeneous datasets (Hey et al., 2009). The growing application of artificial intelligence (AI) has intensified this further: AI-readiness in policy terms reflects not merely computational availability, but the extent to which data are structured, documented and governed within institutional systems (OECD, 2021; International Science Council, 2025). Alongside AI-readiness, concerns about interoperability, long-term stewardship, cybersecurity, and fiscal sustainability have elevated data infrastructure from a technical support function to a core science-system capability (OECD, 2015; UNESCO, 2021; Wilkinson, Dumontier, et al., 2016). Decisions about data storage and sharing increasingly involve national science authorities, research funders and institutional leadership. Infrastructure choices now shape system-wide coordination, long-term sustainability, and the distribution of authority across science systems (Perrier et al., 2020; Borgman, 2015).

Structural variation and the Global South context

Although core storage and sharing technologies are globally available, their institutional configuration varies substantially – shaped by policy authority, resource constraints and the regulatory contexts within which science systems operate. Science systems in the Global South often operate under fiscal constraints, institutional asymmetries, and evolving regulatory frameworks that produce diverse configurations: nationally anchored data centres, federated arrangements, domain-specific platforms, and hybrid models combining institutional infrastructure with commercial services (World Bank, 2021).

These variations confirm that data storage and sharing are institutional configurations shaped as much by governance and fiscal conditions as by technical design. The relevant question for science-system planning is not which technology to adopt but what institutional conditions enable and sustain structured data infrastructure over time.

Scope and framing

This technology profile examines how data storage and sharing technologies function within science systems, with particular attention to institutional design, coordination mechanisms and long-term sustainability. It does not evaluate commercial cloud markets, hardware innovation trajectories or platform competition dynamics, except commenting on how they shape institutional design choices.

Rather than advancing a normative framework, this technology profile analyses observed institutional pathways and derives cross-cutting enabling conditions from eight cases spanning Africa, Asia and Latin America. Five cases are grounded in primary data from direct interviews and structured questionnaire exchanges with institutional representatives. These include: South Africa's Square Kilometre Array (SKA) programme and its precursor, the MeerKAT radio telescope; the Human Heredity and Health in Africa (H3Africa) Initiative platform; the Data Intensive Research Initiative of South Africa (DIRISA), China's National Genomics Data Center and National Space Science Data Center (NSSDC). Three cases are draw on desk research using publicly available documentation, policy sources and published literature. These are: Global Biodiversity Information Facility (GBIF) Global South Nodes; the Latin American Network for Open Science (LA Referencia); and CGIAR's Global Agricultural Data Innovation and Acceleration Network (GARDIAN).

Storage and sharing architectures: A working typology

Research data infrastructures are commonly described using terms – 'federated', 'distributed', 'centralized' – that carry different meanings across technical, policy and institutional contexts. Without a shared reference, these terms obscure rather than clarify strategic choices. Figure 1 establishes the working definitions used throughout this profile.

- **Centralized architecture** places both physical storage and governance authority within a single institution or node. It offers simplicity and clear accountability but creates single points of failure and concentrates political risk.

- **Distributed architecture** spreads data physically across multiple locations – primarily for performance, redundancy or throughput. Governance authority may remain centralized; distribution is a technical design choice, not a governance model.

- **Decentralized architecture** distributes both physical storage and governance authority across autonomous nodes with no single controlling entity. Peer-to-peer systems are the clearest example.

- **Federated architecture** is the most context-dependent term. In commercial and IT contexts, 'federated' typically refers to access control: role-based permissions and authentication across institutional boundaries. In science systems, it more often describes institutional federation – autonomous institutions maintaining their own data under their own governance while interoperating through shared standards and protocols. Recognizing that these are distinct configurations operating at different layers (physical storage, access control, institutional governance) provides an opportunity to enhance clarity and precision in infrastructure planning. Figure 1 maps these distinctions explicitly.

- **Compute-to-data (managed access)** inverts the standard model: analysis travels to the data rather than data being transferred to researchers. Used where datasets are too large to move, too sensitive to share openly, or subject to jurisdictional constraints, it requires coordination across all three layers – physical infrastructure, access control and institutional agreements about who may run what analysis under what conditions.

Most Global South research institutions operate outside these structured configurations. The modal practice is local institutional storage supplemented by commercial cloud, where budgets allow (Mell and Grance 2011), with data shared informally via rsync, email or physical media. The pathways examined in this profile represent the institutional conditions under

which some science systems move beyond this baseline toward more structured infrastructure – and what that transition requires.

