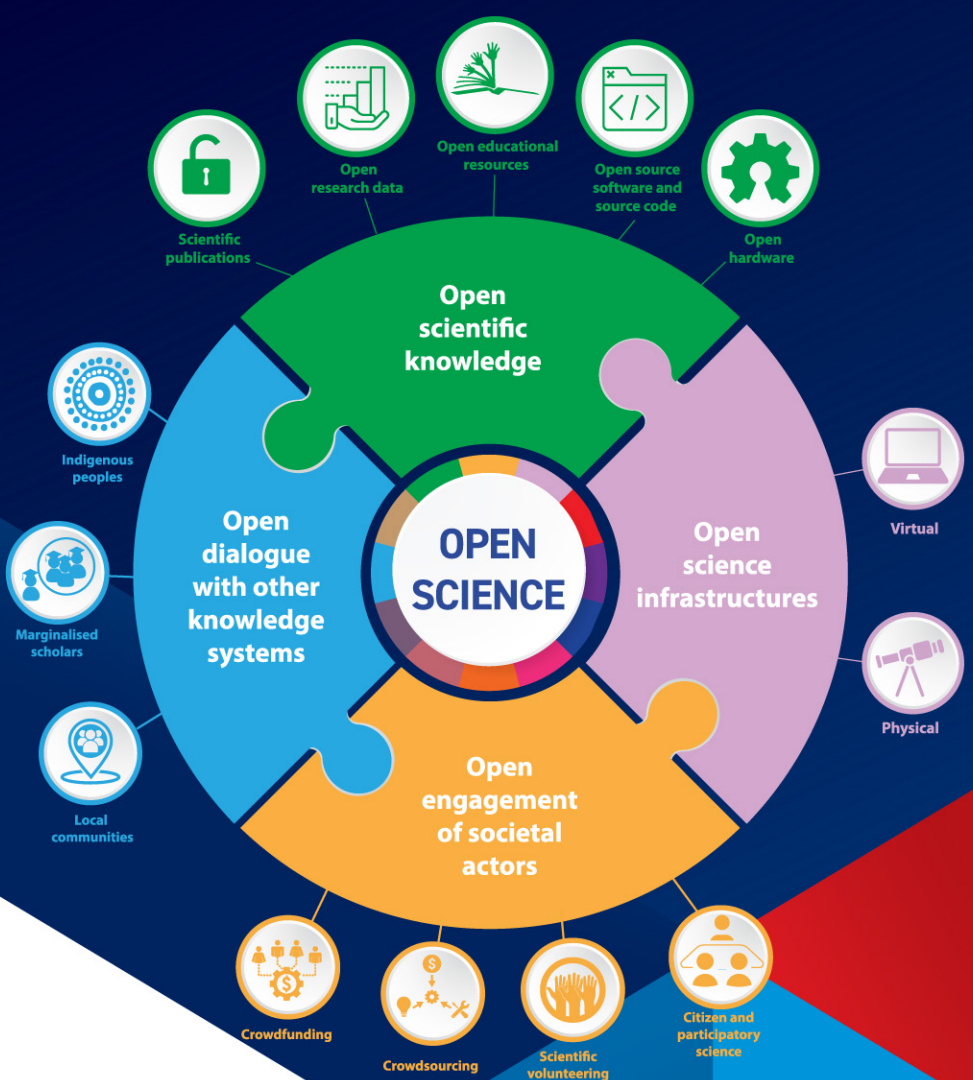
International Conference

on

Open and FAIR Data Ecosystem Principles, Policies and Platforms

Abridged Proceedings

11 - 13 September 2023



Organised By:

INDIA INTERNATIONAL CENTRE IIC

PUNE KNOWLEDGE CLUSTER PKC

International Conference

on

Open and FAIR Data Ecosystem
Principles, Policies, and Platforms

Abridged Proceedings

11–13 September 2023

India International Centre, New Delhi

Organisers

Shree Guru Gobind Singh Tricentenary (SGT) University, Gurugram
India International Centre (IIC), New Delhi
Pune Knowledge Cluster (PKC), Pune

June, 2024

International Conference on Open and FAIR Data Ecosystem Principles, Policies, and Platforms: Abridged Proceedings

ISBN:

Editorial Team

Prof (Dr) Prabir G. Dastidar – Chief Editor
Ex-Adviser/Scientist, Ministry of Earth Sciences, GoI, New Delhi
Shree Guru Gobind Singh Tricentenary (SGT) University, Gurugram

Dr B.S. Aggarwal – Editorial Consultant
Ex-Senior Scientist, CSIR, New Delhi

Dr Anup Kumar Das – Editorial Support
Jawaharlal Nehru University, New Delhi

Published by
Prof. Prabir G. Dastidar
on behalf of
Shree Guru Gobind Singh Tricentenary (SGT) University, Gurugram
India International Centre (IIC), New Delhi
Pune Knowledge Cluster (PKC), Pune

Printed at:

Contents

Agri-tech

Data Policy, Practices, Ethics & Infrastructure

Ocean & Atmospheric Science

Digital Humanities

Acknowledgements

Organizers are thankful to the following organizations/agencies for financial and other support to the Conference.

Financial Support

Ministry of Earch Sciences, Govt. of India, New Delhi

Indian National Science Academy (INSA), New Delhi

Indian Council of Social Science Research (ICSSR), New Delhi

Organisers

.....

.....

.....

Background

Considering the significant role science plays for the well-being of the society, unfolding the mysteries of ~~universe~~ and its ubiquitous nature in our life, a revolutionary concept of "Open Science" was adopted by the UNESCO in its 29th Session. It stands on four pillars ~~¾~~ Open Scientific Knowledge, Open Science Infrastructure, Open Engagements of Societal Actors and Open Dialogue with other knowledge systems. The concept of open science inherently provides transparency and access to data, information and infrastructure to all. The research outcomes should be accessible to all concerned, transcending the social, intellectual and economic barriers.

It is now recognised worldwide that to take the scientific heritage of mankind forward, the concept of open science along with open data, open access has to be popularised at both government and institutional levels. The concept and basic principles have to be mandated for appropriate storage, retrieval and dissemination of data. Alongwith open data, another concept of data infrastructure is FAIR data, which encompasses the four basic principles of data ecosystem — **F**indability, **A**ccessibility, **I**nteroperability and **R**eusability.

Considering the importance of data and information, the Ministry of Science and Technology, Government of India had formulated a National Data Sharing and Accessibility Policy (NDSAP) in 2012. The policy complements the FAIR data principles. It is felt that a well-structured national guideline will benefit the society by maximising the usage of data, avoiding duplications, creating ~~opportunity~~ to integrate, making ownership information visible, and above all infusing better decision-making based on quality data.

The recently concluded G20 in 2023, New Delhi leaders' declaration has discussed eloquently on various aspects of data generation, its storage and dissemination. The importance of data for development (D4D) for realising the sustainable development goals (SDGs) globally. Some other international summits and organisations that have endorsed the usage of FAIR data include G20 Hangzhou Summit in 2016, the G20 Leaders Statement of 2016, statement issued by a group of Australian organisations on FAIR Access to Australia's Research Outputs, the joint agreement of Germany, Netherlands and France to establish an international office to support the FAIR initiatives**.**

In this backdrop, considering the importance of data and the role it plays, the SGT University, Gurugram, Pune Knowledge Cluster and India International Centre, New Delhi, have come together to organise an International Conference on FAIR and Open Data at India International Centre, during 11-13 September 2023. It is aimed to popularise the concept of Open data in India in line with the government policies.

A well-represented delegates from the prestigious national and international institutes, government departments, Committee on Data of the International Science Council (ISC), etc. are likely to attend the Conference.

About the Conference

Rapidly evolving digital technologies are transforming the way in which research is conducted globally, with data science becoming the most important element across all domains. But, the extensive use of data science brings its own challenges. A major issue in data science is in the analysis of vast quantities of heterogeneous data, both in formats and content. To mitigate this and numerous other anticipated issues, nations, regions, and disciplines have begun to build research infrastructures to aid data management, stewardship, and data analysis. Open science policies and practices have further galvanized these developments by various key stakeholders. Understanding these policies, adapting them to domain-specific and regional requirements and revising them with changing technologies and aspirations, and developing suitable practices are crucial for realising the full potential of data analysis.

Therefore, a need was being felt to bring different stakeholders ~~at~~ a platform for sharing expert insights on the above issues and enable intense community interaction to better understand global trends. The conclusions from the meeting should be useful in identifying existing impediments in some of the major socio-economically relevant disciplines, and mobilising efforts for better data harnessing and data capture. Most importantly, the meeting should help significantly to chart a roadmap, suitable for the Indian ecosystem, to guide the further course of action on Open Data.

The themes included in the conference broadly focussed on open science and open data, FAIR, research data management, data ethics, data licensing, data stewardship, data science education and training, data research infrastructures and data platforms.

Preface

All knowledge-based disciplines including science and technology, medicine, economics and finance, social science and humanities are becoming increasingly data driven, and so is every aspect of human activity. It has therefore become imperative to develop expertise in the collection and curation of data, data processing, data management and data distribution. Many complex problems being faced by humanity, like climate change and environmental sustainability require a multidisciplinary approach. The interdisciplinary work requires data from different domains, with data being quite different in format and content from domain to domain. Interoperability standards and tools therefore have to be developed for the seamless exchange of data. While the generation of mechanisms for the collection and curation in an interoperable manner is vital, it is equally important to make the databases freely available. That has to be done following the FAIR principles of Findability, Accessibility, Interoperability and Reusability FAIR.

The conference, which was organized at the India International Centre, New Delhi, during September 11-13, 2023, provided a forum for sharing expert insights on the above issues and to develop understanding of global trends. Conclusions from the meeting should be useful in the future for identifying and resolving difficulties in some of the relevant disciplines, and for mobilising efforts for better data capturing. The meeting provided some pointers to chart a roadmap, suitable for the Indian ecosystem, and to guide the further actions on Open Data. The themes included in the conference mainly focussed on open science and open data, FAIR, research data management, data ethics, data licensing, data stewardship, data science education and training, data research infrastructures and data platforms. Special attention was paid to the four verticals of National Health, Agri-Tech, Ocean & Atmospheric Science, and Digital Humanities and Social Sciences.

At the conference a number of national and international experts delivered keynote addresses, invited talks and contributed papers; some of the experts addressed the meeting virtually. There was a panel on the last day of Conference on "Open and FAIR Data for Successful Digital Transformation in Academia and Industry". There was a Digital Twins Roundtable organised during the meeting at which there were a number of presentations and discussions on Digital Twins technology applied to a very broad range of topics including AI, chemistry, food and beverages, healthcare, infrastructure and architecture with focus on manufacturing, intelligent infrastructure and bridges, life sciences, logistics, mobility and transportation, natural resources and telecommunications. The conference proceedings include abstracts of some of the talks which were presented at the meeting. Generous support for these organisations was critical for the success of the Conference.

Dr Ajit Kembhavi
Pune Knowledge Cluster
Pune, India

National Organising Committee

Conference Chairman

Prof (Dr) Ajit Kembhavi
Principal Investigator of the Pune Knowledge Cluster, Pune

Conference Director

Dr Usha M. Munshi
India International Centre (IIC), New Delhi

Organising Secretary

Prof (Dr) Prabir G. Dastidar
Shree Guru Gobind Singh Tricentenary (SGT) University, Gurugram

Dr Priya Nagaraj
Pune Knowledge Cluster (PKC), Pune

Conference Team at SGT University

Prof (Dr) Shalini Kapoor
Dr Simranjeet Singh
Sri Hunny Arora

Program Committee

Dr Jagdish Arora, National Board of Accreditation (NBA), New Delhi

Prof (Dr) Prabir G. Dastidar, SGT University, Gurugram, Ex-Adviser, Ministry of Earth Sciences

Shri Rajesh Gera, NIC, New Delhi

Dr Sridevi Jade, CSIR-4PI, Bengaluru

Prof J.P. Singh Joorel, INFLIBNET Centre, Gandhinagar

Dr Sachidanand Joshi, IGNCA, New Delhi

Prof Ajit Kembhavi, Principal Investigator, PKC, Pune

Shri Gopal Iyenger, Advisor, MoES, New Delhi

Dr Vivek Kumar, S&T Office, New Delhi

Dr G. Mahesh, CSIR, New Delhi

Dr Usha M. Munshi, IIC, New delhi

Dr Priya Nagaraj, PKC, Pune

Shri N.V. Sathyanarayana, Chairman & MD, Informatics (India) Ltd., Bengaluru

Shri G. Srinivasan (India), Director, Department of Science & Technology, India

Prof Azizuddin Khan, Psychophyiology Laboratory, Indian Institute of Technology (IIT), Bombay

Dr D.R. Pattnaik, Scientist, Indian Metrological Department (IMD), New Delhi

Dr Uday Bhaskar, Scientist, Indian National Center for Ocean Information Services (INCOIS), Hyderabad

Ms Namita Gupta, Advisor/Scientist G, Knowledge Involement in Research Advancement through Nurturing (KIRAN) Division, Department of Science and Technology

Prof (Dr) Reshu Sanan, Faculty of Dental Sciences, SGT University, Gurugram

The Conference at a Glance

Introduction

An International Conference on Open and FAIR Data Ecosystem: Principles, Policies and Platforms was held at the India International Centre, New Delhi, during 11-13 September, 2023. The Conference on Open and FAIR Data Principles, Policies, and Platforms brought together stakeholders from academia, industry, and government organizations to discuss strategies for implementing Open and FAIR data initiatives. Open data refers to making research data publicly available without restrictions, while FAIR data complies with the principles of **Findability, Accessibility, Interoperability, and Reusability**. The Conference aimed at highlighting the advancements in data-sharing practices, ~~promote~~ open science policies, and fostering collaborations among diverse communities working with open and FAIR data.

The International Conference on Open and FAIR Data Ecosystem: Principles, Policies, and Platforms served as a platform for knowledge exchange among professionals working globally dedicated to ensuring openness in scientific research processes. By discussing emerging trends, identifying challenges, and sharing best practices, the Conference contributed to the ongoing efforts on improving data-sharing methods and fostering global collaborations within the scientific community.

As the first outcome of the New Delhi G20 submission held during 9-10 September, 2023, the Conference was attended by one hundred eighty-one delegates from almost all the states of the country and several foreign countries like ~~USA,~~ UK, South Africa and Germany.

The Conference was inaugurated by Prof. Ajay Kumar Sood, the Principal Scientific advisor to the Government of India. There were 10 Technical Sessions, one Panel Discussion and a Round Table on Digital Twins during the Conference. A total of 50 papers were presented by the experts and discussed threadbare to arrive at suitable recommendations for implementation by national governments, national and international organisations, research institutions and other stakeholders.

The Chief Guest in his inaugural address mentioned, "it is great to know that this Conference is one of the first outcomes of the New Delhi G20 Declaration. **This is a big opportunity for India to take ~~lead~~ in setting a global governance approach to AI."**

He further reiterated that the G20 New Delhi Declaration has stressed "Harnessing Artificial Intelligence (AI) Responsibly for Good and All" and stated that "*The rapid progress of AI promises prosperity and expansion of the global digital economy. We endeavour to leverage AI for the public good by solving challenges in a responsible, inclusive and human-centric manner while protecting people's rights and safety. To ensure responsible AI development, deployment and use, the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, biases, privacy, and data protection must be addressed. To unlock the full potential of AI, equitably share its benefits and mitigate risks, we will work together to promote international cooperation and further discussions on international governance for AI. To this end, we:*

- *Reaffirm our commitment to G20 AI Principles (2019) and endeavour to share information on approaches to using AI to support solutions in the digital economy.*
- *Will pursue a pro-innovation regulatory/governance approach that maximizes the benefits and takes into account the risks associated with the use of AI.*
- *Will promote responsible AI for achieving SDGs. "*

India is the incoming Council Chair of the Global Partnership on Artificial Intelligence. Today, 29 international partners (including USA, and EU members) have joined together to guide the responsible development and use of artificial intelligence, grounded in human rights, inclusion, diversity, innovation and economic growth.

Dr Usha Mujoo Munshi, Conference Director, welcomed the guests, Ambassador Shyam Sharan presented the introductory remarks and Prof. Ajit Kembhavi communicated about the aims and objectives of the Conference.

Mr Simon Hodson, Executive Director, CODATA, ICS, France, while initiating discussions on the theme of the Conference in the first Plenary Technical Session reiterated that as the 'Committee on Data of the International Science Council (ISC)', CODATA supports and promote the ISC's mission of 'advancing science as a global public good' by promoting Open Science and FAIR data. CODATA's vision is of a world in which science is empowered to address universal challenges through the transparent, trustworthy and equitable use of data and information.

The CODATA aims to "Connect data and people to advance science and improve our world". To achieve this, CODATA convenes a global expert community and provides a forum for international consensus building and agreements around a range of data science and data policy issues, from the fundamental physical constants to cross-domain data specifications.

Mr Hodson recommended developing **interoperability frameworks** for FAIR sharing within disciplines and for interdisciplinary research: Research communities need to be supported to

develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure. **To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant.**'

UNESCO Recommendation calls for various stakeholders, to promote Community agreements, concluded in the context of regional or global research communities, and which define community practices for data sharing, data formats, metadata standards, ontologies and terminologies, tools and infrastructure. International scientific unions and associations, regional or national research infrastructures and journal editorial boards each has a role to play in helping develop these agreements. In addition, convergence between the various semantic artefacts (particularly vocabularies, taxonomies, ontologies and metadata schema) is essential for the interoperability and reuse of data for interdisciplinary research.

