Annals of Library and Information Studies
Vol. 71, December 2024, pp. 506-511
DOI: 10.56042/alis.v71i4.14295

Data for Development and the role of CODATA of the International Science Council (ISC)

Daisy Selematsela[ab], Usha MujooMunshi[c*] and Gitanjali Yadav[d*]

[a]University of Witwatersrand, Johannesburgdaisy.selematsela@wits.ac.za

[b]Vice-President, CODATA-ISC

[c]India International Centre, 40, Max Mueller Marg, New Delhi 110003, India ORCID: 0000-0002-2569-063Xu

[d]Scientist, NIPGR, New Delhi, India ORCID: 0000-0001-6591-9964

Received: 09 August 2024; Accepted: 19 September 2024

While the discussion on data for development gains traction in society especially with the focus on Sustainable Development Goals(SDGs), there is the need for more insight into the long-standing global perspectives on data and information for science. The International Science Council (ISC) established in 1932 as International Council for Scientific Unions (ICSU) aims to strengthen international science for the benefit of society. The key principle is the "Universality of Science"[1]which interprets science as a collective effort working for the common good, but a growing number of scientists, policy-makers, and social scientists argue that science is often too isolated from society to fulfil this promise. This brings in the concept of 'Open Science' to close the gap between science and society by democratizing scientific knowledge, for the benefit of everyone. The Committee on Data (CODATA), an interdisciplinary body of ISC is the focus of this paper, along with recent perspectives regarding its role and needs of science in the present information-driven society.

Keywords: Data for development, Data Center, Data distribution, Citizen Science, CODATA, ISC

Introduction

Data constitute the raw material of scientific understanding[1]. Scientific data gathering has a long history – in the past millennia information about solar and auroral activity was chronicled by the Chinese. It is noted that in the Western world, systematic geophysical measurements extends back for centuries and in the 18th and 19th centuries, data were exchanged from the early geomagnetic and seismic observatories through publication of annual station books. Whilst oceanographic and geological data were largely recorded in expedition reports [2]. Our knowledge today of the geomagnetic field, plate tectonics and ocean currents owe respect to these records management processes.

However, systematic measurements dates centuries back, mechanisms for data distribution and exchange are more recent. The first large-scale international scientific enterprises were the International Polar Years (IPY) of 1882-1883 and 1932–1933, and the International Geophysical Year (IGY) of 1957-1958. The International Council of Scientific Unions (ICSU) now known as the International Science (ISC) established the World Data Centre (WDC) system to serve the IGY and further developed data management plans for each IGY scientific discipline.

The data specifications were published in a series of Guides to Data Exchange which were originally issued in 1957 and subsequent updates in 1963, 1973, 1979, and 1987. It is maintained that the IGY planners were diligent in the handling of data. This is attested by the 1955 recommendation "that Data Centers should be prepared to handle data in machine-readable form, which at the time meant punched cards and punched tape". This indicates the precision at which at the time the issues pertinent today regarding the safeguard against catastrophic loss of data, and the pursuance of making data available for the convenience of data providers, users and citizen science [2].

An understanding of the evolution of data for development is transitionary to the institutionalisation of Data Centers through the World Data Centre systems by the respective scientific fields through the following activities:

- Collecting and cataloguing data and information in cooperation with other WDCs.

*Corresponding Authors
umunshi@gmail.com

- Maintaining the data in good condition.
- Providing data to users, at minimum costs to copying and distribution.
- Working with originators of data to improve documentation of data.
- Preserving important old data sets by converting them from tabular to digital form.
- Compiling specialised data sets for small-scale, regional and global geophysical research.
- Making data sets available on a variety of media, enabling users to search large data collections and transfer them to their home laboratory.
- Assessing technical issues of aging, error growth and lifetimes of data storage media.
- Combining data from various sources to derive data products, such as indices of solar or geomagnetic activity.
- Compiling numerical models to describe the time-varying and space-varying geophysical environment, such as the geomagnetic field and the upper atmosphere.
- Maintaining online information services related to the above activities.
- Assisting scientists to locate and access relate data not held in the system [2].

It should be noted that many of the former ICSU original ideas to support basic purpose of providing data and supporting scientists and citizen science across the global north and global south are still relevant in the Committee on Data (CODATA) agenda to guard against attempts by either national or commercial interests to restrict the flow of data and data for public good.

Role of the ISC's Committee on Data (CODATA)

CODATA[3], or the Committee on Data for Science and Technology, is an interdisciplinary scientific committee that works to improve the quality, reliability, and accessibility of data for research. CODATA's work is important for library science and other fields of science and technology because it helps to improve the management of data, which is essential for research.