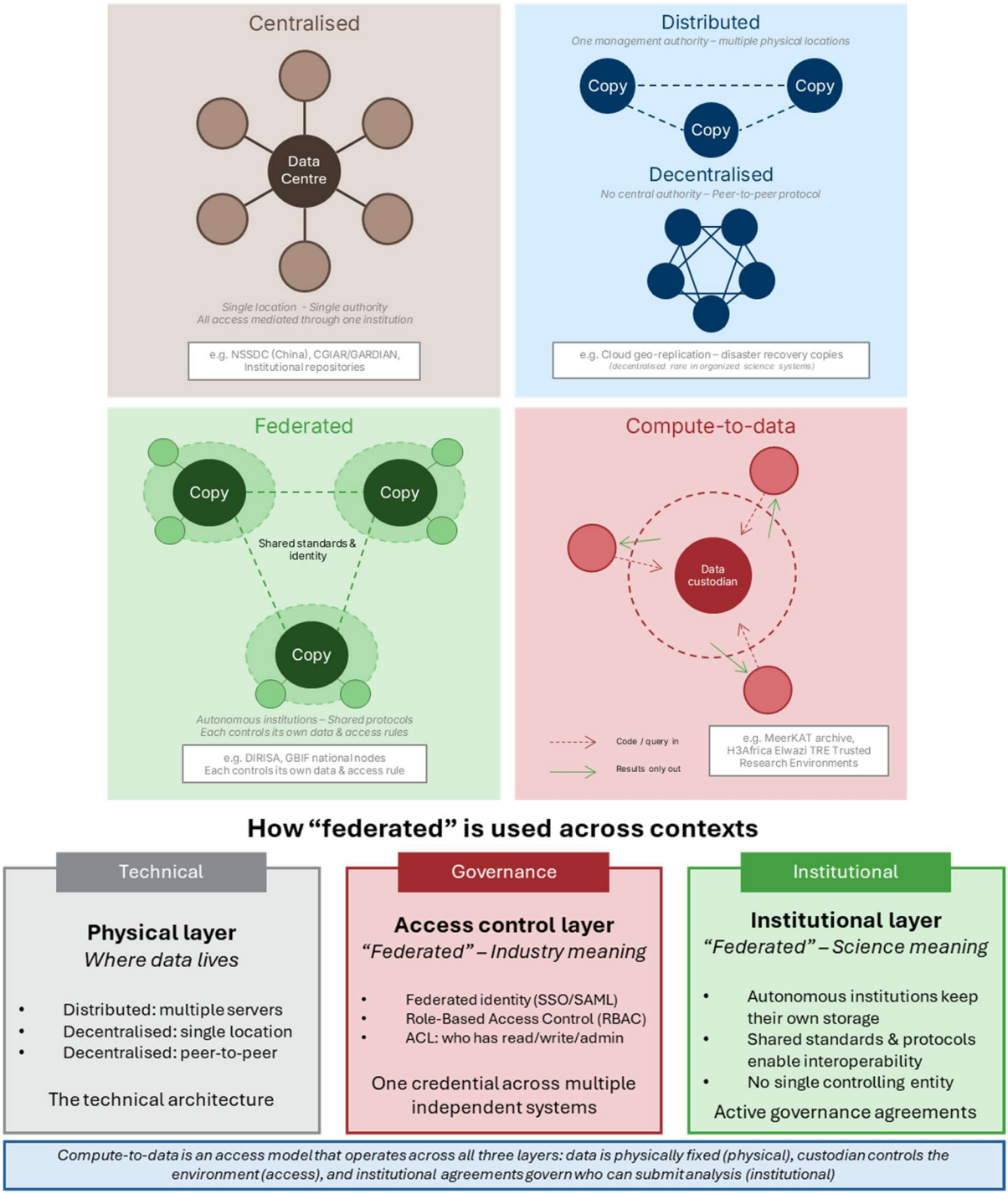

Figure 1: Storage and sharing architecture types and the three layers across which 'federated' is used in science systems. Top panels show four architecture types by physical configuration and governance structure. Bottom strip maps how 'federated' is applied across the physical, access-control and institutional layers – contexts that are frequently conflated in infrastructure planning and policy discussions.

Institutional pathways in data storage and sharing

Four recurring institutional pathways are observed across Global South science systems. Each pathway is defined by the primary constraint it addresses and the institutional configuration that emerges in response – not by geography, scientific domain or developmental stage. The pathways are not mutually exclusive: many systems exhibit elements of more than one, and some shift across pathways as their mandate and capacity evolve. Where a system is assigned to a single pathway, this reflects the dominant institutional configuration observed at the time of study. Figure 2 maps the four pathways and their defining characteristics.

Pathway	Primary constraint	Institutional configuration	Dominant mechanism	Cases
Pathway I *Performance scale integration*	Data volumes and velocities exceed individual institutional capacity for storage, transfer, or processing	Storage, compute, and networking co-integrated around the data source; compute-proximate architecture	Tiered storage (hot disk / cold tape archive); compute-to-data access; high-bandwidth networking	MeerKAT – South Africa's SKA programme
Pathway II *Domain-level consolidation*	Research data fragmented across incompatible systems within a scientific domain; limited cross-institutional reuse	Domain-specific platform concentration curation authority; standardized metadata and structured access control	Centralised repository; mandated deposit policies; tiered access control; standard metadata schemas	Guangong Genomics Data Centre (formerly CNGBdb), China NSSDC
Pathway III *Governance-first custodianship*	Data sensitivity, sovereignty, or strategic significance preludes unrestricted sharing; governance precedes architecture	Legal and access framework established before storage design; neutral custodian enforces depositor-set conditions	Controlled-access repositories; Trusted Research Environments; contractual enforcement; compute-to-data for sensitive data	H3Africa / Elwazi (Africa) DIRISA (South Africa)
Pathway IV *Coordination across distributed systems*	Research outputs dispersed across heterogeneous institutions with no shared discovery or access infrastructure	Coordination intermediary established lightweight shared standards; institutions retain governance of own holdings	OAI-PMH / API metadata harvesting; free publishing tools; policy mandate anchoring; aggregation without centralization	GBIF Global South Nodes LA Referencia CGIAR / GARDIAN

Figure 2: Four institutional pathways for data storage and sharing in Global South science systems, defined by primary constraint, institutional configuration and dominant mechanism. Pathways are not mutually exclusive; systems often combine elements of more than one.

Identifying your institutional pathway:

- If data volumes exceed what standard institutional storage can ingest or process → **Pathway I: Performance-scale integration.**
- If data is fragmented across institutions within a scientific domain → **Pathway II: Domain-level consolidation.**
- If data carries sensitivity, consent or sovereignty obligations that restrict open sharing → **Pathway III: Governance-first custodianship.**
- If research outputs are scattered across many institutions with widely varying technical capacity → **Pathway IV: Coordination across distributed systems.**

These four constraints reflect configurations observed across the cases examined; other primary constraints may apply in science systems not represented in this study.

Pathway I: Performance-scale integration

When research instruments generate data volumes and velocities that individual institutions cannot store, process or transfer through standard means, science systems respond by co-integrating storage, compute and networking around the data. The organizing logic is compute-proximate storage: rather than moving data to researchers, analysis is brought to where the data resides.

MeerKAT and South Africa's SKA Programme

South Africa's 64-dish MeerKAT radio telescope generates approximately 10 petabytes of raw data per year – faster than it can be transferred by network and impractical to cloud-host. Storage is tiered and centralized at Cape Town's supercomputing facility: active working data on disk (approximately 10 PB), with older data migrated to a magnetic tape archive (approximately 40 PB) as disk capacity fills. Critically, the tape tier was not part of the original design – it emerged from operational necessity as data volumes consistently exceeded year-on-year projections.

Researchers access data through a web portal that triggers on-request tape-to-disk restoration; for large international transfers, physical tape shipping is a documented practice. The organizing principle is compute-proximate access: analysis is brought to the data, not the reverse.

The forthcoming SKA will require a distributed network of regional computing centres (SRCNet), managed by the SKA Observatory – this is a future design horizon, not an extension of MeerKAT's current architecture.