The Conference discussed recommending good practices and standards for a set of functions related to the FAIR principles:

- Discoverability (DCAT, schema.org)
- Access, and negotiate access to non-public data (ODRL, DPV, DUO)
- Assess and Integrate/Combine:

  - Understand and transform data structure (DDI-CDI)
  - Understand semantics (SKOS/XKOS, OWL, SSSOM)
  - Determine origin/context (PROV-O, I-ADOPT/O&M)
  - Understand processing (DDI-CDI, SDTL)
- 'Universals' (time, place, units of measurement, authoritative domain recommendations)
- Establish connections to higher-level research and management processes, track citation/use, and connect to ROI assessments (CERIF, GAMSO)

Themes and Sessions

The Conference selected and discussed the following themes for exploring FAIR implementation and the application of the CDIF:

- Health
- Agritech
- Ocean and Atmospheric Sciences
- Policy, Practices, Ethics and Infrastructure.

Health

India has implemented a national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research:

1. Semantic,
2. Standard formalism to represent and store the semantics and the data.
3. Supports the integration and translation of this formalism into existing application-targeted models.

Ayushman Bharat Digital Mission — Creating India's Digital Health Ecosystem.

Ayushman Bharat Health Account (ABHA) number is a 14-digit number that will uniquely identify you as a participant in India's digital healthcare ecosystem. About 23.3 crore accounts have already been created.

Integrated Health Information Program

Mission: To strengthen disease surveillance in the country by establishing a decentralized State-based surveillance system for epidemic-prone diseases to detect the early warning signals, so that timely and effective public health actions can be initiated in response to health challenges in the country at the Districts, State and National levels.

Program for Monitoring Emerging Diseases (ProMED)

The **Program for Monitoring Emerging Diseases** (ProMED) is a program of the **International Society for Infectious Diseases** (ISID). ProMED was launched in 1994 as an Internet service to identify unusual health events related to emerging and re-emerging infectious diseases and toxins affecting humans, animals and plants.

The **'Indian Biological Data Centre (IBDC)'** is the first national repository for life science data in India. IBDC is mandated to archive all life-science data generated from the publicly funded research in India. The data centre is supported by the Government of India (GOI) through the Department of Biotechnology (DBT).

Fundamentally, IBDC is committed to the spirit of data sharing as per FAIR principles.

IBDC would have two major data access types:

Open access/time-release access: Data submitted at IBDC would be freely accessible across the globe as per international open-access standards. The submitter, however, may choose to restrict the data access for a defined period.

Restricted access: The data would not be made accessible freely. It can only be accessed through prior permission through IBDC from the original data submitter.

The Challenges include:

- Data and data access are critical to society during times of crises (e.g., the recent pandemic)
  - Unavailability of data — GISAID
  - Data-sharing hesitancy and/or restrictions
  - Incomplete and missing data
  - Absence of relevant metadata
  - Data-access restrictions
  - Incompatibilities across databases
- Post-pandemic – long COVID … inability to conduct research because of lack of (harmonized) data.
- Inadequate interoperability of data and data systems.

Another major challenge is that Most health records are **handwritten**. Their conversion to a "findable" electronic formal is a formidable problem.

Agritech

Dr Rama Devi Lanka, Government of Telangana, while talking on emergent technologies elaborated on the Vision and Objectives of the Telangana State Open Data Policy. The objective of the Telangana state Open Data Policy is to provide a framework for the proactive disclosure of Government data by the various Public Authorities. She mentioned that open data promotes transparency and accountability in government and other institutions, drives innovation and economic growth by providing a valuable resource for R&D and helps to address social and environmental issues by providing information that can be used to inform policy and decision-making. It helps in maximising use for the benefit of the community, Cost savings in Data collection, Maximum Integration of individual databases with the adoption of common standards, Reduces Information asymmetry and Enables Better Decision Making

The kind of agriculture data being collected and processed include Agricultural credit, Agricultural insurance, Logistics, Storage-related services, Market-related services, Payment services, Farm advisories, Pest and disease management, Water management, Quality assurance, Soil health, Land records, Research, Innovation and Education, Data principal use.

She informed that the Agricultural Data Exchange Platform (ADeX) developed for Telangana State is an open- source platform empowering AI innovators to build solutions for farmers with an exchange of quality data. Yet another project namely the **Saagu Baagu** project has provided 7,000 chilli farmers in the district access to agritech services specific to challenges across the value chain.

Smart Nutrition Management of Soil data system is also a step towards increasing awareness and sustainability, helping farmers understand the nature and the nutrient composition of the soil, This will help farmers make better decisions for the optimal use of the Agri-inputs and increase their yield.

Another platform Data In Climate Resilient Agriculture (DiCRA) has been developed for the farmers with the following features:

- World's First Digital Public Good on Climate-Resilient Agriculture,
- Demystifies Climate Resilience in Agriculture through measurable data parameters,
- Open Geospatial platform curated by data collaborative of 9 organizations and 100+ volunteering data scientists,
- Provides Open Access, Open Data, Open Software, and Open AI,
- Provides high-resolution high-frequency maps that one can zoom to the farm level to understand climate resilience at the village level,
- Provides historical and decadal patterns to understand which are climate-resilient farms and take lessons from them,
- Customized advisory to 21,000 Villages on climate-resilient agriculture using DiCRA,
- Plug & play design can be applied to any geography in the world. Scaling fast to 5 states in India and 2 countries in Latin America.

Dr Amarnath Gupta, Sandiego, USA while sharing his experience elaborated that a survey of about 250 small business/farm owners/aspirants in the San Diego/Imperial Valley, USA region was conducted to

- Discover Actionable Knowledge
    - o Significant knowledge gaps about how to start, sustain and grow fresh food businesses
    - o Lack of data on local market opportunities, funders and supports
- Ability to Interpret and Navigate Bureaucracy
    - o Complex government permits and certifications
    - o The process of finding appropriate funding resources is cumbersome
- Ability to Develop My Resource Network
    - o Need to connect large-impact investors to local microlenders
    - o Weak connections between local farms and restaurants, microlenders and small businesses.

It helped in developing FAIR as an information supplier-centric guideline to develop user- centric guidelines for delivering the information that:

- U → Usable for specific user profiles
- N → Navigable through the network of related information

- I → Intelligible for a user based on his/her profile and preferences
- T → Timely based on the user's evolving context

Ocean and Atmospheric Science

There are two very distinct views of the scientific method, the first of which was expressed quite eloquently by Abraham Flexner more than 80 years ago in the book, 'The Usefulness of Useless Knowledge' and pointed out the role of imagination and curiosity. In the current age of data, this may be supplemented by the humility to rein in the imagination courage and perseverance. The second has been variously called 'The Fourth Paradigm', etc. due to the unprecedented volumes and complex datasets available. The talk described some of the extensive collections of earth science datasets across the world, including those mentioned below.

The talk described the important distinctions between observational data and the data generated by earth system models, which are essentially digital twins of the earth with varying levels of complexity being modelled, from clouds to sea-ice, to ecology, etc. This naturally led to the discussion of some important questions related to earth science data: (1) What is the source of these data? Observations? Model? or A combination of the two? (2) What do these data represent? (Interpretability) – e.g. temperature measurement (easy to interpret); model prediction of temperature at IIC (average over a region) and (3) Do more data give us more information about the system? These were illustrated in the context of weather prediction problems, whose solution depends crucially on the process of data assimilation.

Three key concepts related to earth science data that are crucially associated with data assimilation were described next: (1) Uncertainty — of course, it is important for all datasets, but its representation and quantification are particularly difficult in earth sciences. (2) Dynamics — the time-evolving systems lead to spatio-temporal variability that requires new descriptions and (3) Complexity — the system spans multiple spatial and temporal scales and consists of many sub-systems, many of which themselves are complex systems, such as cloud systems, sea-ice, bio-geo-chemistry, ecosystems, etc. The talk ended by drawing a parallel between the complex earth systems and biological and health systems.

Data Policy, Practices, Ethics and Infrastructure

Dr Neeta Verma, ex DG, National Informatics Centre, New Delhi, while chairing the session spoke about the Future of Data Infrastructure. She opined that Data is becoming essential to our daily life! Data's dramatic expansion **demands new approaches and mindsets** to deliver consistent, secure and timely value. The global mobile application market size was valued at USD 206.85 billion in 2022 and is expected to grow at a compound annual growth rate **(CAGR) of 13.8% from 2023 to 2030**.

This calls for proper data infrastructure in collecting insights from data in a better way. Data infrastructure includes **hardware, software, networking, services, policies, and more—that enable data consumption, storage, and sharing**. Data infrastructure consists of data assets, the organisations that operate and maintain them and processes, policies and guides describing how to use and manage the data. Data infrastructure is the ecosystem of technology, processes, and people responsible for the collection, storage, maintenance, and distribution of data within an organization.

A Robust/ Agile Data Infrastructure is Important for Better Data Storage and Management, its Analysis and Insights, leverage data to make strategic decisions, Enhanced Efficiency and Productivity, Personalized Customer Experience, Real-time Decision Making, Business/ Domain Intelligence and Reporting, Risk Management and Compliance, Innovation and New Opportunities, ensure that the

right data can get to the right users or systems at the right time. The fundamental role of data infrastructures is to provide a platform environment for applications and data that is resilient, flexible, scalable, agile, efficient and cost- effective as well.

The National Data Infrastructure Initiatives: These include Open Data Platforms at International, National and State, City – specific, Sector–specific Data Platforms, Open Data Infrastructure for Researchers, Data Sharing Policies, and Data Protection Acts.

Artificial Intelligence will play an increasingly important role in the further evolution of Data Infrastructure due to the Increasing Complexity of data infrastructure, Cloud-Native Applications and data, Edge Computing and IoT Data Management, Support for DataOps for Data Management and Collaboration, AI-integrated data Infrastructure, Secure & Nuanced mechanisms of Data Sharing, API Ecosystem, Data Privacy and Security, Quantum Data Infrastructure.

A data fabric is an architecture that unifies all of these disparate data sources and applications in a secure, and automated fashion without changing where or how that data is stored. It offers data access without data migration. This connected architecture makes it easier, faster and more secure for the organizations to deploy data-driven applications and automation and exposes more data-driven insights to business users in a self-service fashion. Instead of changing where or how data is stored, the data fabric is an overlay, connecting data with analytics and users wherever needed. **While the data itself remains distributed across several on-premises and cloud resources, the fabric allows it to appear unified to end users.**

P. Vigneswara Ilavarasan stressed on the need to open up the data for economic development. Enabling data-based start-ups or new ventures may contribute to 4% of India's GDP. The enablement process is in progress and faces some challenges. There is a quality symmetry in data provision. The policy directions on data protection are giving mixed signals to the investments. The need for aligning good governance image and data outputs may hamper the data explorations. Data-driven policymaking may gain a lot from the demonstration effect of the private sector.

The Open Government Data Portal was designed, developed and hosted by the National Informatics Centre (NIC), New Delhi. It has been developed using Open Source Stack, the project is one of the initiatives under Pillar 6 (Information for All) of the Digital India initiative. The base Open Government Data Platform India is a joint initiative of the Government of India and the US Government. The Open Government Data Platform India is also packaged as a product and made available in open source for implementation by the countries globally.

The objective of Open Government Data Platform India is to facilitate access to the Government shareable data and information in both human-readable and machine-readable forms in a proactive and periodically updatable manner, within the framework of various related policies, Acts and Rules of the Government of India, thereby promoting wider accessibility and application of government-owned data and unlocking the potential of data for the national development.

The Challenges include Quality asymmetry in data provision, Missing datasets; Outdate, Sub- national government power dynamics, Policy advisory – New - Digital Personal Data Protection Bill, 2022, Need for aligning with good governance image.

To meet these challenges there is a need to improve Awareness — Data/evidence-based policy making, Need capacity building, Clear regulatory environment & advisory bodies, National data management & governance structure, Incentive for data sharing, Encouraging open data-driven start-ups

Mr. Arun Jadhav while presenting on 'Enabling Data Access, Sharing and Reuse' mentioned that technology can improve farmers' lives by providing information and applying scientific expertise to solve problems in agriculture and the environment. The provision of timely access to scientific data will certainly help the farmers to grow more and lose less of what they produce, combating threats to agriculture and the environment from pests and diseases.

FAIR is a globally recognized framework that can help organizations make data-rich investments more sustainable and deliver greater ROI (Return on Investments). By not practising FAIR, organizations are spending more than required every year on data-rich investments. Wrangling poor-quality data 'can take up to 80%' of the total effort in investments, leaving only 20% for analysis. The EU estimates the true opportunity cost of not having FAIR data in research at •10.2 Bn each year. The Benefits of FAIR include Reduced Duplication, Compliance Assurance, Improved Findings, Focus on Value, Enhanced Infrastructure, Foster Collaboration, Boost Citations, Maximize Publications, Enhanced Reputation and Lead by Example.

The topic of the talk of *Eva Victoria* was Encouraging open research data practices within national and funder policies. The research data can include but is not limited to raw data, processed data, software, algorithms, protocols, methods, materials, etc. Transparent research data is a pre-requisite for reproducibility. Open science debate is now emphasizing the importance of access to all research components in the research lifecycle.

The stakeholders including funders, institutions, publishers and researchers must collaborate and align to build a positive research culture that rewards and integrates reproducibility practice. Effective research data management practices must be implemented at the start of the research lifecycle – with benefits for researchers and institutions

We should encourage Embedding policies and practices, using data and collaboration, to achieve national research goals by preparing National policies and their implementation, committing to fund and support national infrastructure/ technological solutions, encouraging ubiquitous use of data and collaborating amongst key stakeholders.

Prof. Shantanu Chaudhuri, Director, IIT Jodhpur, while chairing the Session highlighted the Trustworthiness of Big Data. He stressed on establishing a comprehensive data governance framework that includes data quality, data security, and data privacy policies. He suggested to regularly audit and update this framework to ensure that it aligns with the evolving data standards and regulations. This will help build trust in the big data ecosystem by demonstrating a commitment to responsible data handling.

Transparency and Accountability: Be transparent about the sources of data, data processing methods, and algorithms used in big data analytics. Implement mechanisms for accountability in the case of data breaches or misuse. Encourage organizations to publish reports or documentation explaining how they handle data, enabling the stakeholders to assess the trustworthiness of the data.

Prof. Avik Sarka spoke on the topic of Ethics in Artificial Intelligence – Perspective from India

Localized Ethical Guidelines — Develop AI ethics guidelines tailored to the specific cultural, social, and economic contexts of India. These guidelines should address the issues like bias, fairness, and accountability with consideration for India's diverse population and unique challenges.

Collaborative Oversight — Encourage collaborations between government, academia, industry, and civil society in shaping AI ethics policies. Establish independent bodies or councils responsible for

overseeing AI ethics, comprising experts from various fields to ensure a balanced perspective and accountability.

Dr Victoria Eva, Senior Vice-President, Global Policy & Industrial Relations, Elsevier, spoke on The importance of FAIR data: Key practices and the risks of not implementing

Training and Awareness — Promote awareness and provide training to researchers, data scientists, and data stewards about FAIR data principles. Ensure that data creators and users understand the importance of making data Findable, Accessible, Interoperable, and Reusable.

Incentives and Mandates — Encourage funding agencies, institutions, and journals to require adherence to FAIR data principles as a condition for research grants, publications, or collaborations. Implementing such incentives and mandates can significantly drive the adoption of FAIR data practices.

Ms Alka Mishra opined that India has travelled a long way in its Open Government Data journey over the last one decade or so. The Open Data Technology Division of the National Informatics Centre (NIC), New Delhi has been an active enabler of this journey through the development, implementation and management of the Open Government Data (OGD) Platform India (data.gov.in). They highlighted the challenges and opportunities experienced in the Indian government data ecosystem, its policies, licences and the software development strategies implemented by the NIC for improved proactive disclosure of shareable government data. The OGD Platform is an open-source, microservice-based, cloud-native open government data platform for India. It features a full suite of data cataloguing and publishing workflows powered by the Dublin Core metadata standard; a robust approach to encourage and enforce machine-to-machine data sourcing through APIs for all data-providing entities; and blogging, informatics and visualisation functionalities. The redevelopment of the Platform software using microservices architecture has allowed the NIC to offer the same to various government entities to serve their unique requirements by configuring and customising of the core software components and services. At the same time, the common architecture and codebase of the OGD Platform, when utilised by the multiple government entities for their own data portals across multiple Software-as-a-Service instances, enables harvesting of data and metadata from the independent instances to the central Platform and vice versa.

Dr Vishnu Chandra, Former Deputy Director General & HOG-NIC (Geospatial Technology Services) brought in the discussion on Digital Transformation: Geospatial Perspective

Interoperable Geospatial Standards — Advocate for the development and adoption of standardized geospatial data formats and interoperability protocols. This will enable seamless integration of geospatial data into various digital transformation initiatives, fostering more efficient decision-making processes.

Investment in Geospatial Infrastructure — Encourage public and private sector investments in geospatial infrastructure, such as satellite systems, geospatial data repositories, and advanced mapping technologies. A robust geospatial foundation is essential for leveraging location-based data in digital transformation efforts.