Being at the forefront of promoting creation, access and dissemination of data, CODATA is making constant endeavours to address key issues to make data viability to the researchers and scientists for connecting the dots and evolving science led solutions that ensure sustainable development for the societal good. This is being done through standing committees and strategic executive led initiatives **(**https://codata.org/initiatives/strategic-programme/**)** several Task Groups (https://codata.org/initiatives/task-groups/) and Working Groups (https://codata.org/initiatives/working-groups/) to work towards realising the mission of the CODATA which primarily is to connect data and people to advance science and improve the world we live in. Of late, CODATA's strategic activities are divided into four prioritiesas depicted in Figure 1. The Decadal Programme aims to make data work for cross-domain grand challenges and the Global Open Science Cloud Initiative. The Data Policy activities promote principles, policies and practices for FAIR Data and Open Science; the Data Science program advanced the frontiers of the science of data and lastly, the Data Skills strategy builds capacity for Open Science by improving data skills and the functions of national science systems needed to support open data.

From Vision to Reality: CODATA Decadal Programme Drives FAIR Data Solutions

Decadal Programme on Making Data Work for Cross-Domain Grand Challenges began to come into fruition with the World FAIR project, a set of related Case Studies (https://worldfair-project.eu/12-months-of-the-worldfair-project/), and important work on units (https://codata.org/initiatives/task-groups/drum/), vocabularies (https://codata.org/ initiatives/decadal-programme2/fair-vocabularies/ iussp-codata-working-group-on-fair-vocabularies/) and a Cross-Domain Interoperability Framework (https://worldfair-project.eu/cross-domain-interoperability-framework/). Also notable is new impetus and direction for our International Data Policy Committee (https://codata.org/initiatives/data-policy/international-data-policy-committee/).

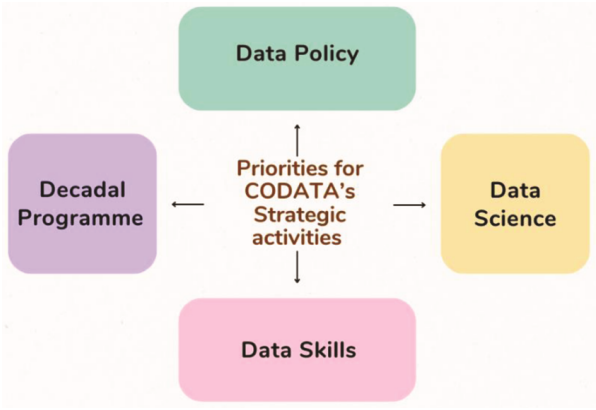

Fig. 1 — Major priorities for strategic activities by CODATA.

Several Indian scientists. administrators and data experts are members of various Task Groups, Working Groups and International Data Policy Committee (IDPC) of CODATA, there by substantially contributing to its various activities.

A Critical Challenge for Global Science

The major scientific and societal challenges of the 21st century—such as climate change, sustainable development, and disaster risk reduction—demand cross-disciplinary research capable of understanding complex systems through large-scale, machine-assisted analysis. However, this potential is hindered by our limited ability to access and integrate heterogeneous data across domains. Currently, inefficient data practices consume an estimated 80% of research budgets, impeding progress.

Addressing these complex issues requires data that is machine-readable and actionable, supported by cutting-edge technologies. For this, data must be accompanied by rich metadata, transparent, and comprehensible, enabling the extraction of meaningful insights from complexity. A key enabler of data-driven science is an ecosystem that ensures data is FAIR (Findable, Accessible, Interoperable, and Reusable) for both humans and machines. This ecosystem must include automated data stewardship and standardized terminologies and metadata specifications[3].

A Global Consensus for Core Interoperability

To achieve this vision, a global consensus on core technologies and semantic solutions is essential for integrating data across disciplines. CODATA, on behalf of the International Science Council (ISC), proposes a decade-long initiative, "Making Data Work for Cross-Domain Grand Challenges" (2020–2030), aimed at overcoming these barriers. The program adopts a three-pronged approach, with collaboration between the following key areas as shown in Figure 2.

1. **Enabling Technologies and Best Practices for Data-Intensive Science**: Collaborating with domain and technology experts, the program will define the enabling technologies and best practices needed for data-driven discovery across disciplines.
2. **Mobilizing Domains and Breaking Down Silos**: Proactive engagement with international scientific organizations and stakeholders will foster data interoperability and collaborative work across fields.