The primary lesson: *at performance scale, archival storage requirements consistently outpace initial planning – cold storage is not an optional addition, it is an operational inevitability.*

Pathway II: Domain-level consolidation

When research data within a scientific domain is scattered across incompatible institutional systems, the response is a domain-specific platform that aggregates data, enforces common metadata and access standards, and creates a unified resource that individual institutions cannot build independently. The primary constraint is fragmentation within a domain; the platform addresses it by concentrating curation authority while maintaining access pathways for the research community.

National Genomics Data Center, China

The National Genomics Data Center, which is a part of the China National Center for Bioinformation and operated by BGI, addresses domain fragmentation within global life sciences research – aggregating sequencing data, genome assemblies and phenotypic records from dispersed institutional producers into a unified platform. As of January 2026, the platform holds 21 PB of data covering 1.39 million samples, 581,805 genome assemblies, and 8,894 species.

A significant infrastructure shift has accompanied this growth: the centre has transitioned from on-premises high-performance computing to a cloud-based architecture, driven by the need for elastic scaling and predictable cost structures as data volumes grow faster than capital procurement cycles. Access operates on a two-tier model: public access for open data; controlled access requiring a formal application and Data Access Committee review for restricted datasets.

International credibility is institutionally anchored: CoreTrustSeal certification (2023) and World Data System membership (2024) signal long-term stewardship commitment to global research communities.

The lesson: *national domain platforms that align with international standards attract more deposits and sustain greater reuse than isolated national repositories.*

National Space Science Data Center, China
China's National Space Science Data Center (NSSDC), established under the Chinese Academy of Sciences, is the primary custodian for data from China's space science missions – covering space physics, space astronomy and lunar and planetary sciences (Earth observation data is managed by a separate dedicated national centre). As of 2025, NSSDC manages over 5 PB of data across more than 4,000 datasets, serving 70,000+ registered users. Holdings are projected to reach 20 PB by 2030.

Storage is centralized and hierarchical at NSSDC's Beijing facility, with three geographically distributed off-site disaster recovery centres (Beijing, Guangdong and Yunnan). Data enters through two channels: direct management of missions where NSSDC is the designated data centre, and mandatory deposit under China's Scientific Data Management Measures, which require submission of all publicly funded space science data. The governing principle is 'as open as possible, as closed as necessary'.

Discovery is supported by the Virtual Space Science Observatory, enabling cross-mission search including natural language queries.

The lesson: *legislated deposit mandates – not researcher goodwill – are the most reliable mechanism for populating national domain platforms at scale.*

Pathway III: Governance-first custodianship

When data involves sensitive personal, national or strategically significant information, the governance framework – access conditions, legal instruments, data sovereignty obligations – must be established before storage architecture can be designed. Infrastructure in this pathway is shaped by what data *cannot* be shared openly, as well as by what it contains.

Human Heredity and Health in Africa platform
The Human Heredity and Health in Africa (H3Africa) consortium generated genomic datasets from more than 50,000 participants across African research institutions – data too sensitive for unrestricted open access and too large for standard institutional repositories. The architecture was designed around governance requirements: a three-tier structure modelled on the European Genome-phenome Archive, with controlled access managed at the international level, curated data at a continental node, and raw data held at the institutional site of collection, with a central copy maintained by the H3Africa archive and data coordinating centre. Enforcement was contractual rather than

cultural: data access agreements – not researcher trust – governed who could access what under what conditions.

The successor platform, eLwazi, extends this through on-premises trusted research environments with a pilot test project operating across partner institutions in Mali, Uganda and South Africa using GA4GH standards. Each institution governs its own data while interoperating through shared access-review protocols. The trusted research environment model – computation travels to the data – means sensitive genomic data never leaves its institutional jurisdiction.

There are three internationally recognized open science frameworks:

- FAIR: findable, accessible, interoperable, reusable;
- CARE: collective benefit, authority to control responsibility, ethics; and
- TRUST: transparency, responsibility, user focus, sustainability, technology.

H3Africa formally engaged with the FAIR framework. FAIR principles were applied systematically and assessed throughout the programme: tools, portals, and data holdings were regularly badged against FAIR criteria, and FAIR compliance became a formal funding requirement in the second round of H3Africa grants. When directly asked about CARE (Carroll, Garba, et al., 2020) and TRUST, the consortium's data lead noted that the CARE terminology had not yet emerged during the design phase.

The lesson: *for sensitive human research data, governance architecture is the primary design challenge. Storage choices follow from legal and ethical frameworks; they cannot precede them.*

South Africa's National Research Data Infrastructure

Data Intensive Research Initiative of South Africa (DIRISA), hosted within the South African National Research Foundation, operates under a federated custodianship model. Mapped against the typology in Figure 1: physical storage is centralized at two national sites (Western Cape and Gauteng), but institutional governance is federated – each depositing institution retains custodianship authority and sets the access conditions for its own holdings, which DIRISA enforces as neutral infrastructure. DIRISA is not open-by-default; access is negotiated on a per-deposit basis, reflecting data sensitivity and funder requirements.

What distinguishes DIRISA institutionally is its proactive policy role. Rather than responding to researcher demand, it actively shapes national data management guidelines and engages with South Africa's evolving research data governance framework. Embedding data infrastructure within the National Research Foundation's mandate – South Africa's primary public research funding agency – provides legitimacy and funding continuity that purely technical infrastructure projects rarely sustain.

The lesson: *custodianship infrastructure requires governance authority, not just storage capacity. Institutions that shape national data policy, rather than merely implement it, demonstrate greater long-term durability.*

Pathway IV: Coordination across distributed systems

When research outputs are scattered across many heterogeneous institutions with widely varying technical capacity, the response is a coordination intermediary that establishes lightweight shared standards and publishing tools – enabling dispersed institutions to contribute to a shared discovery layer without surrendering governance over their own holdings. The intermediary does not own the data; it makes it findable.

Global Biodiversity Information Facility Global South Nodes

Global Biodiversity Information Facility (GBIF) coordinates biodiversity data publishing across thousands of heterogeneous institutions – natural history museums, government environmental agencies, university collections and citizen science programmes – through national nodes including SANBI-GBIF (South Africa), SiBBr (Brazil), and SiB Colombia. The dispersion constraint is severe: occurrence data is scattered across organizations ranging from major research universities to small non-governmental organizations (NGOs) with no server infrastructure.