Mr Sitansu Mahapatra, et al., NIC, India, spoke on A FAIR (Findability, Accessibility, Interoperability and Reusability) and Open Approach to Data beyond Statistical and Geospatial Datasets

Extended FAIR Principles — Extend the FAIR principles to include various types of data beyond statistical and geospatial datasets, such as multimedia, textual, and sensor data. Adapt FAIR principles to fit the unique characteristics of different data types while maintaining their core principles.

Collaborative Data Ecosystem — Foster collaborations between data providers, data users, and data infrastructure developers to create an open and interconnected data ecosystem. This ecosystem should promote data sharing, reuse, and innovation while adhering to FAIR principles to ensure that data remains accessible and usable across domains and applications.

The Session on Data Policy, Practices, Ethics and Infrastructure continued on 13th September, 2023, also. The first presentation titled "Empowering Data Journalism, through an Open and FAIR Data Ecosystem: Strategies, Standards, and Synergies", was delivered by Dr Uma Shankar Pandey. He presented the journeys of narrative journalism, precision journalism, and data journalism. He highlighted the challenges of the digital world, such as the rise of fake news, and lack of trust in public institutions in the post-trust, post-truth and post-expert world today. He highlighted how and why precision journalism evolved from the era of narrative journalism. While precision journalism paves the way for data journalism, everyday data stories add value to the audiences. Accessing data in the right format at the right time is essential for facilitating the work of data journalists. He then highlighted some of the tools he and other media educators used for creating data stories.

The presentation titled 'Science and Technology Driven Urban Forestry Programmes: A Preliminary Assessment on Data Practices" was presented by Mr Shirish Kanetkar of PKC, Pune. He highlighted the roles of urban forestry in the ecosystem services, particularly in climate resilience. He then highlighted the status of urban forestry in India which includes the initiatives of *Nagar Vana Yojana* of MoEFC; C biodiversity parks of various city governments, tree plantation drives, and climate-smart city initiatives. The carbon-neutral campuses, and sapling management, plantation monitoring are some of the new initiatives worth mentioning. He highlighted many initiatives of the Pune Municipality Corporation of urban forestry and afforestation. He suggested having urban forestry data in the proper format for overcoming data challenges.

The next presentation titled 'Big Data from Citizen Astronomy" was presented by Dr Disha Sawant of PKC, Pune. He highlighted how the galaxies are formed, shaped and discovered. Several radio telescopes across the world capture data on galaxies and many professional and amateur astronomers and citizen scientists are engaged in the analysis of galactic data. She also highlighted the various facets of the One Million Galaxies Project being coordinated at PKC (Pune Knowledge Cluster), which is a global project having multi-national researchers' participation.

In the presentation titled 'Determinants of Collaboration for Strengthening the Research Partnerships for Innovation and Need for Integrated RDM", Mr Angad Munshi of IIFT, highlighted the importance of research data management in scientific research. He focused on food technology research in this presentation and emphasized on research collaboration values, which include keeping the collaborators and participants safe, and identifying and managing risks, besides other factors. He also highlighted the contribution of food science research literature from India, which is about 3.5% of the global output and India stands seventh in terms of research literature production. He suggested having a formal data governance structure in R & D projects to address research data management. He further suggested incentivising the researchers when they employ systematic research data management based on a data management plan.

The Session on *Empowering Data Journalism through an Open and FAIR Data Ecosystem: Strategies, Standards, and Synergies*

In an era marked by misinformation and declining public trust in journalism, the need for accurate, evidence-based reporting has never been more critical. Data journalism emerges as a cornerstone in

restoring this trust, serving as a bulwark against ambiguity and speculative narratives. However, the challenges are manifold—ranging from the intricate skills required to the complex architecture of data ecosystems. This paper highlighted the strategies, standards, and synergies that could empower data journalism through an Open and FAIR (Findable, Accessible, Interoperable, and Reusable) data ecosystem. The skill set for successful data journalism is multi-disciplinary, extending from numeracy and spreadsheet fluency to data visualization and statistical analysis. Often, in moderate-to-large newsrooms, data tasks are compartmentalized among various desks and departments. Nevertheless, a holistic familiarity among reporters, editors, designers, and coders remains indispensable for producing compelling stories grounded in data. One salient challenge is the prevalence of data in PDF formats, which inherently stymies analysis. Transitioning towards a FAIR data ecosystem will not only simplify data extraction but also promote the integrity and reusability of the data. The strategies like standardized metadata tagging, open data repositories, and collaborative platforms can foster a conducive environment for data journalism. Data visualization is highlighted as a pivotal skill, functioning as the linchpin between intricate data sets and public comprehension. We discuss emerging tools and frameworks that enable journalists to present data in an engaging, intuitive manner. The key takeaways include the urgency for evolving newsroom skills, the necessity for adopting FAIR data principles, and the transformative impact of data visualization. By harmonizing these elements, journalism can fortify its credibility and make strides towards a more informed and engaged public.

Summary of Key Takeaways

- Restoration of public trust in journalism hinges on the adoption of data-driven reporting.
- A diverse skill set is vital for effective data journalism, with team members benefiting from cross-disciplinary familiarity.
- The prevalence of PDF formats hinders efficient data analysis; transitioning to a FAIR data ecosystem addresses this issue.
- Data visualization serves as a critical interface between complex data and public understanding, necessitating its mastery in newsrooms.
- Adopting open and FAIR data ecosystems could be a game-changer in elevating the quality and credibility of journalism.

Dr Harshal Shah gave a new perspective as to how data can be used more efficiently and effectively. While talking about the importance of data ethics, he mentioned that all human activities leave data trail for being accessed and processed by the agencies. There are quite a few concerns, and issues that are being raised on the ethics and policies around data and data governance in general. Zooming out, if we start analysing the evolution of a data-driven lifestyle, we will see more hope than concerns.

He presented his analysis of the past, present and future of data-driven lifestyle and its impact on data policies, ethics and infrastructure. The data has taken centre stage in the past few decades. However, before Google, this was not as pervasive and hence, the overall new way of life needs some more learnings to be more streamlined and ethical. Today, there are concerns about centralization of data, a few enterprises owning most world data and resulting in misuse of personal data and ownership. This also presents an opportunity to build the right ethical structure and improve the existing infrastructure. More distributed, informed consent-oriented and democratic data exchanges and ecosystems are the future of this more global data-driven world and there are various initiatives currently in the works which will be able to achieve this shortly.

Ms Akansha Nagpal and Dr Arun Jadhav stressed that effective data management and sharing are crucial in the current era of increasing research data volumes and complexity. The FAIR data principles are a leading framework for achieving these goals. However, implementing these in practice presents challenges beyond technical aspects. It requires a people-first approach because individuals' perspectives and expectations about FAIR and shared data vary. Understanding of these factors is vital to realizing the full potential of FAIR data principles in agriculture and other domains. They are developing the FAIR process and specific tools to enable diverse stakeholders to implement FAIR principles in investments. CABI has already delivered the Data Sharing Toolkit, an educational resource for building stakeholder competencies and knowledge. Their study is a call to action for organizations and stakeholders to prioritize implementing FAIR principles, overcome data-sharing constraints, and collectively work towards achieving a thriving and data-driven agricultural sector.

Prof. Sangeeta Trehan intertwined with FAIR data aspirations are the issues of privacy and use of personal data. The author was motivated to look at the variety of responses and the ensuing debate on some of these issues found in the literature and media recently with the passing of the digital personal data protection bill by both Houses of the Indian Parliament

and also do some first-hand reflection on the philosophical aspects of the so-called "personal" data.

She surveyed the issues of privacy and intersecting interests in "personal" information through five broad themes emerging from analyzing the selected literature:

a) Concern for privacy/ its violation and use of personal information/data on digital platforms for wealth creation by third parties,

b) Competing demands of protecting personal data and having to process such personal data for matters connected with or incidental to lawful purposes

c) Responsible practices ("Fair data") in handling of personal information

d) Individual ownership rights for personal data, and

e) Regulation of third-party data collection by companies and legislation to protect personal data.

She concluded that a pragmatic approach would essentially consist of having widely applicable baseline privacy legislation. This would ensure that individual privacy interests are respected as personal information flows to desirable uses.

Dr Krishnamurthy mentioned that the exponential growth of digital information and a growing emphasis on open science has resulted in an urgent demand for competent individuals who can manage, curate, and safeguard research data. Although libraries have always been excellent at managing information, managing research data poses unique and distinctive issues. Libraries are seen as the progressively embracing sophisticated data management systems, metadata standards, and collaborative tools with an emphasis on interoperability and scalability. Hence, the major goal of this research is to identify the obstacles and opportunities faced by the research data repository. In the ever-evolving landscape of academia, research data repositories (RDRs) will certainly play a pivotal role in facilitating open science, collaboration, and knowledge dissemination. His study embarked on a comprehensive exploration of the challenges and opportunities inherent in the management of research data within the academic settings.

Dr Mohit Garg (IIT, Delhi) remarked that the emergence of an open data ecosystem has facilitated the world with freely accessible data for driving innovative solutions. It aimed to break down various data

barriers (e.g., data format and standards, data privacy and security, geographic boundaries, time, etc.) and empower individuals, organizations, and communities to take full advantage of data for socio-economic development, civic management, scientific advancements, resources optimization, etc. Since the revolution of open data started, many portals have been developed by both public and private organizations to enhance public transparency and facilitate enriched public engagement in policy and other analyses. Some popular open data portals are: data.gov, data.gov.in, European Data Portal, World Bank Open Data, and UN Data. These portals enable the public to download multiple formats (e.g., XML, CSV, XSL, JSON, etc.) of data from a single access point. But many applications, like mashups and expert systems, require data from multiple sources in real-time. This cannot be done by

extracting the data from each source and merging it later. This can be achieved with the help of an API (Application Programming Interface) mechanism. It establishes the linkages between the different systems for sharing data in a synchronized way. The APIs play a significant role in the digitally connected world by enabling communication and interaction between software applications, services, and systems. It empowers the development team to build a robust and dynamic application by leveraging other systems' functionality, data, and services. He mentioned various APIs on the open data platforms, the workflow of integrating open data API into the existing or new systems, skills required to build such systems, challenges, and future aspects of API-based data systems and presented a demo of the application developed using API of Open Government Data Platform India and R on Baadal (An academic cloud provided by IIT Delhi).

Dr Ranjeet Kumar viewed that research data, as the outcome of colossal funding, hard work, infrastructure, and resource support which makes it a valuable asset. Creating a DMP (Data Management Plan) enables better management, reuse, and exchange of research data and a DMP tool will facilitate the same. The data management starts with a plan and involves essential measures to govern the complete data life-cycle. The DMP is a document that lays out all the necessary measures for better management of data, covering all the aspects of the data life-cycle. Hence, a DMP tool that satisfies the needs of all the research stakeholders of a particular country is very much needed. Some of the recent advancements in DMP research include:

- maDMP: Machine Actionable Data Management Plans by Research Data Alliance (RDA) working and interests groups on DMPs
- DMP-HUB: DMP Repository
- DMP-ID: Persistent Unique Identifier for DMPs
- DMPFrame: A conceptual metadata framework for DMPs

Dr Kumar recommended that since data management is a complex process, it requires a lot of infrastructure support to store, archive, and make it reusable; therefore, a DMP tool will reduce the cost and efforts of the data managers by streamlining and systematizing processes. The development of a DMP tool can give technical assistance for better RDM in countries like India.

Concluding Remarks

The Conference theme, "Open and FAIR Data Ecosystem," underscores the vital importance of creating a data landscape that is accessible, transparent, and equitable. It is a testament to the fact that data has become the lifeblood of the modern world, permeating every aspect of our lives, from healthcare and education to business and governance. Therefore, it is our collective responsibility to ensure that data is open, accessible, and usable for the betterment of society.

The Conference delved deep into the principles, policies, and platforms that underpin the open and FAIR data ecosystem. The experts, researchers, policymakers, and practitioners shared their insights, experiences, and best practices. It witnessed the power of collaboration and seen real-world examples of how open and FAIR data can drive scientific discovery, spur innovation, and address some of the most pressing global challenges. A few key takeaways from the Conference:

1. **FAIR Data Principles** — The Conference reaffirmed the importance of making data Findable, Accessible, Interoperable, and Reusable (FAIR). These principles serve as a foundation for ensuring that data can be effectively shared and utilized across disciplines and borders.
2. **Data Policies** — Effective data policies are essential to promote openness and transparency while safeguarding privacy and security. We must work collaboratively to develop and implement policies that strike the right balance between these competing interests.
3. **Data Platforms** — The tools and platforms we use to manage, store, and share data play a pivotal role in the Open and FAIR Data ecosystem. Investing in robust data infrastructure is key to supporting data-intensive research and innovation.
4. **Interdisciplinary Collaboration** — Interdisciplinary collaboration is at the heart of Open and FAIR data. Bringing together experts from diverse fields can lead to new insights and solutions that were previously unattainable.
5. **Ethical Considerations** — We must never lose sight of the ethical dimensions of data sharing. Protecting individuals' privacy, ensuring data security, and addressing biases and discriminations are ethical imperatives that must guide our actions.

The Way Forward

Let's carry the torch of Open and FAIR Data principles in our respective domains. Whether you are a researcher, policymaker, data steward, or technologist, you have a vital role to play in shaping the future of data. Let us continue to work together to create an ecosystem where data is a force for good, benefiting society as a whole.

Let us leave here today with a renewed commitment to the principles of openness, fairness, and collaboration. Together, we can build a brighter future where data empowers us to address the most pressing challenges facing humanity. Let the spirit of Open and FAIR data guide our path forward.

Principles, Policies, and Platforms — This conference has been an inspiring and enlightening journey, one that has brought together the finest minds from around the world to deliberate on the crucial topic of data openness and fairness.

Over the past 3 days, we have engaged in meaningful discussions, exchanged ground-breaking ideas, and forged valuable connections with experts, practitioners, policymakers, and advocates in the field of data management. The diverse perspectives and experiences shared during this conference have undoubtedly enriched our understanding of the challenges and opportunities that lie ahead in our pursuit of an open and FAIR data ecosystem.

As we approach the conclusion of this remarkable event, it is essential to reflect upon the key principles, policies, and platforms that have been at the heart of our discussions. Openness and fairness in data management are not just buzzwords; they are guiding principles that have the potential to transform the way we collect, curate, and share data. They are the cornerstones upon which we can build a more transparent, equitable, and collaborative data ecosystem.

Throughout the conference, we have delved deep into the fundamental principles that underpin the open and FAIR data movement. We have explored the concept of Findable, Accessible, Interoperable, and Reusable (FAIR) data and how it can enhance the discoverability and usability of data. We have discussed the importance of data sharing, data privacy, and data governance in achieving these goals. And we have heard inspiring stories of successful initiatives that have embraced openness and fairness in their data practices.

Furthermore, our discussions have touched upon the policies and regulations that shape the landscape of data management. We have explored the role of governments, institutions, and international bodies in crafting policies that promote data openness while safeguarding privacy and security. We have considered the ethical dimensions of data handling and the imperative of data stewardship.

Additionally, we have examined the platforms and technologies that enable us to implement open and FAIR data practices. From data repositories to data standards, from data analytics tools to collaborative platforms, we have witnessed the incredible array of tools and solutions that empower us in our quest for better data management.

As we look back on the knowledge shared and the insights gained during this conference, let us not forget that our journey is far from over. The challenges ahead are significant, but so are the opportunities. The decisions we make, the policies we craft, and the technologies we develop will have a profound impact on the future of data management and, indeed, on the world itself.

As we leave this conference, let us carry forward the knowledge and the connections we have gained here. Let us continue to work together, across borders and disciplines, to build a data ecosystem that is truly open and FAIR. Let us be the champions of change and progress in the realm of data, for the benefit of society, science, and humanity.

Prof (Dr) Ravinder Kumar Chadha
Ex Additional Secretary
Parliament of India, New Delhi, India

Inaugural

CODATA and WorldFAIR: Making Data work for Global Grand Challenges

Simon Hodson

CODATA Executive Director, 5 rue Auguste Vacquerie, 75016 Paris, France

Email: simon@codata.org

First published in 2016, in an article that has now received over 6000 citations, the FAIR (Findability, Accessibility, Interoperability, Reusability) Principles have had a significant influence on policy, practice and thinking about research data management and stewardship.

The catchy mnemonic has been effective, but the article and related work convey an important message. The fundamental purpose of the FAIR principles is to provide guidelines such that data and metadata relevant to all kinds of research outputs are machine readable and machine actionable. The vision is one in which research outputs can be visited online and at vast scale and the potential of machine-assisted analysis can be realized, but with data and metadata that are sufficiently reliable so as to reduce error and quantify uncertainty.

Another catchphrase is that FAIR also means 'Fully AI Ready'.

The FAIR principles do not insist that data or metadata be fully open and make the case that the principles apply just as well to cases where data can only be accessed in controlled circumstances. Yet, the interest in FAIR is related to another major development and movement in the early 21st century research: that of Open Science. Recognising, of course, that in some circumstances access restrictions are necessary, the Open Science movement aims to maximise the public benefit of research (particularly publicly funded research) by insisting that scientific outputs should be as open as possible, and that access restrictions should be 'proportionate and justified'. The nuances of these arguments are frequently lost, and sometimes obfuscated by vested interests. Nevertheless, the UNESCO Recommendation on Open Science, supported by the International Science Council (ISC), represents another major step in the transformation of science policy.