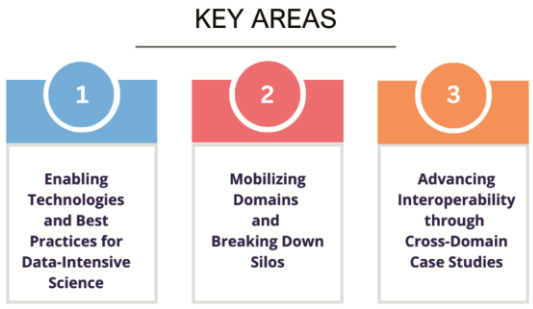

Fig. 2 — Three key areas for overcoming challenges.

3. **Advancing Interoperability through Cross-Domain Case Studies**: The program will apply its findings to real-world case studies, including sustainable development, disaster risk reduction, urban health, resilience, and infectious diseases.

This initiative seeks to transform global data practices, unlocking new potential to solve some of the world's most pressing challenges.

Unlocking the Power of Data: Building a FAIR Data Ecosystem

Our ability to harness the power of data is hindered by fragmented and incompatible data ecosystems. To unlock the full potential of data-driven science, we must create a global infrastructure that enables data to be Findable, Accessible, Interoperable, and Reusable (FAIR). This requires a concerted effort to develop common standards, tools, and practices that facilitate data sharing and collaboration across domains. By investing in data infrastructure and fostering international cooperation, we can empower researchers to tackle complex challenges and drive innovation for a sustainable future.

In the WorldFAIR project (https://worldfair-project.eu/), CODATA (the Committee on Data of the International Science Council) and RDA (the Research Data Alliance) work with a set of 11 disciplinary and cross-disciplinary case studies to advance implementation of the FAIR principles and, in particular, to improve interoperability and reusability of digital research objects, including data. Particular attention is paid to the articulation of an interoperability framework for each case study and research domain.

Digital Representation of Units of Measurement

Yet again, the CODATA Digital Representation of Units of Measurement Task Group (DRUM TG) spearheads a global initiative to ensure accurate and interoperable data across scientific disciplines.

Aligned with CODATA's Decadal Programme, DRUM TG fosters awareness and implementation of standardized digital unit representation (DUR) – the cornerstone of FAIR data.

Current Challenges:

- Scientific data increasingly relies on digital formats, yet unit representation suffers from inconsistency and lack of machine readability.
- Inconsistent unit representation hinders data normalization and interoperability, creating major barriers to cross-domain research.

DRUM TG's impactful mission is to advance data interoperability, DRUM TG is dedicated to developing recommendations for standardized unit representation in digital formats. These efforts include establishing guidelines for unit annotation in data systems and promoting the adoption of the DRUM system. Furthermore, DRUM TG aims to create a Units of Measure Interoperability Service (UMIS) to facilitate seamless unit conversion and exchange across various digital platforms.

The urgency of interoperable units: Units are the lifeblood of scientific data. Without accurate and machine-readable unit representation, data is vulnerable to misinterpretation and AI algorithms may generate erroneous results.

Bridging the Data Chasm: The Cross-Domain Interoperability Framework (CDIF)

The burgeoning field of data-driven science faces a critical bottleneck: fragmented data ecosystems. To unlock the full potential of FAIR data, especially across diverse scientific domains and institutional boundaries, the **Cross-Domain Interoperability Framework (CDIF)** emerges as a transformative solution.

CDIF transcends mere standards by establishing a robust set of practical guidelines for implementing FAIR principles across disciplines. This framework tackles the crucial need for harmonization in how FAIR data is managed, facilitating seamless interactions between research systems. Notably, CDIF targets "grand challenge" research questions that inherently demand interdisciplinary collaboration.

Core Principles, Not Prescriptive Mandates: CDIF eschews the pursuit of an all-encompassing FAIR implementation guide. Instead, it focuses on identifying a core set of functions essential for successful FAIR systems. This pragmatic approach leverages existing, widely embraced standards (e.g., Schema.org, DCAT) while advocating for well-established practices within established FAIR networks. In areas lacking established standards, CDIF outlines potential approaches, fostering ongoing development. This commitment to flexibility underscores CDIF's ability to adapt to emerging technologies and practices in a dynamic data landscape.

Functional Areas: Bridging the Gap Between Data and Discovery: CDIF tackles several critical areas to ensure seamless interoperability as shown in Figure 3.