The coordination mechanism is deliberately lightweight: the Integrated Publishing Toolkit, offered as free hosted accounts for low-resource institutions, converts local data to Darwin Core Archive format for standard harvesting. The Biodiversity Information for Development programme anchors data mobilization to national biodiversity strategy obligations, sustaining participation beyond individual project cycles. The programme has had €9.5 million investment from the European Union since 2015. GBIF now holds over 1.6 billion occurrence records, representing 1,150 percent growth over a decade (Heberling et al., 2021).

A persistent equity tension: biodiversity data from Global South countries flows primarily to Global North researchers and institutions, raising data colonialism concerns that parallel Nagoya Protocol debates on genetic resources.

The lesson: *coordination intermediaries succeed when they lower institutional barriers through free tools and policy anchors – but must engage explicitly with asymmetric data flows to maintain legitimacy in Global South contexts.*

The Latin American Network for Open Science

The Latin American Network for Open Science (LA Referencia: Red Federada de Repositorios Institucionales de Publicaciones Científicas) is a regional federated network of open access repositories founded in 2012, spanning ten countries. The primary constraint it addresses is a structural visibility problem: research published in Spanish and Portuguese is systematically absent from major commercial bibliometric databases, rendering publicly funded science invisible to both the global community and the societies that funded it.

The coordination mechanism is a three-tier OAI-PMH federation: institutional repositories at universities and research centres feed national aggregators, which feed a regional discovery layer. Per-node hardware requirements are deliberately minimal – two-core processor, 8 GB RAM, 500 GB storage[1] – making the model accessible across the region's varied IT capacity. As of 2025, the network provides open access to over 1.43 million documents, with 5.5 million records accessible through integration with OpenAIRE. A 2025 Memorandum of Understanding with OpenAIRE and RedCLARA extends South–North interoperability.

LA Referencia operates within Latin America's diamond open access model: no author-facing article processing charges, publicly funded infrastructure, content freely accessible to readers. National open access mandates in Argentina and Brazil consistently produce higher participation and metadata quality.

The lesson: *regional coordination across heterogeneous institutional infrastructure is financially sustainable when anchored to national policy and built on lightweight open* standards – a replicable model for other Global South regions facing the same language-visibility gap.

[1] https://www.lareferencia.info/en/services/tecnologia

CGIAR and the Global Agricultural Research Data Innovation and Acceleration Network

CGIAR is one of the largest international agricultural research partnerships, with a mandate focused on food and nutrition security in the Global South. The Global Agricultural Research Data Innovation and Acceleration Network (GARDIAN), operated within CGIAR's data and digital research initiatives, aggregates publications and datasets from multiple distributed centre repositories via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The architecture separates discovery from storage: GARDIAN functions as a metadata aggregation and discovery layer, while full-text content remains in originating repositories such as CGSpace, the system's shared DSpace-based document repository.

CGIAR's Open and FAIR Data Assets Policy (2021) requires research outputs to be managed according to FAIR principles, with CC BY 4.0 as the default license for most outputs. A distinct governance layer covers over 700,000 crop genetic resource accessions held across CGIAR genebanks; these are governed under the International Treaty on Plant Genetic Resources for Food and Agriculture – not the Nagoya Protocol – placing them within a multilateral benefit-sharing framework rather than bilateral national agreements.

The structural tension CGIAR makes visible: data about smallholder farmers, crop varieties and ecological systems across Africa, Asia and Latin America is collected and published by an international organization, with limited capacity for national agricultural research services to co-govern or build on their own research base. The One CGIAR reform (from 2022) has acknowledged this gap; closing it requires investment in national data infrastructure that international organizations can support but not substitute for.

Enabling conditions

Five structural conditions recur across the cases examined, consistently more decisive than technological choice or funding volume alone. They are not best practices or prescriptions but conditions that shape feasibility, legitimacy and longevity across all four pathways. Systems that exhibit all five tend to sustain and expand their infrastructure over time; weakness in any one constrains even technically sophisticated implementations.

1. **A mandate or policy anchor.** The most durable data infrastructures are those where participation is sustained by policy obligation, not researcher goodwill. China's Scientific Data Management Measures make submission of publicly funded research data legally mandatory – China's NSSDC deposit volumes reflect this mandate, not discretionary compliance. The Biodiversity Information for Development programme tied GBIF data mobilization to national biodiversity strategy obligations. LA Referencia operates within national open access legislation across ten countries. Where mandates are absent or unenforced, deposit remains discretionary, quality is variable and institutional investment in infrastructure is difficult to justify.

2. **Governance architecture before technical design.** The consistent sequence in durable infrastructures is: establish who can access what, under what conditions, before procuring storage. H3Africa's three-tier model was a governance design before it was a technical one – the European Genome-phenome Archive architecture was chosen because it matched access-control requirements, not the other way around. The National Genomics Data Center underwent multiple rounds of architecture adjustment as human genetic resources regulation evolved; in each case, regulatory requirements drove technical change. Institutions that design governance after the fact consistently face harder and costly adaptation problems.

   Governance in practice materializes as concrete policy instruments: data access conditions and sharing agreements, licensing terms (Creative Commons waivers for open datasets; controlled access agreements for sensitive data) and institutional sustainability plans. These are the documents through which governance commitments become visible to prospective depositors and users evaluating whether to engage with a repository.

   Three internationally recognized frameworks provide reference points for governance design:
   - **FAIR:** findable, accessible, interoperable, reusable (Wilkinson et al., 2016). FAIR addresses data discoverability and reuse readiness, and has achieved the broadest formal adoption across the cases. CGIAR mandates FAIR compliance with CC BY 4.0 as

the default license; NSSDC assigns persistent identifiers and Creative Commons licenses to maximize FAIRness.

- **CARE:** collective benefit, authority to control, responsibility, ethics (Carrol et al., 2020). CARE addresses the rights and interests of communities whose data are collected, particularly where data originates from historically marginalized or Indigenous populations, as in H3Africa.

- **TRUST:** transparency, responsibility, user focus, sustainability, technology (Lin et al., 2020). TRUST addresses the institutional properties that make repositories credible long-term stewards. DIRISA interviewees explicitly invoked TRUST principles by name as a framework they apply and actively advocate for. The National Genomics Data Center received CoreTrustSeal certification in 2023 – an international certification scheme aligned with the TRUST principles – though the institution refers to the certification rather than the underlying framework by name. DIRISA interviewees further noted that CARE and TRUST, while increasingly advocated, are not yet uniformly adopted across depositing institutions – a gap DIRISA addresses through its annual national research data workshops and stakeholder engagement programme.