Why does this matter? Why should we care about these developments in high level policy and in data practice? The major global scientific and human challenges of the 21st century (including climate mitigation and adaptation, sustainable development, disaster risk research

and reduction, smart cities and energy systems, precision medicine or agriculture) can only be addressed through cross-domain research that seeks to understand complex systems through machine-assisted analysis at scale. Our capacity for such analysis is currently constrained by the limitations in our ability to access and combine heterogenous data within and across domains. The FAIR principles and the frameworks set by Open Science provide a significant part of the solution. Attention needs to be paid to the interfaces where data is used between disciplines, but the engagement with key research disciplines around data (re-)usability is essential.

To help address these issues, CODATA has been entrusted by the International Science Council to develop a program of activity: 'Making Data Work for Cross-Domain Grand Challenges'. After some exploratory work, the flagship activity is the WorldFAIR project which focuses on the implementation of the FAIR principles both within and across 11 different domain and cross-domain case studies, with a central effort to understand and guide cross-domain FAIR. It is the first broad-based effort to understand the issues around cross-domain and cross-infrastructure FAIR implementation through a case study driven methodology. Ultimately, it hopes to provide guidance for FAIR implementation both within specific domains and infrastructures and across them.

The I and the R of FAIR pose particular challenges but are significantly important in addressing the complex issues where datasets need to be combined and in enhancing scientific rigour and reproducibility. Consequently, increasing attention is being paid to semantics, the maintenance of referenceable vocabularies, terminologies, and ontologies and to metadata profiles—and to tools that facilitate the tracking of provenance and process, or that use variable level metadata and semantics to facilitate data integration. The outputs will include recommendations for FAIR implementation for each case study, as well as a proposed Cross-Domain Interoperability Framework (CDIF). The CDIF identifies a set of functional requirements for interoperability, particularly for steps in data combination, and recommends good practices for each of these requirements, in relation to the use of existing or emerging standards and specifications. The CDIF is categorically not a new standard, but is intended to act as a lingua franca across domain data practices and encourage the incorporation of a number of standards that perform important and specific functions across domains. We are keen to test this approach with colleagues from as many disciplines and application areas as possible.

This presentation explored these developments in detail, made a case for the importance of the FAIR principles and invited Indian institutions and research communities to participate in the WorldFAIR initiative.

Health

India's healthcare sector is highly dependent on data-driven decision making. While it has great scope for further growth and expansion, it is currently hampered by inadequate policies. It would help to take appropriate measures urgently, particularly in areas of health data management and exchange, with a focus on patient data privacy and security, for future sustainability.

The Power of Data for Driving Healthcare Innovations

Renu Swarup

Former Secretary to Government of India, Department of Biotechnology, Ministry of Science & Technology, New Delhi, India
Email: renuswarup01@gmail.com

This presentation focused on the Power of Data, as reflected in the statements and strategies of G20 leadership, particularly for driving healthcare innovations. The G7 Leaders Statement 2021 asserts that Data can play a transformative role in supporting effective early warning and rapid response to health crises. On the other hand, the G20 Leaders Statement 2023 recognises the role of digital transformation, AI, data advances, and the need to address digital divides. They endorse the G20 Principles on Harnessing Data for Development (D4D) and welcome the decision to launch the Data for Development Capacity Building Initiative, and other existing initiatives. The responsibility of Data Sharing for the public funded research comes from the research funders, data producers, sources, and users. It highlighted the Biotech-PRIDE (Promotion of Research and Innovation through Data Exchange) Guidelines, prepared by the Department of Biotechnology, Government of India. These Guidelines provide a well-defined framework and guiding principle to facilitate and enable sharing and exchange of biological knowledge, information and data, and is specifically applicable to high-throughput, high-volume data generated by the research groups across the country. These aim to help the universities and research laboratories to manage and maintain data; and to make accessible all shareable data for other researchers, students and startups to take up newer research activities.

The paper highlighted the functions and activities of Indian Biological Data Centre (IBDC), housed in the Regional Centre for Biotechnology, Faridabad, in the realization of biological data sharing in India. The presentation featured the India-HBGDki Data, which integrates health information developed through clinical trials, observational studies, surveys and the like data to develop a knowledge compendium. The database facilitates 24 harmonized datasets for an Indian data repository at the Global Health Analytics Program. The presentation also featured the Indian SARS-CoV-2 Genomics Consortium (INSACOG), which is an inter-ministerial consortium of 28 Laboratories. The ecosystem is created to ascertain status of new SARS-Cov-2 variants; and to establish a hospital network for clinical correlation, sequencing of infection breakthroughs, and sewage surveillance. The paper highlighted the Agenda for Action, recommending to build robust data banks with data compatibility, to harmonise data sharing policies globally with specified quality standards for data operability, to focus on capacity building of human resources and infrastructure, and to identify immediate priorities for applications.

Human Health and Genomics: Data Imperative, Opportunities and Challenges

Partha Majumder

National Science Chair
SERB, Government of India, Emeritus Professor, Indian Statistical Institute
Email: parmaj2023@gmail.com

The Recent Global Experience Underscored the Essentiality of FAIRification of Health Data

Data and data-access are critical for the society during times of crises (e.g., recent SARS-CoV-2 pandemic). On the other hand, there was a wide unavailability of data on a timely basis. One available data source was GISAID (https://gisaid.org/). There were widespread data-sharing hesitancy and/or restrictions to deposit data in the public databases. For the available data, there were considerable incompleteness and missingness; for example, a lot of metadata of relevance were missing. A lot of restrictions were imposed on data-access. The available data from multiple sources were not harmonized and hence there remained a lot of incompatibility.

The paper mentions that during the post-pandemic period, possibly a heterogeneous set of factors affected many infected individuals to suffer from the long COVID infection. However, it was difficult to conduct research to identify the factors because of lack of harmonized data. Further, there is inadequate interoperability of data and data systems.

A FAIRification Framework

The paper stresses on the need of a framework for data FAIRification (https://w3id.org/faircookbook/FCB079). Since data types can be extremely heterogenous, implementation of the various steps can be rather difficult. The specific use cases — personalized health, quality of care, public health — would require specific contextual information.

A Three-pillar Strategy

To enable interoperable usage of health data for research, the paper outlined a national, semantic-driven, three-pillar strategy as suggested by Gaudet-Blavignac et al. [JMIR Med Inform 9(6): e27591. doi: 10.2196/27591]. The paper considered this strategy to be widely applicable and adoptable in various national contexts, and outlined its different pillars

The first pillar defines the semantics. For example, allergy, as a concept, is composed of (i) 'reaction type'" and (ii) 'substance category'[See Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT, https://www.snomed.org/)].

The second pillar introduces a standard formalism to represent and store the semantics and the data. Some formalisms have been developed. For example, widely adopted in several bioscience projects as a standard framework for data aggregation and exchange to facilitate knowledge building and data

exploration, is the Resource Description Framework (RDF, https://www.w3.org/RDF/), which is an open and flexible standard.

The third pillar supports the integration and translation of this formalism into the existing application-targeted models.

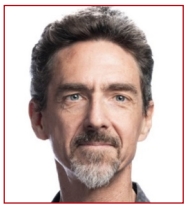
The paper mentioned that most healthcare records in India and in many other countries are handwritten and non-uniform. And therefore, to introduce a uniform system for the maintenance of electronic medical/health records (EMR/EHR) by the hospitals and healthcare workers, the Ministry of Health & Family Welfare (MoHFW), Government of India, released the EHR Standards for India in December 2016. The ultimate goal is to enable a unique EHR for every citizen of India. The EHR Standards for India document states "… with adoption of these standards properly, the data capture, storage, view, presentation, and transmission will be standardized to levels that will achieve interoperability of both meaning and data contained in the records".

With the mission to strengthen the disease surveillance in the country by establishing a decentralized State-based surveillance system for the epidemic-prone diseases to detect the early warning signals, so that timely and effective public health actions can be initiated in response to health challenges in the country at the District, State and National levels, an integrated health information platform has also been launched by the Ministry of Health and Family Welfare, Government of India, New Delhi.

The paper informed that a few years ago, the first national repository for life science data in India, called the 'Indian Biological Data Centre (IBDC)' has been launched. The IBDC is mandated to archive all life science data generated from the public-funded research in India. The data centre is supported by the Government of India (GoI) through the Department of Biotechnology (DBT). Fundamentally, IBDC is committed to the spirit of data sharing as per FAIR principles.

The IBDC would have primarily two data access types: (i) Open access/time-release access: Data submitted at IBDC would be freely accessible across the globe as per international open-access standards. The submitter, however, may choose to restrict the data access for a defined period of time. (ii) Restricted access: The data would not be made accessible freely. It can only be accessed through prior permission through the IBDC from the original data submitter. These access modalities are consistent with the principles enunciated in the National Data-Sharing and Accessibility Policy, Government of India, 2012.

Global Biodata Resources: Challenges to Long-term Sustainability of a Crucial Data Infrastructure



Chuck Cook

Program Manager, Global Biodata Coalition

Email: ccook@globalbiodata.org; www.globalbiodata.org

The presentation outlined the functioning of the Global Biodata Coalition (GBC) and highlighted the issues around the sustainability of biodata resources. It presented the global inventory and Global Core Biodata Resource projects and mentioned in particular contributions from within India to the global infrastructure.

Life scientists have for decades shared research data through open-access data resources that now form a globally distributed infrastructure of thousands of interlinked resources, providing access to deposited and curated data across many data types and life science domains. These biodata resources are critical to biological, life science, and biomedical research across academia and industry; they have substantial impacts both in fundamental and applied research, and they underpin and amplify the research investments made by the funders. Researchers, clinicians, and private sector scientists depend on the existence of biodata resources and the health of the infrastructure as a whole.

Despite their importance, individual biodata resources — and hence the infrastructure as a whole — show limited sustainability; unlike other scientific infrastructures, there is no overall coordination and new biodata resources emerge in organic fashion as the scientific community responds to its data needs. Resulting in uncoordinated and often misaligned funding—typically through short-term and competitive grants—the ability of resources to plan strategically, to secure appropriate technical and scientific expertise, to procure infrastructure and to develop fully according to the needs of the user community are limited. Furthermore, there is risk for many data resources important to the scientific community that funding will be reduced or become unavailable.

The Global Biodata Coalition (GBC) brings together organisations from around the world that fund life science research. Recognising the fragility of biodata resources, the GBC members commit to work together to better coordinate support for the life science data infrastructure and to address the challenges of sustainability. The current work between GBC members includes the development of principles that will help to structure future models for biodata resource sustainability.

Amongst GBC's current priorities is the profiling of the biodata resource infrastructure: Which/where are the data resources? Where lies the greatest fragility? What are the greatest risks and where should GBC prioritise? Two lines of work are helping GBC to address these questions. First, GBC has compiled an inventory, including where possible funding and further information, of some 3,112 unique and globally distributed biodata resources supported by many hundreds of funders. Second, GBC has identified 37 Global Core Biodata Resources (GCBRs) through an independent peer-review process. GCBRs, like keystone species in an ecosystem, are crucial for the functioning of the entire ecosystem, and their identification allows GBC, and the funders who support biodata resources, to understand which resources are most crucial for the infrastructure and by inference whose funding should be most carefully monitored. The GBC has established a forum for the leaders of the GCBRs that will share expertise and work with funders to develop and implement mechanisms for sustainability.

Health and Environment: Collaborative Approaches to Data Generation and Sharing

Priya Nagaraj

Chief Executive Officer, Pune Knowledge Cluster (PKC), Pune
Email: priya.nagaraj@pkc.org.in

This presentation focused on the program and activities of the Pune Knowledge Cluster (PKC). The PKC is one of the six city clusters that are a part of the Principal Scientific Adviser (PSA) Office's Science and Technology Clusters' Initiative. The PKC is focusing on five interlinked areas: health, sustainability and environment, sustainable mobility, big data analytics, artificial intelligence, and capacity building, including that in STEM education. The PKC works with over 70 partner organizations to build and implement S&T projects in the areas of Health, Sustainability and Environment, Sustainable Mobility, Big Data & Artificial Intelligence (AI), and Capacity Building. The Health initiatives of PKC focus on building collaborations to facilitate generation of data critical for health decisions, such as sero-surveys, clinical, immunological, and environmental public surveillance and creating organized platforms with comprehensive health information for the Pune region (India) which includes access to real time data.

Over the past 2 years, the PKC has enabled the creation of a disease surveillance ecosystem for infectious diseases in the Pune region (India). The goal is to set up an early warning system for disease outbreaks. The ecosystem enables a collaborative approach towards building a disease surveillance machinery involving local hospitals, R&D Labs, and civic bodies. The PKC has been creating several capacity building programmes utilizing their infrastructure and technical expertise. The PKC is rolling out an integrated data management system integrating disease, environment, weather, and mobility aspects; an interface with civic bodies; and carries out an impactful policy advocacy programme.

The PKC has initiated a COVID19 Surveillance Consortium, involving several partner hospitals, research collaborators, lab partners, sequencing centres, and research sponsors. The PKC has also initiated the Disease Surveillance for COVID19, Wastewater Surveillance (for detection of viral strains in wastewater), and Clinical Surveillance (analysing population level data, genomic sequencing, and clinical patient data). The PKC facilitated SARS-CoV-2 genome sequencing of clinical samples that accounts for 2% of all the positive samples from Pune. The PKC then shared data with the government on a weekly, and biweekly basis. Through the Wastewater Surveillance, the PKC created the Pune Wastewater Surveillance Dashboard for COVID19 and facilitated the identification of the first SARS-CoV-2 Omicron variant of concern sample in the Pune city. Consequently, PKC has worked with hospitals to collect, collate, and digitize clinical data of COVID19 hospitalized patients. The PKC was also involved in creation of the Epidemiological and Clinical Database for COVID19 partnering with M/S Strand Life Sciences. They are also involved in the creation of the COVID19 retrospective clinical database for Pune, utilising their access to 10% of hospitalised patient data in Pune including access to handwritten patient files.

The PKC is an active member of the One Health Mission initiated by the Principal Scientific Advisor Office, Government of India, New Delhi. One Health aims at integrating multiple streams of data from numerous governmental and non-governmental sources to ensure achieving overall pandemic preparedness and integrated disease control against priority diseases of both human and animal sectors.

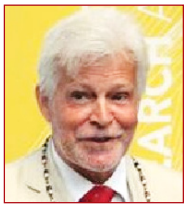
Role of Open and FAIR Data Ecosystem in AI/ML-driven Innovations in Bioscience and Biotechnology

Debasisa Mohanty

National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi - 110067
Email: deb@nii.ac.in

This presentation provided a general overview on the artificial intelligence and machine learning-driven innovations in biological sciences and biotechnology. It provided an overview of data-driven research in biology, bioinformatics-driven translational research, and bioinformatics software. The presentation discussed whether data-driven research has led to new discoveries, particularly after the discovery of DNA double helix structure. Here, the fundamental principle was when data from many experiments are pooled together and analyzed, inherent pattern becomes apparent. When a focus on genomics was given around the world, many projects of big data in biology evolved that led to many discoveries in the biosciences, including the personalized medicines, biomarkers, drug repurposing, computational approaches for protein structure prediction, AI-based methods for protein structure prediction, besides few other groundbreaking discoveries. The presentation mentioned that the The Nobel Prize 2013 in Chemistry showcased how the exascale computing would facilitate computational microscope to zoom from "atoms to cells", covering different scales of length and time. This presentation also highlighted the use of Digital Twins in biosciences. A digital twin is a virtual representation of an object or system that spans its lifecycle, is updated from the real-time data, and uses simulation, machine learning and reasoning to help decision-making. The digital twins of the living hearts, human livers, and human immune systems have been evolved in recent times facilitating the clinical biologists in their experiments.

Data Sharing for Public Health Emergencies - The GloPID-R Funders Perspective



Hans-Eckhardt Hagen

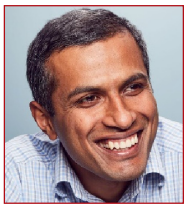
**GloPID-R*
Charité University Medicine, 10117 Berlin,Germany
Email: hans-eckhardt.hagen@charite.de

Alice Norton

Head of GloPID-R Research & Policy Team
Centre for Tropical Medicine and Global Health
Oxford University, Peter Medawar Building for Pathogen Research
South Parks Road, Oxford, OX1 3SY, UK

This presentation highlighted the functioning of GloPID-R (the Global Research Collaboration for Infections Disease Preparedness) is a global alliance of funding organization supporting research on new or re-emerging infectious diseases to increase preparedness and facilitate rapid research response to outbreaks with pandemic potential, funded by the European Commission. The paper mentioned that currently GloPID-R has 35 full members and 7 observers, including organisations such as the WHO, CEPI, and EDCTP. Since January 2023, the main Secretariat is based at Charité University Medicine in Berlin. Currently, the alliance is undergoing major changes, moving from a convening to an enabling power, offering practical support to its members such as coordinated funding call to address identified and agreed research gaps. The paper mentioned that one of the major projects is called Pandemic PACT, led by the GloPID-R Research & Policy Team, University of Oxford in the UK. This four-year initiative is designed to collate and analyse global funding tracking data and evidence across multiple disciplines on an ongoing basis. A funding tracking database and interactive tool would track research funding against research domains and research agendas. It further mentioned that a living Rapid Research Needs Appraisals (RRNAs) platform would quickly and rigorously appraise the existing clinical evidence to identify the key knowledge gaps across pre-defined clinical domains.