- **Discovery of Data:** Facilitating the identification of both static datasets and queryable services that provide valuable data.
- **Data Integration:** Establishing a framework for describing data precisely to enable seamless integration across research domains.
- **Controlled Vocabularies and Ontologies:** Promoting the adoption and use of standardized vocabularies and ontologies for unambiguous data interpretation.
- **Provenance and Process Description:** Ensuring traceability of data by capturing its origin and processing history.
- **Universal Temporal and Spatial Information:** Establishing a standardized approach for conveying temporal and spatial data across domains.
- **Machine Learning Training Data:** Providing clear guidelines on preparing and describing data for utilization in machine learning algorithms.

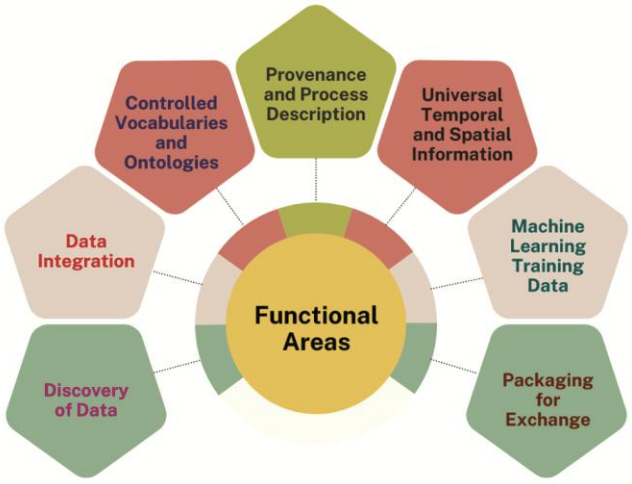

Fig. 3 — Functional areas for bridging the gap between data and discovery.

- **Packaging for Exchange:** Enabling the efficient and standardized exchange of data and associated metadata across systems.

CDIF in Action: A Pioneering Profile for Data Discovery: CDIF's initial profile, focused on the discovery of static data sets, serves as a testament to the framework's practical application. This profile showcases the level of detail and guidance researchers can expect in forthcoming areas of focus. We encourage feedback on this initial profile to further refine and empower CDIF's transformative potential.

By fostering a practical and collaborative approach, CDIF positions itself as a cornerstone for bridging the data divide in scientific research. This framework paves the way for a future where interoperable data empowers researchers to tackle the most critical challenges facing our world.

Shaping the Future of Data: The CODATA International Data Policy Committee

The **CODATA International Data Policy Committee (IDPC)** stands as a cornerstone for advancing responsible and effective data governance on a global scale. As a subsidiary of CODATA, the IDPC plays a pivotal role in promoting open science and FAIR data principles while supporting CODATA's broader strategic objectives.

A Global Forum for Data Policy: Comprised of a diverse group of experts, the IDPC serves as a platform for collaboration, research, and advocacy. By fostering dialogue among policymakers, researchers, industry leaders, and civil society, the IDPC develops data policies that address the complex challenges of our digital age.

The IDPC's current focus areas reflect the evolving landscape of data science and its profound impact on society as shown in Figure 4:

1. **Data Quality and Integrity:** Ensuring the reliability and trustworthiness of data through robust policy frameworks.
2. **Science in Crisis:** Developing data policies that support scientific research during times of crisis, such as natural disasters or public health emergencies.
3. **Data Education:** Promoting data literacy and responsible data practices through educational initiatives.
4. **Artificial Intelligence:** Addressing the ethical and societal implications of AI development and deployment through data-centric policies.

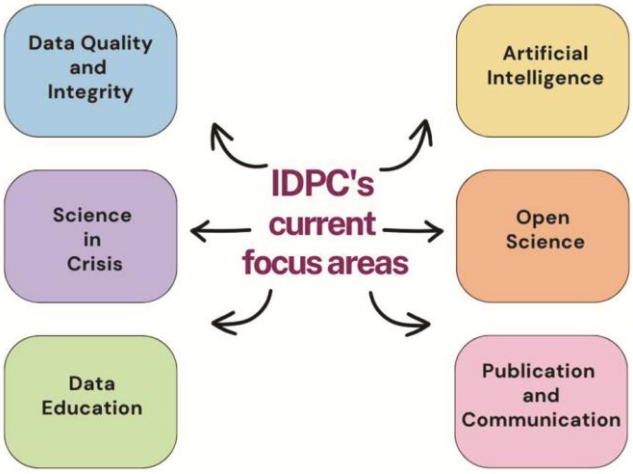

Fig. 4 — The focus areas adopted by IDPC for making data policies.

5. **Open Science:** Fostering open access to data, research, and publications to accelerate scientific progress and innovation.
6. **Publication and Communication:** Enhancing the dissemination of scientific knowledge through effective data policies.