3. **Institutional embedding in national governance.** Data infrastructure actors that hold formal authority within national science governance – not just technical expertise – demonstrates greater long-term durability. DIRISA's embedding within the South African National Research Foundation gives it legitimacy in national policy discussions and a funding baseline that is part of the National Research Foundation's mandate rather than a grant within it. NSSDC operates under the Chinese Academy of Sciences with a Ministry of Science and Technology funding line; GBIF nodes anchored to national biodiversity agencies sustain participation across funding cycles better than nodes hosted within universities or NGOs alone. Technical capacity is necessary but not sufficient; institutional authority is the stabilizing factor.

4. **Interoperability through lightweight open standards.** The most scalable coordination models adopt internationally recognized open standards as the mechanism for connecting institutions, rather than building bespoke integration solutions. OAI-PMH enables LA Referencia, CGIAR/GARDIAN, and GBIF to harvest metadata from distributed repositories without centralizing data or requiring custom interfaces at each node; Darwin Core Archive format allows institutions with minimal IT capacity to publish data through GBIF's free hosted Integrated Publishing Toolkit. CoreTrustSeal certification (The National Genomics Data Center) provides international credibility without bilateral validation agreements. Standards adoption reduces the cost of participation and enables interoperability at scale; bespoke integration achieves neither.

5. **Funding continuity over funding amount.** The volume of funding is less decisive than its source and structural continuity. NSSDC cited stable Ministry subsidy funding alongside project-based income as the combination that sustains operations through periods of rapid data volume growth; DIRISA's National Research Foundation mandate provides a funding baseline that does not depend on grant renewal cycles. CGIAR's Platform for Big Data – funded as a cross-cutting platform rather than a project – is more insulated from the fragmentation that characterizes centre-level funding. Short funding cycles create a specific failure mode: infrastructure is procured but not staffed, software is deployed but not updated, and institutional knowledge dissipates when contracts end, regardless of the initial investment scale. The cases also illustrate a range of fiscal models that extends well beyond user-fee revenue: stable ministry or agency subsidies (NSSDC, DIRISA), project-based co-funding tied to mission participation (NSSDC, National Genomics Data Center), regional donor programmes anchored to national policy obligations (GBIF/Biodiversity Information for Development), and publicly funded diamond open access models with no charges to authors or users (LA Referencia). The National Genomics Data Center explicitly noted the importance of diversified revenue and moving toward basic cost balance rather than dependence on time-limited project funding. User-fee models, while viable for large platforms with established user bases, introduce equity barriers and revenue uncertainty that are particularly acute in resource-constrained settings; the cases that demonstrate the greatest funding continuity are those anchored to institutional mandates rather than transactional revenue. For a systematic treatment of fiscal model options and their trade-offs for data reuse infrastructure, see Hooft and Roos (2025).

Future directions: Considerations for science-system planning

AI-driven data management. Large language models and automated classification tools are beginning to reduce the marginal cost of metadata generation, duplicate detection and quality assessment. For Global South institutions managing large volumes of heterogeneous legacy data with constrained curation capacity, these tools offer real leverage in lowering the operational barriers to making legacy datasets discoverable and reusable. However, automated curation produces outputs that are difficult to audit, can embed disciplinary biases, and do not resolve the underlying governance questions: who authorizes metadata schemas, what constitutes sufficient documentation and how quality is defined across scientific communities. Institutions that adopt AI-assisted curation without governance frameworks for validating and auditing outputs risk trading one curation bottleneck for a less visible one.

Federated learning and compute-to-data. Compute-to-data models – where analysis runs at the data source rather than data being moved to analysis – are being more widely adopted as an alternative to centralized repositories in sensitive domains. For health and genomic data, this approach can satisfy sovereignty and sensitivity constraints that would otherwise block sharing entirely. The governance challenge is that compute-to-data displaces rather than resolves questions about access, auditability, and benefit-sharing: who controls the analysis environment, what outputs can be extracted, and how results are attributed or contested. The Elwazi Trusted Research Environment model is an early institutional instantiation of this logic; whether it can be extended across jurisdictions and resource settings remains an open question for science-system planners.

Data sovereignty regimes beyond biodiversity. The Nagoya Protocol framework for genetic resources and traditional knowledge established a precedent for benefit-sharing obligations that travel with data. Analogous pressures are building in genomics (H3Africa and related pan-African initiatives), agricultural data (through CGIAR reform and seed-system governance debates), and increasingly in climate and earth observation data. Institutional configurations built today – whether centralized national repositories, federated custodianship models or third-party platforms – will face these obligations under evolving regulatory regimes. Science systems that treat sovereignty as a downstream compliance requirement rather than an upstream design parameter will face costly retrofits as these regimes mature.

Cloud-institutional hybrid sustainability. Commercial cloud services have reduced entry costs for data infrastructure across Global South science systems, and several cases in this profile involve hybrid or cloud-native architectures. The long-term sustainability questions are not primarily technical: they concern cost trajectories as data volumes grow, vendor lock-in at the data and application programming interface layer, and institutional capacity to negotiate, audit and exit commercial arrangements. Procurement decisions made under short-term cost logic can foreclose options that matter for sovereignty, reproducibility and long-term continuity – with consequences that become visible only on decade-long time horizons.

These directions converge on a common challenge: The technological advances create new options for data infrastructure, but the governance questions they raise are not resolved by the technology itself. Science systems that invest in governance capacity alongside technical capability will be better positioned to adapt as these directions mature, regardless of which specific technologies prove durable. Pathway choices made now will either preserve or foreclose that adaptability.

Conclusion

This profile argues that data storage and sharing in Global South science systems are institutional configurations shaped as much by governance and fiscal capacity as by technology. The cases examined identify no single infrastructure model as universally appropriate; what distinguishes more sustainable configurations is not their technical architecture but the coherence between institutional mandate, governance design and pathway choice. Two structural findings hold across all four pathways. First, the most common failure mode is sequence error: procuring storage before resolving access conditions, funding infrastructure without mandating deposit, or replicating a peer institution's model without diagnosing the primary constraint.