Data Integration in One Health: Challenges and Opportunities



Sindura Ganapathi

Office of the Principal Scientific Adviser to the Government or India, New Delhi-110011

4063, Prestige South Ridge, Outer Ring Road, Hosakerehalli, BSK 3rd Stage
Bengalure, 560 085
Email: sindurabg@gmail.com

This presentation highlighted that data needs to be liquid enough to be useful to serve multiple purposes, all of which cannot be foreseen up front. But, the sensitive data needs to be stored in a way it doesn't leak. This opens important conundrums to balance these opposing issues. It further mentioned that India is launching 'One Health Mission' to address the integrated disease control and preparedness for the pandemics by bringing together diverse sectors such as human, animal, and environment. To be able to realize this goal, information linkage across these sectors is critical. This need can be most exemplified by diseases such as Avian Influenza or Nipah. In the case of Avian Influenza, the virus can move across migratory birds, domestic poultry and humans, posing the threat of becoming the next pandemic. In the case of Nipah it is fruit bats, pigs, and humans. This means information needs to be accessible to decision-makers at the local, state, and national levels across the sectors for an effective control, not to mention to researchers so that they can work on the right solutions, be it a diagnostic or a vaccine.

The presentation was of the view that the need for data sharing is clear, but the mechanism that balances the need and overcomes the challenges is less clear. One Health Mission is working on a 'federated' approach of data sharing, borrowing from other uccessful examples where it has worked. The idea is that each sector or a subsector has specific needs for collecting, storing, and managing data in a certain way that serves the needs of the sector. This has been iterated over time to serve that sector. Therefore, trying to change this fundamental aspect by insisting on putting all the data in one central place is not practical, nor warranted.

According to the presentation, the second aspect is that the need for sharing the data is based on use cases. And it is often at the margins for any specific purpose. There is no apparent need to link cardiovascular or cancer outcome data from the hospitals with the wildlife sector as it doesn't lead to improved decision-making (even if one could make a case, it would likely be at the margins). However, sharing data on anthrax outbreaks between human and wildlife sector is critical for an effective control. Considering this, the federated approach ensures that data can live in its respective 'data islands' serving the needs of that sector in the intended way and not impinging on the ownership. However, if the data is machine readable, and annotated properly, it can be linked across the data islands for specific use cases. This means however that mechanisms for such linkage are developed, be it technical solutions such as APIs, and more importantly policies that govern effective sharing of the data while also ensuring privacy and security also. In the last, the presentation pointed that there would be many challenges along the way in realizing this, but it is important to start with the right intent and iterate along the way so that we learn to address them better since challenges like future pandemics are not likely to go away and we need to be better prepared.

A Patient Level Prediction Network Study of Testing Practices Used in Integrated Disease Surveillance and Response Systems by Several Countries in the Africa Region during the Pandemic

Jay Greenfield

322 Watkins Hill Road, Walpole, NH 03608, USA

Email: nightcleaner@gmail.com

The paper presented the design for a network study that explores the actual surveillance undertaken using the Integrated Disease and Response System (IDSR) for COVID-19 during the pandemic era. The study design follows two cohorts: The suspected individuals were identified by the community-based health care workers from the symptoms. The individuals were selected because their positions put them at an increased risk of exposure like hospital workers. In the design, we captured a person-level CRF called the individual case-based reporting form, administered through the ministries of health. Using CRF (Chronic renal failure) synthetic data, we created an ETL (Extract, Transform and Load) program that loads CRF data into a common data model – the OHDSI (the Observational Health Data Sciences and Informatics), OMOP CDM (the Observational Medical Outcome Partnership Common Data Model). Next, using an OHDSI data analysis workbench called ATLAS, we specified and tested a predictive analysis to determine the factors that predicted IDSR (Integrated Disease Surveillance and Response) protocol adherence and IDSR protocol exceptions. The adherence and exceptions went to specimen data collection and testing. All this was in preparation for receiving ministry of health CRF data first from Malawi and in the future from other ministries. The paper mentioned that the future for this study is to use a new feature of ATLAS that would allow us to include disaggregated SDG-3 (Sustainable Development Goal-3) indicators as predictors.

Agri-tech

The agri-tech business in India is expected to grow significantly in the coming years as it is driven by rising food demand and a need to improve agricultural productivity. Data science and machine learning offer valuable support to the industry as there exist possibilities for extensive data analysis. The important challenges in the field can be better addressed by learning from successful models existing in global ecosystems.

Open Science Practices Relevant to Earth Sciences and Agri-Tech

Sridevi Jade

Head & Outstanding Scientist, CSIR-4PI
NAL Belur campus, Wind Tunnel Rd, Bengaluru, Karnataka 560037
Email: head@csir4pi.in

The Council of Scientific and Industrial Research (CSIR) with its 37 labs is a Pan-India scientific and industrial organization covering a wide spectrum of areas, ranging from oceanography, geophysics, chemicals, drugs, genomics, biotechnology and nanotechnology to mining, aeronautics, instrumentation, environmental engineering, information technology. For the past 80+ years, it has been generating significant data across these varied and wide disciplines. The CSIR-Fourth Paradigm Institute addresses the data science and super computing needs of CSIR with niche domain expertise in the Earth and Engineering Sciences. The Data Science research involves both domain and data expertise in data and science supported by the computing and data visualization platform. It addresses the interdisciplinary nature of the data science and challenges i.e the five V's of data, viz. Volume, Velocity, Value, Veracity and Variety; and the associated large-scale open scientific data. The paper discussed in details the exponential growth, performance, scalability, security and observability of the BIG data. The best practices for open and FAIR data specific to scientific domain were presented, and these included providing ample metadata about scientific data, adopting open standards and formats for scientific data sharing, and possible inclusion of codes (computer programs) and tools used for the generation of data and scientific results. The paper highlighted that due to the rapid increase in data generation and consumption rate in the domain of earth sciences, the open data framework of this domain would demand more sophistication in the future. Some of the evolving challenges are: Seamless and real-time handling of missing data, realizing a more pervasive framework for open data leading to open science, catering to the needs of real-time replication of massive open data and disaster recovery framework. The paper covered specific aspects associated with promoting and practising open data for Agri-Tech and included dealing with concerns on exposing traditional farming knowledge, which is predominantly considered as a proprietary property of farmers.

Agriculture Data Exchange Platform (AdeX)

L. Ramadevi

Room # 24, 3rd Floor, Telangana Secretariat, Hyderabad-500022, Telangana, India
Email: osd_itc@telangana.gov.in

The paper presents the Telangana's vision to become a leader in the domain of emerging technologies (ET). The ET Wing of Government of Telangana focuses on the emerging technologies, such as AI (Artificial Intelligence), Internet of Things (IoT), cloud computing, blockchain, big data, space tech, additive manufacturing, drones, and robotics. It adopts a two-pronged approach, namely (i) Ecosystem Development for developing actionable policy frameworks, and (ii) Government Adoption of developing Institutions of Excellence. In consultation with industry, academia, startups, and other stakeholders, the Actionable Policy Frameworks are formulated which lay down the State's strategic initiatives and focus areas. The State has developed several sectoral policy frameworks, namely Cyber Security Policy, e-Waste Management Policy, AI Framework, Open Data Policy, Drone Framework, SpaceTech Framework, IoT Policy, Blockchain Framework, and Robotics Framework. The objective of Telangana state Open Data Policy is to provide a framework for the proactive disclosure of Government data by the various public authorities. The OD Policy focusses on maximising the use for the benefit of community, cost savings in data collection, maximum integration of individual databases with adoption of common standards, reduction in information asymmetry, and better decision-making. The State Open Data Portal (Data.telangana.gov.in) launched in 2016, is active now. Several startups and non-profit organisations are actively using these datasets made available on the open data portal by the government. The presentation also highlighted the Agricultural Data Exchange (ADeX) platform of the state, which is India's first digital public infrastructure project in agriculture. The contribution of the state-focused institutions/centres of excellence, ecosystem enablers, digital public institutions, digital public goods, GovTech adoption, and AgriTech solutions for the benefit of society were also highlighted.

A Situation-Aware Conversational Resource-Recommendation Strategy to Promote Business Growth in Food Deserts

Amarnath Gupta

6076 Roselle Meadows Trail, San Diego, CA 92130, USA

Email: a1gupta@ucsd.edu

The domain of FAIR principles, developed originally for scientific research data, has expanded widely in scope. Any organization that generates information for the consumption of a specific audience should ensure at least the findability and accessibility of the information. However, as the audience for the information generalizes from scientists to non-technical groups, and as the nature of the disseminated information becomes more diverse, other factors, not included in the original FAIR principles, become significant. The paper presented a case where the information provider is the Digital Government and the audience constitutes of small businesspeople, and showed that a complementary set of requirements, dubbed UNIT (Usability, Navigability, Interpretability and Timeliness), must be developed and implemented. In this case study, it was shown that non-incorporation of the UNIT principles can lead to information gaps and these information gaps could hinder businesses (including farms) in achieving their potential growth, and in a larger scale could prohibit socioeconomic progress. The presentation briefly discussed a project called NOURISH, where the UNIT principles would applied to an information dissemination and recommendation system for the people living in food and nutrition-insecure regions of the US.

Data Policy, Practices, Ethics & Infrastructure

As the evolving digital technologies are transforming the way research is conducted globally, the landscape of data science has been changing correspondingly, adapting to the challenges. Today our major challenges reside in the analysis of vast quantities of heterogeneous data from multiple sources. In recent times, many countries, regions, and domains have developed research infrastructures to assist with data management, stewardship, and analysis. Importantly, Open Science policies and practices have further galvanized these developments by various key stakeholders by providing a congenial environment that nurtures systematic needful reforms.

These principles and supporting practices seek to accelerate the process and unlock the potential of data analysis at the scale of computation and technology. This session would thus focus on sub-themes of particular interest related to open science and open data, FAIR data, research data management, data ethics, data licensing, data stewardship, data science education and training, data research infrastructures, and data platforms.

Stop Data Sharing

Barend Mons

GO FAIR Foundation, Rijnsburgerweg 10, 2333 AA Leiden, The Netherlands
Email: barend@gofair.foundation
https://www.gofairfoundation.org

This presentation focuses on FAIR (Findable, Accessible, Interoperable, Reusable) Data, FAIR Digital Objects, distributed analytics, and how these helped out in times of the COVID-19 crisis. It highlights the FAIR guiding principles for scientific data management and stewardship, which enhance the ability of machines to automatically find and (re)use the data.

With enormous amounts of research data, computers do the hard work of processing, pattern identification and machine learning. Human researchers, and the science system as a whole, will benefit from machine-actionability, as less time will be spent data munging. When data is stewarded and processed properly, ambiguity and non-reproducibility will be less of a problem as well. In addition, many datasets and resources are now either too large or too privacy-sensitive, or both, to be effectively routed around the globe for multidisciplinary and data-intensive science projects. Therefore, distributed machine learning is a new paradigm that I refer to as 'data visiting' rather than the classical model of 'data sharing'.

Data, Development and Policy in India: Potential & Challenges

P. Vigneswara Ilavarasan

Email: evignesh@gmail.com, vignes@iitd.ac.in

The paper focused on the need for opening up the data for economic development. It mentioned that enabling data-based start-ups or new ventures may contribute to 4% of India's GDP. The enablement process is under progress and facing some challenges. There is quality symmetry in data provision. It outlined that policy directions on data protection are providing mixed signals to the investments. The need for aligning good governance image and data outputs may hamper the data explorations. The paper mentioned that the data driven policy making may gain a lot from the demonstration effect from the private sector.

Enabling Data Access, Sharing and Reuse

Arun Jadhav

CABI, NASC Complex, 2nd Floor, CG Block, Dev Prakash Shastri Marg, Pusa, New Delhi, 110012

Email: A.Jadhav@cabi.org

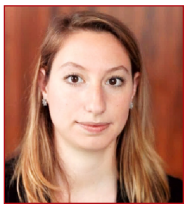
The paper highlighted the significance of a well-structured framework in ensuring the successful implementation of FAIR principles. The FAIR Findable, Accessible, Interoperable and Reusable) principles can help to unlock greater food security in South Asia through better access to information on soil health, agronomy, and fertilizers.

The paper outlined that one area where CABI has been working to apply the FAIR principles is the Data Sharing Toolkit funded by the Bill & Melinda Gates Foundation. The resource aims to equip development practitioners (Foundation programme officers, their grantees, partners, and other donors) with the necessary skills to develop better grants that will foster more access to agricultural data.

The paper mentioned that by not practicing FAIR principles, the organizations are spending more than required every year on data-rich investments. It cautioned that "wrangling poor-quality data" can take up to 80% of the total effort in investments – leaving only 20% for analysis. The paper informed that "The EU estimates the true opportunity cost of not having FAIR data in research to be at EUROS •10.2 Bn each year.

According to the paper, the benefits of FAIR data include reduced duplication, compliance assurance, improved findings, greater focus on value, and enhanced infrastructure. The FAIR principles also foster greater collaboration, boost citations, maximise publications, enhance reputation and show leadership by example.

Research Data Policy and Principles: Encouraging Open Research Data Practices within National and Funder Policies



Victoria Eva

Elsevier, 125 London Wall, London, EC2Y 5AS, United Kingdom

Email: victoria.eva@reedelsevier.com

This presentation provided an overview of the global policy landscape related to the open research data policies, before exploring definitions, best practice and publisher efforts to support open research data. The presentation addressed the suggestions for the future to support improved research data practices via collaborations between policymakers, industry, and the wider research community.

Trustworthiness of Big Data

Santanu Chaudhury

Director, Administration Building (West), Indian Institute of Technology Jodhpur
N.H. 62, Nagaur Road, Karwar, Jodhpur 342030,Rajasthan (India)

Trustworthy data refers to data that is accurate, reliable, and can be depended upon to be true and unbiased. Trustworthy data is essential for making informed decisions, conducting research, and developing policies. In today's digital age, where data is abundant and easily accessible, ensuring the trustworthiness of data is crucial to avoid misinformation or making decisions based on faulty information. This is why it is important to critically evaluate the sources of data, consider any potential biases, and verify the accuracy of the information before relying on it.

The measures of data trust include data quality, data health, , fairness of data, data sustainability, data diversity, and data privacy. Data Quality considers factors such as accuracy, completeness, consistency, timeliness, uniqueness, and validity or conformity of data. Data Health considers factors such as completeness, transparency, verifiability, and traceability of data. While we talk about responsible data science, the data scientists have the responsibility to make sure that insights from data are meaningful. The outcomes of data science processes may affect peoples' lives and the world fundamentally and permanently. responsibilities of the data scientists include fairness, diversity, sustainability, safety and privacy preservation of data subjects. It is important to also note the significance of digital foot-prints as data. All the digital transactions, authenticated or not, have their trace. The individualized data traces and digital health records are digital resource for the country. These still have ambiguity related to their ownership. Thus, the protection, security, and safety of authenticated personal or private data can make our digital footprints trustworthy. The paper pointed out that the ethical question that arises here is whether these individual data assets are tradable in the marketplace. The Trusted Applications using data for artificial intelligence should consider the factors such as fairness, robustness, explainability, and lineage of data. The Application Integrity of machine learning systems should consider the factors such as consistency, explainability, security, and reproducibility of results.

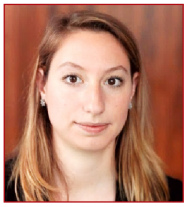
Ethics in Artificial Intelligence — Perspective from India

Avik Sarkar

Paik Para - BT Road Crossing
Kolkata 700002
Email: avik.tan@gmail.com

The presentation started with the statement that there has been an acceleration in the recent developments of artificial intelligence (AI) and its adoption across various business sectors. The AI applications primarily focus on model accuracy as a measure of performance. The paper pointed out that several ethical aspects are overlooked when we focus only on accuracy. The most prominent example is the model bias, which can occur based on protected variables that should not be used for model evaluation, like race, gender, etc. It was mentioned that many AI models are 'black box', making them challenging to explain to people or lacking explainability. The paper highlighted the need to focus on adopting responsible AI practices, which would help AI models to have improved outcomes. The paper focused on the various principles for responsible AI, as suggested by different forums like IEEE, OECD, ISO, UNESCO, GPAI, etc. It was highlighted that many countries advocate the need for responsible AI practices in their geography through principles declared in the public domain. The paper then mentioned on a few geographies like the European Union which has created legislation for regulating AI, followed by a discussion on the same. The presentation opined that it would be interesting to see the regulatory framework that would be developed for AI going forward in India and other nations.

The Importance of FAIR Data: Key Practices and Risks of Not Implementing



Victoria Eva

Elsevier, 125 London Wall, London, EC2Y 5AS, United Kingdom

Email: victoria.eva@reedelsevier.com

This presentation provided the publisher's perspective on FAIR data practices, including publisher's tools and initiatives to support making research data findable, accessible, interoperable and reuseable (FAIR). The presentation outlined the challenges and mentioned their implications to implementing FAIR data principles, before providing suggestions and ideas to mitigate these risks.