A Vision for the Future: Beyond these immediate focus areas, the IDPC's long-term vision is to create a data-driven world that is equitable, sustainable, and beneficial to all. By shaping data policies that promote innovation, protect privacy, and advance ethical practices, the IDPC is making a significant contribution to the future of science and society.

Case Study: The CoARA ERIP Working Group: In view of the increasing dependence on Data in all fields of science, members of CODATA IDPC, including several Indian representatives have joined the Coalition for Advancing Research Assessment (CoARA), with the vision to advance assessment of research, researchers and research organisations, recognising diverse outputs, practices and activities that maximise the quality and impact of research. Within this, placing a special priority on basing the evolution of research assessment primarily on qualitative judgement (for which peer review is central), members of the CODATA IDPC have joined forces to create a new working group on ‘Ethics and Research Integrity Policy in Responsible Research Assessment for Data and Artificial Intelligence’ (ERIP). ERIP builds global expertise to address the transformative cross-disciplinary impact of data and AI on research culture (values, processes, structures, perceptions) supported by data and AI integrity for the ethical development of AI in

research and institutional assessment framed in human-centric quantitative and qualitative metrics/indicators for data/AI research activities. The mission is to develop policy, guidance, and tools for advancing research assessment that promote a responsible culture for the assessment of data and AI in research, fostering responsibility, transparency, and societal benefit. The CoARA ERIP also aims to foster stakeholder engagement through facilitated dialogue and collaboration among researchers, policymakers, funding agencies, and other actors across boundaries of nationality, themes and cultures to ensure diverse and global perspectives in the development and implementation of ethical research assessment policies. As of September 2024, the ERIP co-chairs have met approximately 12 times, deliberating on the organizational structure, objectives and workstreams. ERIP was presented by our own team (UM and GY) during India's National Conclave on 'AI and Ethics: Perspectives from Industry and Academia' on 20th February 2024. This was followed by participation in the United Nations University first AI Conference and contribution to the founding of the UNU AI Network; Macau in April 2024. The CoARA ERIP was presented during the AI and Climate Expert Meeting at the United Nations University (UNU EHS) in Bonn in July 2024. In summary, ERIP mediated by the CODATA IDPC serves as an excellent case study of a team actively recruiting member institutions from the Global North as well as the Global South, while engaging a broad global community, encompasing all continents and many countries already represented in CoARA.

Conclusion

It should be noted that the economic, legal, and technological restrictions that are placed on public-domain scientific data especially in developing countries poses challenges in the use of data for development. The trends to privatise for example, governmental public good functions and to commercialise more of the academic sectors research outputs – from an economic perspective this trend can support significant research advances and economic benefits, however they are not without their own social costs to knowledge generation and societal impact.

There is legitimate public-policy reason for limiting access to certain types of data, including appropriate national security restrictions, the protection of privacy and confidentiality, and the protection of private (as opposed to government) intellectual property rights[4]

With regards to scientific data resources, most databases and data centers are managed directly or funded by government ministries and are often subject to restrictions. There is often a challenge to the adoption of open-access model because of either cultural, institutional, and political factor.

In summary, CODATA and its various domains are intricately engaged in delivering methods and tools to enable digital contributions to science and knowledge in research programs, while ensuring that research ethics and integrity of scientific outputs are at pace with the advancing use of data and impact of AI. The design of flexible community-based working group structures ensure accommodation of diverse needs, expertise, and perspectives of members, allowing for fair representation and meaningful contributions across time zones that are adaptive to workloads of the experts. The outputs are constructed on the basis of open and transparent inter-institutional, inter-actor, cross-cultural, and multi-regional research and discussion.

CODATA's involvement in advocating for data accessibility for public good strengthens arguments in favor of greater unrestricted access to data and offers policy guidelines supporting 'open availability.' The case for revising access policies—particularly for governmental scientific data—can be justified by considering national self-interest and through comparisons with policies in other countries, which can further enhance data accessibility for development.

References

1. "Based on presentation by Carthage Smith, International Council for Science, available at http://www7.nationalacademies.org/usnc-codata/CarthageSmithPresentation.ppt.
2. Guide to the World Data Center System. International Council of Scientific Unions: Panel on World Data Centres (Geophysical, Solar and Environmental), APRIL 1996
3. https://codata.org/about-codata/
4. Strategies for Preservation of and Open Access to Scientific Data in China. Summary of a workshop. U.S. National Committee for CODATA – Board of International Scientific Organisations. Paul F. Uhlir and Julie M. Esanu: Rapporteurs 2004http://newton.nap.edu/catalog/11710.html