Second, biodiversity, agricultural, and genomic data from Global South countries consistently flow through infrastructure governed by international consortia or Global North institutions, with limited capacity for national systems to co-govern or build on their own research base. International support that transfers tools without transferring governance capacity perpetuates rather than resolves this asymmetry. The recommendations below address four stakeholders and are designed to be actionable within different institutional positions, not as a coordinated implementation agenda.

Recommendations

Science-system leaders and research institutions

- Diagnose the primary constraint before selecting an infrastructure model – performance-scale, domain fragmentation, data sensitivity and institutional dispersion each call for a different pathway and different investments.
- Resolve governance questions (access conditions, sensitivity classifications, sovereignty obligations) before procuring storage; retrofitting governance to existing infrastructure is consistently more costly.
- Anchor data management requirements to institutional mandates and funding conditions, not to project-specific data management plans that expire with the grant.

National research funders

- Mandate data deposit rather than incentivize it. Voluntary frameworks systematically underpopulate national platforms and cannot justify sustained infrastructure investment.
- Fund operational continuity alongside capital procurement – equipment without staffing, maintenance and curation capacity is not infrastructure.

- Recognize coordination intermediaries (metadata aggregators, standards nodes, lightweight publishing tools) as infrastructure investment, not as administrative overhead; they are frequently the highest-leverage intervention in dispersed institutional environments.

International partners and donors

- Align funding timelines with infrastructure lifecycle requirements. Project-cycle funding for multi-year operational infrastructure creates the specific failure mode of procurement without maintenance.
- Build national governance capacity alongside technical infrastructure – tool transfer without governance transfer perpetuates external dependency.
- Engage explicitly with data sovereignty and benefit-sharing obligations when supporting biodiversity, genomic or agricultural data systems. Open data defaults are not neutral in these domains.

Regional coordination bodies

- Invest in shared standards adoption before attempting platform harmonization: OAI-PMH-based metadata federation consistently outperforms centralized platform projects in heterogeneous institutional environments at lower cost.
- Develop South–South learning and adaptation mechanisms so that proven models (GBIF national nodes, Elwazi trusted research environments, LA Referencia) can be adapted rather than reinvented across contexts.
- Advocate for national data mandates as a prerequisite for effective regional coordination – regional infrastructure cannot substitute for absent national policy foundations.

Glossary of terms

- **AI-driven data management:** The use of large language models and automated classification tools to support metadata generation, duplicate detection, and quality assessment of research data. Identified in this profile as a forward-looking consideration that can lower the operational cost of curating large legacy datasets but displaces rather than resolves governance questions about validation, auditability, and bias.

- **AI-readiness:** The extent to which data are structured, documented, and governed within institutional systems in a way that supports the responsible application of artificial intelligence methods in research. Distinguished in this profile from raw compute availability - AI-readiness is a property of data and the institutions that hold it.

- **Benefit-sharing:** Obligations - established by frameworks such as the Nagoya Protocol and the International Treaty on Plant Genetic Resources for Food and Agriculture - requiring that value derived from data, genetic resources, or associated knowledge be equitably shared with the communities, countries, or institutions of origin.

- **BID programme:** The Biodiversity Information for Development programme - an EU-funded GBIF initiative supporting biodiversity data mobilisation in Global South countries, anchored to national biodiversity strategy obligations.

- **CARE Principles:** A framework for Indigenous and community data governance - Collective Benefit, Authority to Control, Responsibility, Ethics - developed to address the rights and interests of communities whose data are collected in research. In this profile, CARE complements FAIR by foregrounding rights, authority, and benefit-sharing.

- **Centralised architecture:** A data infrastructure model in which both physical storage and governance authority reside within a single institution or node. Offers simplicity and clear accountability but creates single points of failure and concentrates political risk.

- **CGIAR:** An international agricultural research partnership with a mandate focused on food and nutrition security in the Global South. In this profile, CGIAR's data infrastructure is examined through its GARDIAN platform (see entry) and its governance of over 700,000 crop genetic resource accessions under the ITPGRFA.

- **Cloud-based architecture**: An infrastructure model in which storage and compute are provided through commercial cloud services, supporting elastic scaling while introducing considerations related to vendor dependency, recurring costs, and data sovereignty. The

Guangdong Genomics Data Center's transition from on-premises high-performance computing to cloud illustrates this configuration in the profile.

- **Compute-proximate storage:** An architectural principle in which analysis is brought to where the data resides, rather than data being moved to the researcher. Used where data volumes exceed practical transfer capacity; the organising logic of Pathway I (Performance-Scale Integration).

- **Compute-to-data (managed access):** An access model in which analysis code or queries are executed inside the data custodian's controlled environment, and only approved outputs are returned to the researcher. Used for datasets too sensitive, too large, or subject to jurisdictional constraints that prevent open sharing.

- **Coordination Across Distributed Systems (Pathway IV):** An institutional pathway in which a coordination intermediary establishes lightweight shared standards and publishing tools that allow dispersed institutions - with widely varying technical capacity - to contribute to a shared discovery layer without surrendering governance over their own holdings. Exemplified by GBIF, LA Referencia, and CGIAR/GARDIAN.

- **CoreTrustSeal:** An international certification scheme aligned with the TRUST Principles that signals a repository's long-term stewardship commitment to global research communities. Achieved by the Guangdong Genomics Data Center in 2023.

- **Darwin Core Archive:** A standardised biodiversity data format used by GBIF that allows institutions with minimal IT capacity to publish occurrence data through lightweight tools such as the Integrated Publishing Toolkit (see entry).

- **Data access agreement:** A formal instrument governing who may access controlled-access datasets, under what conditions, and for what purposes. Enforcement varies across cases: H3Africa, eLwazi, and the Guangdong Genomics Data Center administer agreements through Data Access Committees (see entry), while DIRISA uses per-deposit negotiation. In all cases, agreements substitute for reliance on researcher goodwill alone.

- **Data Access Committee (DAC):** An institutional body that reviews applications for access to controlled-access datasets and authorises release under the terms of a data access agreement (see entry). Used by H3Africa and the Guangdong Genomics Data Center; DIRISA operationalises controlled access through per-deposit negotiation rather than a standing committee.