Digital Transformation: Geospatial Perspective

Vishnu Chandra

Advisor, Ministry of Panchayati Raj & Former Deputy Director General NIC
Mahagun Mezzaria, Flat No. 1320, Ferrera Tower, Sector-78, Noida -201301,
District Gautam Budh Nagar, Uttar Pradesh
Email: vishnu@nic.in

For more than five decades, the Information and Communication Technology (ICT) has played a major role and has impacted a wide range of activities dealing with data and information affecting citizens, planners, decision-makers, and governance professionals in both government and corporate sectors. The paper mentioned that the whole process of data and information management had started with digitization and computerization of manual systems to provide access for decision-making and subsequently, with advancement of technology, the process of digitalization started integrating workflows in the governance systems, followed by processes to generate value which further led to Digital Transformation for the economic development. The National Geo-Vision is continuously evolving taking "Geography or Geo-Spatial" as a core foundation of Digital Transformation, integrating end-to-end processes from Survey/Capture of Data, Data Standards & Organization, Integration with Governance Workflows and further delivery of services. This is generating value and adding location-specific dimensions in the context of Digital Public Infrastructure. The "Geo-Enabled Digital Transformation Ecosystem" encompasses a range of digital technologies such a Web, Cloud, Mobile, Geospatial, Remote Sensing, Drone, LIDAR, AI/ML, Big Data and so on to deliver value to the customers and hence, in turn creating digital platforms for the Environmental Social Governance (ESG) to facilitate Sustainable Economic Development of the country.

This paper covered the principles, policies and platforms in the context of "Geospatial" along with specific case studies and future plan of action. The recent policy reforms such as National Geospatial Policy-2022, Drone Policy-2021 and Space Policy-2023 have made a significant impact on Digital Transformation and in turn, on Digital Public Infrastructure. According to the presentation this has further enhanced Open and Fair Data Regime, as envisaged in the National Data Sharing and Accessibility Policy-2012 with respect to Geospatial Content.

A FAIR (Findability, Accessibility, Interoperability and Reusability) and Open Approach to Data beyond Statistical and Geospatial Datasets

Alka Mishra

Email: amishra@nic.in

The history of harnessing data for monitoring, planning and management of national development activities is not a recent one. What has changed since the beginning of this millennium, however, is "an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies," as described by a 2014 report by the UN. The paper discussed the need to look beyond, although not away from, the challenges and opportunities of statistical and geospatial datasets available from various government and non-government sources and to holistically address and utilize the possibilities enabled by the ongoing 'data revolution' i.e., business and social value creation possibilities brought about by transformations in generation, storage and analysis of data driven by the rise of the global digital economy. The paper appealed to the stakeholders of India's data ecosystem to deliberate and make recommendations on how to enable applications of the emerging technologies for harnessing of born-digital or digitized media (audio, image, object, text and video) information generated daily by or from connected devices (such as Internet-of-Things and sensors) and people across the country, as well as on the vast possibilities of open humanities/science/social science/technology research data to drive innovations within and beyond the academic sphere. In the concluding remarks, the paper shared some of the images and text datasets available on the Open Government Data (OGD) Platform India (data.gov.in) for application of emerging technologies and exploration of innovative use cases.

Empowering Data Journalism through an Open and FAIR Data Ecosystem: Strategies, Standards, and Synergies

Uma Shankar Pandey

Associate Professor and Head, Dept. of Journalism & Mass Communication
Surendranath College for Women, Kolkata
Email: mailusp@gmail.com

In an era marked by misinformation and declining public trust in journalism, the need for an accurate, evidence-based reporting has never been more critical. 'Data journalism' has emerged as a cornerstone in restoring this trust, serving as a bulwark against ambiguity and speculative narratives. However, the challenges are manifold—ranging from the intricate skills required to the complex architecture of data ecosystems. This paper delved into the strategies, standards, and synergies that could empower data journalism through an Open and FAIR (Findable, Accessible, Interoperable, and Reusable) data ecosystem. The paper mentioned that the skill set for the successful data journalism is multi-disciplinary, extending from numeracy and spreadsheet fluency to data visualization and statistical analysis. Often, in moderate to large newsrooms, data tasks are compartmentalized among various desks and departments. Nevertheless, a holistic familiarity among reporters, editors, designers, and coders remains indispensable for producing compelling stories grounded in data. The paper observed that one salient challenge is the prevalence of data in PDF formats, which inherently stymies analysis. Transitioning towards a FAIR data ecosystem would not only simplify data extraction but also promote the integrity and reusability of the data. According to the paper the strategies like standardized metadata tagging, open data repositories, and collaborative platforms would foster a conducive environment for data journalism. Data visualization is highlighted as a pivotal skill, functioning as the linchpin between intricate data sets and public comprehension. The paper discussed the emerging tools and frameworks that enable the journalists to present data in an engaging, intuitive manner. The key takeaways include the urgency for evolving newsroom skills, the necessity for adopting FAIR data principles, and the transformative impact of data visualization. By harmonizing these elements, journalism can fortify its credibility and make strides towards a more informed and engaged public. According to the paper, the Key Takeaways were: (i) Restoration of public trust in journalism hinges on the adoption of data-driven reporting, (ii) A diverse skill set is vital for effective data journalism, with team members benefitting from cross-disciplinary familiarity. (iii) The prevalence of PDF formats hinders efficient data analysis; transitioning to a FAIR data ecosystem addresses this issue. (iv) Data visualization serves as a critical interface between the complex data and public understanding, necessitating its mastery in newsrooms. (v) Adoption of open and FAIR data ecosystems could be a game-changer in elevating the quality and credibility of journalism.

Science & Technology Driven Urban Forestry Programs - A Preliminary Assessment on Data Practices

Shirish Kanitkar

Senior Research Associate (Urban Forestry), Project Manager, Trees Project, Pune Knowledge Cluster (PKC)
Email: shirish.kanitkar@pkc.org.in

Urban forestry has emerged as a significant academic field and practical focus area since the latter half of the 20th century. In India, it has gained national recognition only in the last 5-10 years. The FAIR principles on data offer an opportunity to view urban forestry in a more holistic manner, necessitating an understanding of the FAIR principles and their relevance to Urban Forestry. This understanding can be best achieved by examining some of the initiatives under Urban Forestry within the Pune Knowledge Cluster (PKC) and its network of partners.

Urban forestry involves the management and conservation of trees, forests, and green spaces within urban and peri-urban areas. It encompasses the planning, planting, maintenance, and sustainable management of trees and vegetation in cities and towns. The goal is to enhance environmental quality, human well-being, and urban aesthetics while minimizing the impact of resource exploitation.

Despite tree plantations, road-side landscaping, and the development of gardens being common in urban areas, the coinage "urban forestry" is still important. This term encourages us to view trees not in isolation but as part of a larger ecosystem in which we coexist. While tangible benefits of trees are not always easily measurable, they play a crucial role in mitigating the negative impacts of urbanization on the environment and improving the quality of life in cities.

In India, urban forestry is particularly important due to rapid urbanization, which has increased pressure on natural resources and the environment. Trees and green spaces in urban areas can help improve air quality, reduce urban heat, mitigate flood risk, improve biodiversity, provide recreation, aid in water conservation, and prevent soil erosion. These benefits have direct social and economic consequences.

However, urban forestry in India faces challenges. For example, data collection efforts lack transparency and currency, making them unsuitable for research and time-sensitive applications. Pune Municipal Corporation conducted a tree census mandated by a court order, but the data collected has limitations.

To address these challenges, PKC initiatives are analyzing the data collected and developing tools and models to improve plantation and conservation strategies. These initiatives aim to benefit civic authorities, urban planners, policy makers, research institutions, civil society organizations, and students alike by providing better access to data and analysis and through their involvement.

The vision is to evolve a data standard for urban forestry that is transparent, citizen-friendly, geographically accessible, real-time, scientifically credible and contributes to conservation decisions. Adhering to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) will help achieve these goals. However, challenges remain, including the need for capacity building and education to integrate subject-matter expertise with the data ecosystem.

Big Data from Citizen Astronomy

Disha Sawant

Pune Knowledge Cluster
3rd floor, Placement Cell, Savitribai Phule Pune University
Ganeshkhind Rd, Ganeshkhind, Pune, Maharashtra 411007
Email: disha.sawant@pkc.org.in

With emerging technologies, we have advanced instruments and telescopes observing the sky and continuously sharing data of varied astronomical entities. Enormous computational and human resources are being invested into analysing the data and draw scientific insights. In this process of annotating and filtering data, citizens can play a crucial role. In this way a country's large population can be leveraged for scientific discoveries along with initiating a dialogue between the scientific and non-scientific communities. The Pune Knowledge Cluster focuses on providing training to the curious and vigilant eyes and minds of citizens in order to eventually curate training datasets for building AI algorithms. These algorithms would help in classifying galaxies and generate even bigger catalogues. The presentation discussed the vision of Pune Knowledge Cluster vis-à-vis the Citizen Science Flagship program. It aims at strengthening the community aspect for science with the help of data analysis and providing technical backing to scientific and data domain experts to utilize the platform to open all sorts of data to the public.

Data Policy & Practices, Ethics, and Infrastructure

Harshal Shah

1112 Almeria Bend, Leander, TX 78641 US

Email: harshal.shah@ohsl.us

To highlight the economic aspect of data, the paper termed data as the new oil. It mentioned that are there quite a few concerns, issues that are being raised on the ethics and policies around data and data governance in general. Zooming out, if we start analyzing the evolution of data-driven lifestyle, we will see more hope than concerns.

The discussed, the past, present and future of data-driven lifestyle and analyzed its impact on data policies, ethics and infrastructure.

The paper mentioned that 'Data economy' existed even before the hype. In the Indian context, a specific community in northern India kept the ancestry data in something called as the Redbook. A well indexed, secured, and detailed database of a huge ancestry was passed along generations within this community. This was pre-digital era and hence, this was done manually but the major database principles were followed unknowingly. The issue of individual dependence, scalability, and vulnerability to fail existed due to the early nature of data economy.

The paper mentioned that in the present world, data has taken a central stage during the past few decades. However, before Google, it was not as pervasive and hence, the overall new way of life needs some more learnings to be more streamlined and ethical. Today, there are concerns about centralization of data, a few enterprises owning most world data and resulting in misuse of personal data and ownership. This also presents an opportunity to build the right ethical structure and improve the existing infrastructure.

The paper opined that a more distributed, informed consent oriented and democratic data exchanges and ecosystems are the future to this more global data-driven world and there are various initiatives currently in works which will be able to achieve this in the near future.

Data Ethics Policy in Higher Education

Suchith Anand

Professor of Practice in Science Policy at the University of Exeter, UK

Email: S.Anand@exeter.ac.uk

Data drives problem-solving and better decision-making. While the digital revolution has transformed the way we live, solving many issues; it has also given rise to numerous questions and challenges that affect everyone. This presentation highlighted the importance of educating researchers and aspiring data practitioners about the ethical considerations involved in collecting, using, re-using, and storing data during their training. In today's increasingly connected and digitized world, data is everywhere, and its significance cannot be overstated. With the increasing prevalence of big data and artificial intelligence, it has become clear that data ethics is essential for both individuals and organizations. This is especially true in the field of scientific education and research. Therefore, it is crucial for the universities to include data ethics in their curricula. The African Association of Universities (a University Association counting over 400 Universities across the African Continent), the University of Nottingham, and the Ethical Data Initiative launched the 'Campaign for Data Ethics in Education' in Summer 2023. The Campaign advocates for the integration of data ethics in all higher academic courses with a focus on data science and research. It aims to educate the next generation of data and research professionals about their legal and ethical obligations when it comes to using, re-using, and sharing of data.

Challenges and Opportunities of Open Government Data Platform — A Standard-enforcing SaaS (Softwares-as-a-Service) Solution for India's Data Ecosystem

Alka Mishra, Durga Prasad Misra*, Usha Saxena, Sanjay Mahendru, Ritu Ghai, Varun Gupta, Jaskaran Singh, Sitansu Mahapatra, Sumandro Chattapadhyay, Akhilesh Srivastava

**Email: dpmisra@gov.in*

India has travelled a long way in its Open Government Data journey over the past one decade or so. The Open Data Technology Division of the National Informatics Centre (NIC), under the aegis of Ministry of Electronics and Information Technology (MeitY), Government of India, has been an active enabler of this journey through the development, implementation and management of the Open Government Data (OGD) Platform India (data.gov.in). This paper highlighted the challenges and opportunities experienced in the Indian government data ecosystem, its policies, licences and the software development strategies implemented by the NIC team for improved proactive disclosure of shareable government data. The paper mentioned that the OGD Platform is an open-source, microservice-based, cloud-native open government data platform for India. It features a full suite of data cataloguing and publishing workflows powered by the Dublin Core metadata standard; a robust approach to encourage and enforce machine-to-machine data sourcing through APIs for all data providing entities; and blogging, informatics and visualization functionalities. The paper mentioned that the redevelopment of the Platform software using microservices architecture has allowed the NIC team to offer the same to various government entities to serve their unique requirements by configuring and customizing of the core software components and services. At the same time, the common architecture and codebase of the OGD Platform, when utilized by multiple government entities for their own data portals across multiple Software-as-a-Service instances, enables harvesting of data and metadata from the independent instances to the central Platform and vice versa.

INSPIRE Network: Building a Platform for FAIR Population Health Data

Jim Todd

Professor of Applied Biostatistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT
Senior Technical Advisory, National Institute for Medical Research, Mwanza Centre, PO Box 1462, Mwanza, Tanzania
Professor of Biostatistics, Catholic University of Health and Applied Sciences, Mwanza, Tanzania
Email: Jim.Todd@lshtm.ac.uk

This presentation focused on the INSPIRE Network (Implementation Network for Sharing Population Information with Research Entities), based on population health data in Africa. It is designed to be scalable and extensible as platforms for different diseases and health conditions. This is meant to harmonize, share and analyse FAIR data in an ethical way that acknowledges the work done by all partners. This network will include a data hub that will produce FAIR data that can be used by researchers and policymakers to answer important policy relevant questions. The presentation outlined the Health and Demographic Surveillance Systems in Africa (HDSS) that monitor vital events (births, deaths, and migration) and health outcomes. The HDSS enable large datasets covering up to 30 years of health in the rural and urban populations. The HDSS provide a set of platforms for clinical trials, disease-specific studies and public health interventions. The Network of HDSS shares good practices, new ICT tools, and training resources. The objective of the FAIRification is to enable mature data harmonisation models and processes for scalability and FAIR data sharing. The Platform can be used simply as a FAIRification service by an individual research centre. The research centres can use the platform to conduct and analyse a single study across geopolitical boundaries. The presentation mentioned that HDSS Platform can be used to keep each dataset in the national boundaries or sharing the person-level data in a secure data enclave or open data repository. Alternatively, research data centres can just share study aggregates. The Network ensures the standard platform available to other users (not just HDSS), while hospitals and research institutions can combine population and clinic data utilizing the infrastructure and resources. The INSPIRE is learning to develop platforms for different diseases, with high quality oversight from a Steering Committee, adapting vocabularies and concepts from OHDSI, linking with landscaping studies by disease specific researchers, and preparing study packages for analysis of data. The INSPIRE is working closely with CODATA and WorldFAIR to enable sharing of FAIR resources: metadata, data and data science products, networking with data producers, Ministries of Health (MoH), and data users.

In Tune with the Times — Pitching for a Pragmatic Approach to FAIR Data Handling and More

Sangeeta Trehan

Email: sangeeta.trehan@schoolofbusinessdesign.com

This presentation intertwined the FAIR (findable, accessible, interoperable, and reusable) data aspirations with the issues of privacy and use of personal data.

The presentation advocated the consideration of broad trends and approaches to protecting privacy of personal data and its usage and doing critical reflection on what would be a pragmatic approach in the context of a developing country.

According to the presentation, the thematic analysis of the selected literature revealed:

(i) Concern for privacy/ its violation and use of personal information/data on digital platforms for wealth creation by third parties, (ii) Competing demands of protecting personal data, (iii) Responsible practices ('Fair data') in handling of personal information, (iv) Individual ownership rights for personal data, and (v) Regulation of third-party data collection by companies and legislation to protect personal data.

The presentation further mentioned that the initiatives like Datacoup (in 2012), the app called CitizenMe and more recently the so-called 'zero party' data platform Caden are instances of development of a 'privacy marketplace'. Some lawmakers in the United States of America even endorsed and advocated institutionalization of privacy marketplace. Although the idea of such a marketplace looks appealing, treating personal data as a commodity to be licensed and sold by the 'owners' themselves does not seem to be a pragmatic solution. Instead, it is imperative to develop and implement a widely applicable baseline privacy legislation empowering individuals through transparency and individual rights to know, correct, and delete personal information in databases held by others.

People at the Forefront of FAIR Implementation: Enhancing Data Governance and Decision-making in Agriculture through a People-First Approach and FAIR Principles

Akanksha Nagpal

Project Coordinator South Asia
CABI, NASC Complex, 2nd Floor, CG Block, D P, Dharamdas Shastri Marg, Pusa, New Delhi, 110012
Email: a.nagpal@cabi.org

This presentation emphasized the importance of placing people at the forefront of the journey towards FAIR implementation.