- **Data colonialism:** A structural pattern in which data from Global South countries, communities, or ecosystems are primarily processed, governed, or used by external institutions, often limiting equitable participation, local control, or benefit-sharing. Cited in this profile in relation to biodiversity, genomic, and agricultural data flows.

- **Data sovereignty:** The principle that data generated by or about a community, nation, or population should be governed according to the laws, norms, and interests of that originating context, rather than by external institutions or commercial platforms.

- **Decentralised architecture:** A data infrastructure model in which both physical storage and governance authority are distributed across autonomous nodes with no single controlling entity. Peer-to-peer systems are the clearest example.

- **Diamond open access:** A publishing model in which research outputs are freely accessible to readers and authors pay no article processing charges; infrastructure is publicly funded. Characterises LA Referencia's model, which operates within Latin America's broader diamond open-access ecosystem.

- **DIRISA:** The Data Intensive Research Infrastructure of South Africa, hosted within the South African National Research Foundation (NRF). DIRISA operates a federated custodianship model - physical storage centralised at two national sites, governance authority retained by depositing institutions - and participates in broader national coordination and policy discussions related to research data infrastructure.

- **Distributed architecture:** A data infrastructure model in which data are stored physically across multiple locations for performance, redundancy, or throughput, while governance authority may remain centralised. Distribution is a technical design choice, not a governance model.

- **Domain-Level Consolidation (Pathway II):** An institutional pathway in which a domain-specific platform aggregates fragmented data from incompatible institutional systems within a scientific domain, enforcing common metadata and access standards to create a unified resource. Exemplified by the Guangdong Genomics Data Center and China's NSSDC.

- **eLwazi:** The successor platform to H3Africa, extending genomic data access governance through on-premises Trusted Research Environments (see entry) operating across partner institutions in Mali, Uganda, and South Africa under GA4GH standards.

- **FAIR Principles:** Guiding principles for scientific data management - Findable, Accessible, Interoperable, Reusable - that address data discoverability and reuse readiness. The framework with the broadest formal adoption across the cases examined in this profile.

- **Federated architecture:** In science-system contexts, a model in which autonomous institutions maintain their own data under their own governance while interoperating through shared standards and protocols. Distinct from the term's use in commercial and IT contexts, where 'federated' typically refers to access control and identity management across institutional boundaries.

- **Federated custodianship:** An institutional model - exemplified by DIRISA - in which physical storage is centralised but each depositing institution retains custodianship authority and sets access conditions for its own holdings, enforced by a neutral infrastructure operator.

- **Federated learning:** A machine-learning approach in which model training occurs locally at distributed data sources and only model updates, not raw data, are aggregated centrally. Referenced in this profile as a sovereignty-compatible analytical approach for sensitive domains.

- **Funding continuity:** The structural condition in which infrastructure funding is sustained across project cycles - through institutional mandates, ministry subsidies, or long-horizon grants - rather than depending on short-cycle project grants that create specific failure modes when they end. Identified in this profile as more decisive than funding volume.

- **GA4GH standards:** Technical standards developed by the Global Alliance for Genomics and Health to support responsible, interoperable genomic and health data sharing. Used in the eLwazi Trusted Research Environment to enable cross-institutional access governance.

- **GARDIAN:** The Global Agricultural Research Data Innovation and Acceleration Network - a CGIAR platform that aggregates publications and datasets from distributed centre repositories via OAI-PMH. Functions as a metadata discovery layer while full content remains in the originating CGIAR repositories.

- **GBIF:** The Global Biodiversity Information Facility - an international coordination network that aggregates biodiversity occurrence data from thousands of heterogeneous institutions through lightweight shared standards and national nodes, including nodes in Global South countries. Holds over a billion biodiversity occurrence records. The asymmetric flow of

these data to Global North researchers is identified in this profile as a persistent equity tension (see Data colonialism).

- **Global South:** A geopolitical shorthand used throughout this profile to refer to countries in Africa, Asia, Latin America, and the Caribbean, whose science systems operate under diverse institutional, fiscal, and regulatory conditions that differ structurally from those assumed in most technology policy frameworks.

- **Governance-First Custodianship (Pathway III**): An institutional pathway in which legal and ethical frameworks for data access - consent, sensitivity, sovereignty - are established before storage architecture is designed; infrastructure choices follow from governance requirements, not the reverse. Exemplified by H3Africa, eLwazi, and DIRISA.

- **Guangdong Genomics Data Center (GDC**): A genomics data platform operated by BGI Research, also referred to as CNGBdb, examined in this profile as a Pathway II case. Operates at petabyte scale; transitioned from on-premises high-performance computing to cloud-based architecture for elastic scaling; CoreTrustSeal-certified in 2023.

- **H3Africa:** Human Heredity and Health in Africa - a genomic research consortium that generated datasets from over 50,000 participants across African institutions. Designed around a three-tier governance architecture - controlled access at the international level, continental curation, and institutional data custody - modelled on the European Genome-phenome Archive.

- **Institutional embedding:** The presence of data infrastructure actors within national science governance - not merely as technical operators, but as actors holding formal authority within ministry, funder, or academy structures. Identified in this profile as one of the enabling conditions; DIRISA (within South Africa's NRF) and NSSDC (under the Chinese Academy of Sciences) are the leading examples.

- **Integrated Publishing Toolkit (IPT):** A free, open-source tool developed by GBIF that converts local biodiversity data into Darwin Core Archive format for standard harvesting, enabling institutions with minimal IT capacity to contribute to global biodiversity data networks.

- **Interoperability:** The technical and institutional capacity of disparate data systems to exchange, interpret, and reuse one another's data, achieved through adoption of shared open standards rather than bespoke integration.

- **ITPGRFA:** The International Treaty on Plant Genetic Resources for Food and Agriculture - the multilateral legal framework governing over 700,000 crop genetic resource accessions held across CGIAR genebanks. Distinct from the bilateral Nagoya Protocol framework.

- **LA Referencia:** Red Federada de Repositorios Institucionales de Publicaciones Cientificas - a regional federated network of open-access repositories founded in 2012, spanning ten Latin American countries. Addresses the systematic invisibility of Spanish- and Portuguese-language research in global bibliometric databases through OAI-PMH-based metadata federation, operating within a diamond open-access model.

- **MeerKAT:** A 64-dish radio telescope in South Africa managed through a tiered storage architecture at Cape Town's supercomputing facility. A precursor to the Square Kilometre Array (see entry).