It mentioned that agriculture is a 'high potential' sector for open data and the application of FAIR principles is now a 'moral imperative'. The presentation added that since 2018 CABI has built relationships, tools, methodologies, and skills, aligned with the Foundation's existing open policies, to help improve data sharing and governance. For instance, the CABI was awarded a $1.07 million contract from the Foundation to conduct a comprehensive study of national Soil Information Systems (SIS) in several countries, including India, Rwanda, Ethiopia, and Tanzania. This also captured learning from the USA, UK, countries within the European Union, Australia, New Zealand, Brazil, and Bangladesh. This was to support the Foundation's three to five-year goal on soil health and agronomy that would see integrated SISs established in several countries in Africa and the Indian states.

The paper also highlighted how CABI – using frameworks, tools, and use cases – helped the Australian Centre for International Agricultural Research (ACIAR) to think towards the change in data practices working across different levels of stakeholders. The ACIAR funded a Small Research Activity (SRA) to assess data management and FAIR data implementation across its research portfolio.

The SRA evaluated data management practices in the ACIAR-funded research projects and identified challenges and opportunities for improving these practices. The SRA found that while some projects were already implementing good data management practices, there was a need for a strategy to implement FAIR practices across the portfolio.

As a follow-up to the SRA, ACIAR awarded CABI a follow up contract to develop a FAIR data strategy for the Soil and Land Management (SLAM) program, a key research program under ACIAR. The goal of the contract was to ensure that the data produced by the program is effectively managed, shared, and reused, following the FAIR data principles.

Enhancing Data Governance and Decision Making in Agriculture through a People-first Approach and FAIR (Findable, Accessible, Interoperable, Re-useable) Principles

Akanksha Nagpal

CABI (Centre for Agriculture and Bioscience International), South Asia

NASC Complex, DPS Marg, New Delhi

Email: A.Nagpal@cabi.org

Arun Jadhav

CABI (Centre for Agriculture and Bioscience International), South Asia

NASC Complex, DPS Marg, New Delhi

Email: a.jadhav@cabi.org

According to this on-going research project presentation, an effective data management and sharing are crucial in the current era of increasing research data volumes and complexity. The FAIR data principles lead the framework for achieving these goals. However, implementing these in practice presents challenges beyond technical aspects. It requires a people-first approach because individuals' perspectives and expectations about FAIR and shared data vary. Understanding these factors is vital to realizing the full potential of FAIR data principles in agriculture and other domains.

The paper mentioned that the study is using the CABI's human-centered design approach, prioritizing user needs and problem-solving, alongside systems-thinking, that fosters stakeholders understanding and engagement. The iterative methodologies allow adaptation to dynamic stakeholders needs. This approach is being followed with major donors, supporting FAIR data principles in the Bill & Melinda Gates Foundation's AgDev investments, and ACIAR's Soil and Land Management portfolio.

The presentation argued that a FAIR-centric, FAIR-principle aligned, people-first approach would yield multiple positive outcomes. The involvement of stakeholders would build collaborations, trust, and engagement in data management practices, while enhancing data findability, accessibility, interoperability, and reusability. The study expects to observe improved data-sharing, better data quality, and greater potential for data-driven decision-making in the agricultural sector.

The people–centred FAIR data strategy and partner capacity building approach would increase returns on investment, utilize legacy data effectively and foster greater confidence in future investments and more valuable datasets long-term.

According to the presentation, the deliverables of the study would include a replicable FAIR process and specific tools to enable diverse stakeholders to implement FAIR principles in investments. It was mentioned that the CABI has already delivered Data Sharing Toolkit, an educational resource for building stakeholder competencies and knowledge.

Research Data Repositories in Academic Settings: Issues and Challenges

M Krishnamurthy

Documentation Research and Training Center (DRTC)
Indian Statistical Institute, Mysore Road, Bangalore 560059, India
Email: mkrishnamurthy1304@gmail.com; krish@isibang.ac.in

Vikas Kumar

Documentation Research and Training Center (DRTC)
Indian Statistical Institute, Mysore Road, Bangalore 560059, India
Email: krvikas.1998@gmail.com

This presentation outlined a diverse array of issues encountered by the research data repositories. The paper delved into topics such as the assurance of data quality, safeguarding data privacy and security, adhering to metadata and standards, ensuring sustainability and long-term preservation, addressing technical compatibility and maintenance concerns, and promoting interoperability.

Stressing on the need of training programs for librarians, the presentation mentioned that the academic libraries will see an improvement in their ability to manage research data efficiently if the librarians go thorough training programs specifically designed to address the issues and challenges they encounter. According to the study's findings, librarians' competence and confidence in data management significantly increased in libraries with strong training programs.

In the conclusion, the presentation mentioned that in the ever-evolving landscape of academia, research data repositories (RDRs) can play a pivotal role in facilitating open science, collaborations, and knowledge dissemination. It presented a comprehensive exploration of the challenges and opportunities inherent in the management of research data within the academic settings.

Deciphering Open Data Application Programming Interfaces: Fuelling Innovation and Collaboration for Building Data-driven Solutions

Mohit Garg

Assistant Librarian, Central Library, IIT Delhi Hauz Khas, New Delhi - 110016
Email: gargmohit@library.iitd.ac.in

The paper mentioned that the revolution of open data has led to the development, many portals by both public and private organizations to enhance transparency and facilitate enriched public engagement in policy and other analysis. Some popular open data portals are: data.gov, data.gov.in, European Data Portal, World Bank Open Data, and UN Data. These portals enable the public to download multiple formats (e.g., XML, CSV, XSL, JSON, etc.) of data from a single access point. But many applications, like mashups and expert systems, require data from multiple sources in real-time. And this cannot be done by extracting the data from each source and merging later. The paper argued that this could be achieved with the help of an API (Application Programming Interface) mechanism. It establishes the linkages between different systems for sharing data in a synchronized way. The APIs play a significant role in the digitally connected world by enabling communication and interaction between software applications, services, and systems. It empowers the development team to build a robust and dynamic application by leveraging other systems' functionality, data, and services.

The presentation mentioned that we use APIs in our day-to-day life and the best example of an API is a single sign-on system, like using one email id to create an account on the e-commerce sites like Amazon/Flipkart, social networking sites Facebook, Twitter, or any other website. In academia, we have been using persistent identifiers like ORCID for Sign-in/OAuth process, electronic thesis and dissertation submission, journal papers, data repository, etc. In all these cases, we unknowingly use API-based services at the front end, and the platform uses the desired API in the backend.

The presentation further pointed out that bridging mechanisms of APIs are valuable and essential for fostering open data initiatives or building data-based solutions. The majority of the open data platforms provide access to data through an API. The successful implementation of APIs require an understanding of the workflow of the integration between the system. The presentation discussed the availability of various APIs on the open data platforms, the workflow of integrating open data API into the existing or new systems, skills required to build such systems, challenges, and future aspects of API-based data systems. Further, a demo of the application developed using API of Open Government Data Platform India and R on Baadal (An academic cloud provided by IIT Delhi) was also presented.

Open Research Data Sharing among Indian Academics: A Cursory Overview

Pragya Chaube

DST-Centre for Policy Research, Indian Institute of Science

Participatory Research in Asia

Email: chaube@gmail.com

Moumita Koley

DST-Centre for Policy Research, Indian Institute of Science

International Science Council

Email: moumitakoley@iisc.ac.in

Open research data sharing is an essential component of open science which enables transparency, reproducibility, cooperation and advancement in science. However, the complexity of sharing scientific data is influenced by technical, cultural, and infrastructural aspects. This study aimed to gauge awareness, practises, motivations, concerns and barriers to adoption of open data sharing practices among Indian researchers.

The presentation revealed that faculty interviews were conducted at IISc Bengaluru across STEM fields and career stages to capture variations across these parameters. The interviewees were questioned on their awareness of open science, opinion, motivation or barriers to research data sharing, and their individual practices on data management. Attempt was made to find variations in open research data sharing practices among Indian and western counterparts.

The study has observed a noticeable cultural difference across different STEM fields. While practices have reached maturity in certain fields, like the biological fields, some other disciplines, such as chemistry, are still catching up. Moreover, even within specialised subdomains, there are differences. Additionally, variations also exist in terms of data formats, volume, and requirements for reuse. The presentation opined that advocating for a universal requirement that all researchers deposit their research data in public repositories might not be productive. It argued that discipline leaders could decide which data should be made available for reusability and reproducibility.

On the topic of repositories, the paper mentioned that at present researchers use global repositories like Zenodo, Figshare and Dryad. These repositories generally do not charge any fees for depositing data up to a certain limit. Also, Indian researchers generally deposit their research data only when mandated by the journals. But with growing awareness about data sharing a few global repositories won't be sufficient. Our data shows a clear need for better storage and data management facilities. There is also a significant need for capacity-building efforts in data management and record-keeping. The researchers also highlighted concerns about data governance, seeking more flexibility in embargo policies and greater control over their data. Lastly, the cultural aspect of data scooping is a significant reason for hesitation.

Data Management Plan (DMP): A Catalyst for Open and FAIR Data and an Accelerator of Research Data Management Processes

Ranjeet Kumar Singh

Senior Research Fellow at Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile Mysore Road, RVCE Post, Bangalore-560059, India & Department of Library and Information Science, University of Calcutta, Kolkata, West Bengal-700073, India
Email: ranjeet@drtc.isibang.ac.in

Devika P. Madalli

Director, INFLIBNET, Gujarat, Professor at Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile Mysore Road, RVCE Post, Bangalore-560059, India
Email: devika@drtc.isibang.ac.in

The presentation focused on the exploration of Data Management Plans (DMPs) as a catalyst for Open and FAIR data and an accelerator of Research Data Management (RDM) processes. The study aimed at finding answers for the following questions: (i) How DMPs support the overall implications of RDM and related processes? (ii) How DMPs contribute Open Data and FAIR data? and (iii) What are the recent achievements in DMP research?

Adopting a systematic literature review (SLR) process, the study found that RDM encompasses varied concepts like data creation, data preservation, data infrastructure, data sharing, data re-use, and data management skills, i.e. management of data at each step of the research data lifecycle. Although all these concepts are crucial, data sharing is at the heart of all RDM operations. The DMP by its philosophy, provides a strategic plan for all these activities and supports the overall implementation of RDM and related processes.

The presentation mentioned that DMP supports open science through management of research data for access, reuse, and long-term preservation. A significant impact of RDM is the establishment of reproducibility and replicability through facilitation of verifiable results. In the process of achieving reproducibility, social elements such as openness and desire to share and learn are also involved. The presentation added that DMP, being crucial for the RDM, contributes both directly and indirectly to the reproducibility and replicability processes, acting as a catalyst for open science and FAIR data.

The presentation listed some recent advancements in DMP research

- maDMP: Machine Actionable Data Management Plans by Research Data Alliance (RDA) working and interests groups on DMPs.
- DMP-HUB: DMP Repository

- DMP-ID: Persistent Unique Identifier for DMPs
- DMPFrame: A conceptual metadata framework for DMPs

The presentation concluded that data management being a complex process, requires considerable infrastructural support to store, archive, and make it reusable and a DMP tool can reduce the cost and efforts of data managers by streamlining and systematizing the processes. Hence, development of a DMP tool can provide technical assistance for a better research data management in countries like India.

Discovering Research Trends in Artificial Music Intelligence: Perspectives of Digital Humanities through Computational Text Mining



Suman Das **Anup Kumar Das*** **Manorama Tripathi**

**Email: anup_csp@jnu.ac.in; librarian@mail.jnu.ac.in*

The research on music and computer application, deep learning or machine learning related studies are paramount for understanding the humanistic problems faced by music producers, composers, listeners, as it useful for planning and producing automated machine produced music. The AI has transformed the way the music industry was functioning a decade ago. This study presented its investigations into AI, ML/DL and music over the past three decades. The aims of study included: (i) identification of knowledge structure of Artificial Music Intelligence literature by developing schematic diagrams of relationship between keywords from a macro perspective, (ii) investigating the main topic of artificial intelligence and music correlation during 1991-2022; and(iii) exploring the future direction of music and computational application for music producer-to-listener through digital humanities perspectives.

The presentation revealed the twin goals of artificial music intelligence subject discovery through quantitative content study using text network analysis, especially topic modelling. Using topic modelling approach, the study plans to analyze a corpus of research articles on AI and automatic music performance (Web of Science database), identify the main topics within the field and evaluate their evolution over time, co-occurrence, impact (based on citations), and prestige (based on journal ranking and Impact Factor). The presentation mentioned that the LDAvis and Structural topic modelling (STM) techniques have been found instrumental in comprehending the developed and emerging topics.

The presentation mentioned their study would help the music researchers and educators in comprehending the trends in music intelligence education. It would help educate future composers in creating impressive, human cognition music intelligence. At the end, the presentation did mention that the automated method cannot replace human importance, and therefore human-centered computing is crucial and the relationship between machine and human is complementary.

Research Data Repositories and their Metadata: A Comparative Study

Kavya Asok

Department of Library and Information Science
Central University of Haryana, Jant-Pali, 123031
Email: kavyadvg4@gmail.com

Snigdha Dandpat

Department of Library and Information Science
Central University of Haryana, Jant-Pali, 123031
Email: snigdhadandpat97@gmail.com

The presentation mentioned the necessity of common metadata elements to ensure interoperability among research data repositories (RDRs). The examination of various RDRs indexed in re3data revealed that the specific practices employed by these RDRs remained unknown. Therefore, the study centred on two key research questions: (i) How many metadata elements are available in total? and (ii) What are the common metadata elements and diverging features that research data repositories contain? The study adopted selected RDRs, including Zenodo, Figshare, Dryad, Open Science Framework and Harvard Dataverse, all of which were briefly introduced in the poster presentation.

Metadata elements of RDRs were compared in the study. Figshare (18), Zenodo (15), Open Science Framework (13), and Dryad (12) were discovered to have claimed the second-highest number of metadata elements, after Harvard Dataverse (22).

The study also observed that several metadata elements were common across all the RDRs under investigation. These included titles, keywords, descriptions, author name, institution name, license information, DOI, and publication date.

A comparison of RDRs' diverging features was made based on four parameters, viz., sign-up method, persistent identifiers, share, and repository software.

The study concluded that the Harvard Dataverse stood out with the highest number of metadata elements. It also underscored the importance of standardizing terminology and creating more generalized metadata that could be broadly applied to meet the diverse requirements of different research data repository cases. According to the presentation, this standardization was crucial for enhancing interoperability and facilitating the exchange of research data across various platforms.

Acknowledgement

The authors gratefully acknowledge the supervision provided by Prof. Dinesh Kumar Gupta, Professor, DLIS, Central University of Haryana and Dr. Prashant Shrivastava, Librarian Grade-I, B.B. Dikshit Library, All India Institute of Medical Sciences, New Delhi.

Ocean & Atmospheric Science

Recent initiatives like Blue Economy have given a great boost to research and innovation on marine resources. Having a vast coastline (7500 km), India itself has several research institutes for ocean sciences and a significant population thriving on ocean-based industries like fisheries, shipping, etc. These establishments are heavily dependent on data for routine operations and produce significant data sets on their own. Managing the enormous data sets coming from global research and business, and using them for addressing challenging problems will require advanced applications of data science. This meeting will help in bringing experts together, to discuss and perhaps resolve some of these key impediments.

Aspects of Data and Information Related to Ocean and Atmospheric Sciences: Sources and Users



D. Sivananda Pai* M. Mohapatra D.R. Pattanaik*****

India Meteorological Department, Ministry of Earth Sciences, Government of India, Lodi Road, New Delhi 110003

*Email: *dspai2005@gmail.com; **m.mohapatra@imd.gov.in; ***directorgeneral.imd@imd.gov.in, dgmmet@gmail.com*

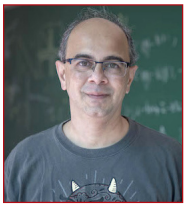
There are many types of data and information related to the ocean and atmospheric sciences and these are readily available for various applications. These include the observed data based on Surface and Environmental Observatories, Automatic Rain Gauges (ARGs), Automatic Weather Systems (AWSs), Upper Air Observatories, Doppler Weather Radars (DWRs), satellites, ships and buoys data, aircrafts, etc., various data analyses and reanalysis data, model simulations used for generating forecasts at various time scales, advisories and warnings issued for various weather and climate sensitive sectors, climatology, and climate change information, etc.

In this presentation, details and sources of various atmospheric and ocean data, provided by the India Meteorological Department (IMD) to its stakeholders, were provided. The climate data consisting of station level surface, autographic, automatic weather/rain gauge stations, upper air, radiation, marine and environmental data is provided through its online climate data service portal (https://dsp.imdpune.gov.in/). As per the recent data policy, 99.8% of the climate data provided by the IMD to its users is of free of cost. In addition, IMD provides the raw and image data from its network of 37 Doppler Weather Radars (DWRs) across the country, imager and sounding data from two meteorological satellites (3D and 3DR), and numerical model hindcasts and forecasts at different time scales on request from the user community. Further, IMD provides several derived products from the above data like drought indices, climate indices, heat and cold wave information, rainfall averaged over various time and spatial scales, climatology information, maps, meteograms, diecast model outputs, etc.

The users of IMD data include various operational agencies like Indian Air force, Central Water Commission, Watershed Organisation Trust, National and State Disaster Management Authorities, Water Resource Departments, Centre for Snow and Avalanche Study, etc. These data are provided through specific purpose websites, ftp servers, Application Programming Interfaces (APIs), etc.

Associated with the global warming scenario, improved model forecasts, easy accessibility to and cost-free availability of data, potential decision-making opportunities in various climate sensitive sectors, there has been a significant increase in the number of stakeholders using these large-scale oceanic and atmospheric data and its quantity.

Dynamical Models and Data for Complex Systems: An Earth Science



Amit Apte

Chair, Department of Data Science, Indian Institute of Science Education and Research IISER Pune, Dr. Homi Bhabha Road, Pune 411008, India
Email: apte@iiserpune.ac.in

There are two very distinct views on the scientific methodology. (i) The first of these views was expressed quite eloquently by Abraham Flexner more than 80 years ago in the book, 'The Usefulness of Useless Knowledge' and it points out the role of imagination and curiosity. In the current age of data, this may be supplemented by the humility to rein in the imagination and courage and perseverance domain. (ii) The second view has been variously called 'The Fourth Paradigm', etc. due to the unprecedented volumes and complex datasets available on it. This presentation described some of the extensive collections of earth science datasets across the world.

The presentation described the important distinctions between observational data and the data generated by the earth system models, which are essentially digital twins of the earth with varying levels of complexity being modelled, from clouds to sea-iee, to ecology, etc. This naturally leads to discussions on some important questions related to earth science data: (1) What is the source of these data? — Observations?, Model?, or A combination of the two? (2) What do these data represent? (Interpretability) – e.g. temperature measurement at IIC (easy to interpret); model prediction of temperature at IIC (average over a region) 3. Do more data give us more information about the system? These were illustrated in the context of weather prediction problems, whose solution depends crucially on the process of data assimilation.

The presentation described the three key concepts related to earth science data that are crucially associated with data assimilation: (1) Uncertainty —it is of course important for all the datasets, but its representation and quantification is particularly difficult in earth sciences. (2) Dynamics — the time-evolving systems lead to spatio-temporal variability that requires new descriptions. (3) Complexity — the system spans multiple spatial and temporal scales and consists of several sub-systems, many of which themselves are complex systems, such as cloud systems, sea-ice, bio-geo-chemistry, ecosystems, etc. The paper concluded by drawing parallels between the complex earth systems and the biological and health systems.

The paper listed some extensive data collections:

https://rda.ucar.edu/;
https://iri.columbia.edu/resources/data-library/;
https://cds.climate.copernicus.eu/;
https://www.ncei.noaa.gov/cdo-web/; https://www.climate.gov/;
https://www.mosdac.gov.in/; https://rds.ncmrwf.gov.in/;
https://www.earthdata.nasa.gov/

Role of Open and FAIR Data in Formulating Strategic Ocean Services

TVS Udaya Bhaskar

Scientist G, INCOIS, MoES, Government of India, New Delhi, India
Head Ocean Data Management (OD),INCOIS, MoES, Govt. of India
Pragathinagar (BO), Nizampet (SO), Hyderabad – 500090,Telangana

This paper started with the statement that adherence to FAIR principles with open, free, timely and unrestricted access to data would be beneficial for all in the field of Earth Sciences. It mentioned that the operational centres with missions to provide the Ocean Information and Advisory Services to the society, industry, government agencies and scientific community through sustained ocean observations and constant improvements through 'Systematic and Focused Research' are automatically adhering to the principles of open and FAIR data. This paper discussed the role of open and FAIR data in formulating strategic ocean services. The institutes like Indian National Centre for Ocean Information Services (INCOIS) of the Ministry of Earth Science (MoES), Government of India, are involved in providing operational services like Potential Fishing Zones (PFZs), Ocean State Forecasting (OSF), Hazards related services like Tsunami, Strom Surge, Multi Hazard Vulnerability, Harmful Algal Blooms, Oil Spills, Search and Rescue Aided Tools, etc to various stakeholders including fishermen, Indian Navy, Indian Coast Guards, Disaster Management authorities, etc. All these services thrive on the openness and FAIRness of data. These two principles make these services reach to a large number of people far and wide out across the world. Making these data open and FAIR enabled collaborations with various stakeholders like fishermen who provide valuable feedback on the service. Several NGOs help in making the reach of these services to more number of people than anticipated. Many startup companies are coming forward to build tools using this data. Several mobile-based apps have been built so that users can use them effectively. The researchers can make use of the open data to undertake studies on the impact of climate change in the services being rendered. The paper presented two case studies, viz. (i) *in situ* data from the International Argo Program, and (ii) Potential Fishing Zone advisories and mentioned that enormous amount of benefits were derived by these two programs by making data open and FAIR. The Argo Program has fostered a multinational collaboration resulting in cutting edge research and ~4000 publications. The PFZ program has benefitted fishermen brought NGOs, startup companies and researchers in the field. Overall, the Open and FAIR data builds an ecosystem that would be beneficial for the society at large.

Towards Open Access for Climate Justice

Gitanjali Yadav

Email: gy@nipgr.ac.in

The paper started with the statement that we all understand 'Science' to be a collective effort working for the common good, but a growing number of scientists, policymakers, and social scientists argue that science is often too isolated from the society to fulfil this promise. They advocate the need for Open Science to close the gap between science and society by democratizing scientific knowledge. To address this need, this paper described semanticClimate, a global movement for liberating knowledge from climate-related reports. According to the paper, it represents a community based on collaboration and communal values, an Open platform-independent client-side technology, semantic resources for TDM (Time Diversion Multiplexing), a dictionary system based on Wikidata & Multilinguality. The paper regarded this effort to be a mechanism for capacity building in terms of creation of youth data practitioners, by means of open notebook science in climate action. The paper informed that founded during the coronavirus pandemic period as an online remote internship program at the NIPGR in India, in collaboration with the University of Cambridge, UK, the #semanticClimate has now evolved into a student authored toolkit for extracting/liberating knowledge from the Literature. The #semanticClimate is a continually expanding citizen science movement, bringing together undergraduate, postgraduate, and PhD students to work across boundaries of rural/urban geographies, nationalities and language boundaries. The atmosphere is electric, and the team along with global volunteers has achieved significant progress with #semanticClimate. The barrier to entry is intentionally kept zero and anyone can join this Citizen Science Initiative by simply filling an interest form on the website "https://semanticclimate.github.io/p/en/"

Digital Humanities

Technological advancements are rapidly remodelling Digital Humanities, which includes disciplines like social sciences, and cultural & natural heritage. Digital methodologies and platforms foster interdisciplinary collaboration to resolve complex issues and/or develop innovative solutions while AI/ML-mediated proficient analysis provides better insights into social phenomena. In studies of cultural & natural heritage, data science has led to many positive changes to promote newer avenues of research. However, a wide chasm exists in the field arising from highly heterogeneous sources of data in terms of both formats and contents. This conference aims to provide insights to researchers and stakeholders on how to organise and resolve data using digital technology. It will help formulate data capture and storage methods that are compatible with multiple sources and to develop more effective methods for preservation, vigilant maintenance, and exhibition.

Artificial Intelligence in Healthcare

Belinda Seto*

U.S. Department of Health & Human Services
National Institutes of Health, Office of Data Science Strategy
Email: setob@mail.nih.gov

The presentation started with a statement that Big data has fueled the development of artificial intelligence (AI) and machine learning (ML) techniques. Such techniques were built on algorithms and models, many of which inform clinical decision-making ranging from diagnosis, treatment and prevention of diseases. The increasing volume and speed of producing genomic, proteomic, metabolomic and microbiome data added invaluable dimensions to fine-tuning AI models toward personalized and precision medicine.

The paper mentioned that there are numerous use cases of AI-based applications in precision oncology, with detailed characterization of the tumor microenvironment and intra-tumor heterogeneity. These characteristics may predict treatment responsiveness for cancer, prognosis or progression of the disease. The AI models have also been used for numerous other diseases. However, it is important to note that disease prevention and improvement in community health will have the most impact on the overall health of a nation.

The AI technology has led to significant changes in the pharmaceutical arena. This technology has quickened the pace in identifying the drug targets, the active site of target-ligand interactions, screening of small molecules and biologics, as well as the strategies for drug repurposing. The U.S. National Institute of Health maintains several open access databases in this arena, including PubChem (PubChem (https://pubchem.ncbi.nlm.nih.gov), Drug Bank, Protein Data Bank and UniProt (https://www.uniprot.org/). These are just a few examples of many that NIH is firmly committed to FAIR and open access to data.

The paper mentioned that in 2023, the NIH implemented a data sharing and management policy for all the NIH-funded research awards. In issuing this policy, the NIH has tried to encourage researchers to consider data sharing options including: 1) domain-specific repositories (134) that are typically limited to data of a certain type or related to a certain discipline and 2) generalist repositories (9) that accept data regardless of data type, format, content, or disciplinary focus. These nine generalist repositories feature open data submission, open timeframe for data deposit and NIH funding support.

To optimize data sharing and upholding the FAIR (findable, accessible, interoperable, and reusable) principles, the NIH continues to develop common data elements and promote their adoptions for translational and clinical studies/trials in order to enhance interoperability. In this regard, the NIH collaborates with grantee institutions to build their capability in using standardized Fast Healthcare Interoperability Resources (FHIR®) in exchanging clinical data.

The paper concluded with the remark that This 2023 conference on **Open and FAIR Data Ecosystem: Principles, Policies, and Platforms** is consistent with the NIH policies on open access and data sharing and management.

Data Science for Social Impact

Partha Pratim Das

Visiting Professor, Department of Computer Science and
Professor, Indian Institute of Technology (IIT), Kharagpur
Director, Center for Data Science and Analytics
Ashoka University, Sonipat, Haryana 131029
Email: partha.das@ashoka.edu.in

The aim of this presentation was to revisit the power of Data Science, to analyse the scope of using quality data in Social Impact projects, and to propose a framework for capacity building in Data Science for Social Change.

Data Science is the computational science of working on the data, with the data, and by the data using Artificial Intelligence, Machine Learning and Data Science per se. And this paper focused on the Social Impact. That is, the paper intended to understand the social issues from data, devise solutions based on data, innovate with the assistance of data, and measure the outcome of solutions using the data (of course, the use all the traditional tools and techniques continued as and when applicable).

The paper mentioned that the social sector has been using data in research, development, and extension for a long time. However, unlike the science or engineering or business sectors, the data in the social sector are usually sparse, incomplete, inaccurate, grossly unstructured, limited in availability, and low in quality. Hence, the social scientists often have to use them in a pilot or sample basis with limited confidence and poor reliability. The paper informed that Data sciences (and all the tools and techniques that fuel it) now offer the means to overcome many of these challenges posed by data in the social sector. Both quantity and quality have been improving, and most importantly techniques to use unstructured and disperse data are now in the coffers of the practitioner.

Since Data Science has evolved mostly from the community of computer scientists and mathematicians, the skills for its use have not percolated deeply into the social sector. Also, it needs strong interdisciplinarity in the overall approach where data scientists (who are masters of harnessing data for problem solving) and social sector scientists (who are masters of various domains of social impact like climate and health) can work in deep mutual collaboration using a language that is understandable for both. This leads to the question – should data scientists learn the social sector, or should the social scientists be trained to use data science? The former is formidably complex because the social sector is too varied, wide and deep. The latter is manageable because data science has been packaged in terms of tools that can be used by the social scientists with an effective short-duration training. An analogy would be – for computer-aided delineation of tumors, a good computer scientist cannot be taught oncology, rather a doctor can be trained in data-driven image inferencing tools.

The paper informed that at the Center of Data Science and Analytics at Ashoka University, such a capacity building program has launched for working professionals. It is titled Professional and Executive Development Program in Data Science for Social Impact, is driven by AshokaX, and currently focusses on interplay of Climate Change and Health. This 6-months' online part-time program can provide a blueprint for quality training in the sector. The program outline and its outcomes were discussed in the presentation.

Building Local Data for Social Impact Workforce: A Network of Networks Model

Priyank Hirani

Email: priyank@data.org

This presentation focused on the Data.org web-portal that exists to shape the DSI (Data for Social Impact) field. Data.org enables collaboration across several sectors and stakeholders to drive action and results. The presentation mentioned that since founding in 2020, Data.org has successfully launched data science initiatives in three focus areas. The portal enabled a Capacity Accelerator Network (CAN) and developed an India Data Capacity Accelerator, which is among the three existing hubs supporting a local network of partners focused on collaboratively building DSI workforce. The paper informed that within India, accelerators and local partners identified through an open request for proposal (RFP) process by a panel of locally grounded and globally informed experts. Multi-disciplinary, experiential programs built with a focus on climate, health and ethical considerations of data science (6-11 months, online/hybrid, undergraduate and postgraduate programs). Diverse learner cohorts created, with learners representing 40%-50% women, from over 20 Indian states, with varying work experience, academic training and socio-economic backgrounds, charting career trajectories for such talent through internships and fellowships curated in the impact space. The paper mentioned that as in India, Data.org has networks of partners in the US and Africa, with upcoming networks in Southeast Asia and Latin America. These partners connect regularly to share learnings and resources. Data.org seeks to further expand such a network of networks by providing avenues to share resources, curriculum, tools, knowledge, and facilitating collaborations amongst like-minded partners across continents with shared values. The CAN has enabled an engaged, global network of diverse DSI (Data for Social Impact) training providers, learners and organizations with an appreciation for open and FAIR data ecosystems.

Digital Humanities and Digital Scholarship Transformation: A South African Perspective

Lazarus Gallant Matizirofa

Email: lazarus.matizirofa@wits.ac.za

Daisy Selematsela

Email: daisy.selematsela@wits.ac.za

The convergence of digital humanities and digital scholarship transforms the digital information landscape and user experiences that transpose libraries to the 4IR (Fourth Industrial Revolution) and keep them focused on maintaining their relevance in the hybrid and dynamic scholarship environment. The University of Witwatersrand Libraries has made efforts towards providing new facilities and developing new services to meet humanities scholars' changing research behaviors by employing Digital Scholarship Centres, Makerspaces, Media Labs, and Digital Fabrication Laboratories (FabLabs). It provides digital experience to clients using a variety of digital tools and methods. The presentation further mentioned that effectively collaborating the Digital Humanities and STEM community is still a challenging task for the University Libraries without all the accessories, technologies, and human capacity to serve as a digital library. The presentation outlined a proposal regarding opportunities and challenges for collaboration with all the relevant stakeholders at WITS (World Integrated Trade Solution) to develop both DH and DS, by embedding digital spaces, archives, special collections, museums, and digitization as one multiple-service delivery of hi-tech library service. The WITS Digitization Centre is at the heart of a national digital preservation strategy that will train various stakeholders to curate, enable discovery, and expose Africa's rich, complex heritage which boasts extensive collections from historical archives, museums, and special collections, including rock art, fossils, artefacts, and ancient stone tools.

Unlocking the Potential of Digital Twins for a Sustainable and Better Future

Souvik Barat

Principal Scientist, Tata Consultancy Services Research, Pune, India
C-1101, Raheja Vista, Phase 2, Mohhamadwadi, Pune, Maharashtra, 411060
Email: souvik.barat@tcs.com

This presentation mentioned that today there is a growing need for technologies capable of supporting evidence-based informed decision-making by comprehensively addressing the concerns of involved stakeholders with conflicting and collaborative objectives in the face of uncertainty and dynamism. The key stakeholders in social systems, including city and state administrators, public policymakers in fields such as healthcare, water planning, agriculture, traffic control, etc., envision having access to technology to navigate an uncertain and evolving future and understand the consequences of their decisions before applying them in reality.

This presentation shed light on the transformative potential of digital twin technology as a powerful tool for enabling in-silico experimentation towards evidence-based informed decision-making in the face of uncertainty and dynamism. The paper delved into a diverse array of case studies, ranging from business scenarios showcasing how digital twins have helped conceptualize and tailored products and services to meet the ever-changing consumer preferences of a telecommunications company, to impactful use cases such as effectively controlling the spread of infections during the peak of the Covid-19 pandemic in an Indian city. The presentation explored the substantial potential of digital twin technology as an instrument for creating a brighter, more sustainable future for both enterprises and society as a whole. This included opportunities like controlling diseases, setting up sustainable food supply chains to reduce waste, managing floods, ensuring continuous water supply, improving healthcare in rural areas, and making crop production more efficient, among others.

Digital Twins for Control and Optimization

Aditya A Paranjape

Senior Scientist, TCS Research

The paper explained how model-based control and optimization techniques have matured over the past fifty years and have been deployed on real-world systems of ever-increasing complexity. In parallel, model-free control techniques have also evolved under the umbrellas of reinforcement learning (RL) and adaptive control. They have been a subject of growing interest, partly because of a shared lineage with the field of artificial intelligence, although their practical use has been somewhat limited. The paper further explained that the two approaches can be brought together gainfully for a broad range of practical problems wherein a nominal model is readily available in the form of a digital twin or a high-fidelity simulator. This is true even if the underlying system is prone to modelling and environmental uncertainties. The paper illustrated this paradigm using examples based on business problems from TCS Research. The examples demonstrated not only the utility of digital twins as aids for control design but also the challenges associated with developing one for control in a practical setting.

cover image source: UNESCO-Open science-pillars.