- **Nagoya Protocol**: An international access and benefit-sharing framework under the Convention on Biological Diversity, establishing obligations that travel with genetic resources and associated traditional knowledge. Creates data sovereignty implications for biodiversity and genomic research.

- **NSSDC:** The National Space Science Data Center of China - the primary custodian for data from China's space science missions, operated under the Chinese Academy of Sciences. Manages petabyte-scale holdings populated under mandatory deposit through China's Scientific Data Management Measures.

- **OAI-PMH:** The Open Archives Initiative Protocol for Metadata Harvesting - a lightweight, open standard enabling metadata aggregation from distributed repositories without centralising data or requiring custom interfaces. Used by LA Referencia, CGIAR/GARDIAN, and GBIF.

- **Open-science frameworks:** Internationally recognised sets of principles guiding responsible data management and stewardship - FAIR (Findable, Accessible, Interoperable, Reusable), CARE (Collective Benefit, Authority to Control, Responsibility, Ethics), and TRUST (Transparency, Responsibility, User Focus, Sustainability, Technology). See individual entries.

- **Performance-Scale Integration (Pathway I):** An institutional pathway in which storage, compute, and networking are co-integrated around a data source whose volume or velocity exceeds the capacity of individual institutions. Organised around compute-proximate access; exemplified by MeerKAT and South Africa's SKA Programme.

- **Persistent identifier (PID):** A standard, globally unique, machine-resolvable reference (such as a DOI or Handle) that identifies a dataset, publication, or resource over time. Persistent identifiers underpin Findability and Reusability under the FAIR Principles; assigned by NSSDC and CGIAR/GARDIAN to maximise reuse and citability.

- **Policy anchor (mandate):** A formal institutional or legal obligation - such as a national data deposit requirement or open-access legislation - that sustains participation in data infrastructure beyond individual researcher goodwill or project cycles. Identified in this profile as the most decisive of the enabling conditions.

- **Scientific Data Management Measures (China):** The legislative instrument that makes deposit of publicly funded research data legally mandatory in China. Cited in this profile as the policy anchor behind NSSDC's deposit volumes (see also: NSSDC, Policy anchor).

- **Sequence error:** The most common infrastructure failure mode identified in this profile: procuring storage before resolving access conditions, funding infrastructure without mandating deposit, or replicating a peer institution's model without first diagnosing the primary constraint.

- **SKA (Square Kilometre Array):** A major international radio telescope project for which South Africa's MeerKAT is a precursor. The forthcoming SKA will require a distributed SKA Regional Centre Network (SRCNet) managed by the SKA Observatory - a distinct future architecture rather than an extension of MeerKAT.

- **Standards adoption:** An enabling condition in which institutions adopt internationally recognised open standards - such as OAI-PMH, Darwin Core, CoreTrustSeal, and FAIR - as coordination mechanisms, reducing the cost of interoperability at scale compared with bespoke integration.

- **Tiered storage:** A data storage architecture in which data are migrated between storage media based on access frequency and cost: active data on fast disk, older or less-accessed data on lower-cost magnetic tape archive ('cold storage'). Used by MeerKAT/SKA and NSSDC; treated in the profile as operationally necessary at petabyte scale.

- **TRUST Principles:** A framework - Transparency, Responsibility, User Focus, Sustainability, Technology - specifying the institutional properties that make digital repositories credible long-term stewards. Explicitly invoked and advocated by DIRISA.

- **Trusted Research Environment (TRE**): A secure, typically on-premises computational environment in which analysis of sensitive data is performed without that data leaving its institutional jurisdiction. Used in the eLwazi platform for genomic data.

- **Vendor lock-in:** Dependence on a specific commercial cloud provider's proprietary data formats, application programming interfaces, or contractual terms, foreclosing future options for migration, sovereignty, reproducibility, and long-term continuity. A central sustainability concern in cloud-institutional configurations.

- **Virtual Space Science Observatory (VSSO):** A discovery system operated by China's NSSDC enabling cross-mission search across space science datasets, including natural-language query support.

References

Borgman, C. L. 2015. Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, Massachusetts, MIT Press.

Carroll, S. R. et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal,* Vol. 19, No. 1, p. 43. https://doi.org/10.5334/dsj-2020-043.

Heberling, J. M. et al. 2021. Data Integration Enables Global Biodiversity Synthesis. *Proceedings of the National Academy of Sciences,* Vol. 118, No. 6, e2018093118. https://doi.org/10.1073/pnas.2018093118.

Hey, T., Tansley, S. and Tolle, K. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA, Microsoft Research.

Hooft, R. W. W. and Roos, M. 2025. Financing Models for Sustainable Data Reuse Infrastructure. *Data Science Journal,* Vol. 24, p. 29. https://doi.org/10.5334/dsj-2025-029.

International Science Council. 2025. Data and AI for Science: Key Considerations. Working Paper, Centre for Science Futures. https://doi.org/10.24948/2025.11.

Lin, D. et al. 2020. The TRUST Principles for Digital Repositories. *Scientific Data,* Vol. 7, p. 144. https://doi.org/10.1038/s41597-020-0486-7.

Mell, P. and Grance, T. 2011. *The NIST Definition of Cloud Computing*. Special Publication, Gaithersburg, Maryland, National Institute of Standards and Technology, pp. 800–145. https://doi.org/10.6028/NIST.SP.800-145.

OECD. 2015. Making Open Science a Reality. *OECD Science, Technology and Industry Policy Papers* 25. Paris, Organisation for Economic Co-operation and Development (OECD) Publishing. https://doi.org/10.1787/5jrs2f963zs1-en.

OECD. 2021. *OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity*. Paris, Organisation for Economic Co-operation and Development (OECD) Publishing. https://doi.org/10.1787/75f79015-en.

Perrier, L., et al. 2020. The Views, Perspectives, and Experiences of Academic Researchers with Data Sharing and Reuse: A Meta-Synthesis. *PLOS ONE,* Vol. 15, No. 2, pp. 1–21. https://doi.org/10.1371/journal.pone.0229182.

UNESCO. 2021. *UNESCO Recommendation on Open Science*. Paris, United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54677/MNMH8546.

Wilkinson, M. D. et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18.

World Bank. 2021. *World Development Report 2021: Data for Better Lives*. Washington, DC, World Bank. https://doi.org/10.1596/978-1-4648-3625-1.