Making data work for cross-domain grand challenges

The major global scientific and human challenges of the 21st century (including climate change, sustainable development, disaster risk reduction) 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. Suboptimal data practices are a major and costly limiting factor on research: it is estimated that 80% of research expenditures are used to prepare inconsistent data for use.

Solutions to complex and difficult problems require data to be assessable and actionable by machines using big data in combination with the most advanced hardware and software technologies. Data must be richly described with metadata, well-documented, transparent and ultimately humanly comprehensible so as to facilitate extraction of meaning from complexity. The fundamental enabler of data-driven science is an ecosystem of resources that enable data to be FAIR (Findable, Accessible, Interoperable and Re-usable) for humans and machines. This ecosystem must include effective, maximally automated stewardship of data, and effective terminologies and metadata specifications.

Achieving this goal requires a major consensus-building effort, in particular to gain agreement about the core technologies and semantic solutions which will allow data to be combined for cross-domain research. **CODATA, with the support of and on behalf of ISC, proposes a major, global, decadal programme—'Making data work for cross-domain grand challenges'—to address these challenges.**

Over the last two years, a CODATA-led pilot project has developed, tested and refined methods for aligning metadata specifications, taxonomies and ontologies to address these problems in a consensual fashion. Applying this approach, the Decadal Programme will work with other international data organisations and their groups, with cross-domain programmes and research initiatives, with the organisations and communities that create metadata specifications and terminologies, and with other stakeholders to enable an ecosystem for FAIR data for cross domain research to be developed and implemented. This ecosystem will include a) maximally-automated stewardship of data; b) easily-applied terminologies and metadata specifications to assist the combination of data across domain boundaries; and, c) facilitation of the more effective application of data intensive, machine-learning and visualisation techniques for discovery and analysis.

The programme will take a three-pronged approach (though engagement and cross-fertilisation between these areas of activity will be essential):

Theme 1: Enabling Technologies and Good Practice for Data-Intensive Science: working with domain and technology experts, the programme will identify and define enabling technologies and good practices for data intensive scientific discovery that is applicable across disciplines.

Theme 2: Mobilising Domains and Breaking Down Silos: the programme will pro-actively engage international scientific unions and associations, and other stakeholders, in programmes of work designed to promote interoperability of data and related services.

Theme 3: Advancing Interoperability Through Cross-Domain Case Studies: the programme will work with a number of cross domain case studies (including, but not limited to the Sustainable Development Goals, disaster risk reduction and reporting, urban health and resilience and infectious diseases).

The impact of the programme will be to assist the scientific and innovation communities to accelerate scientific understanding through a step change in the application of interdisciplinary data-intensive methodologies and thereby enable more efficient and transparent science to address global challenges. The programme will achieve this by helping reduce the proportion of effort dedicated to data cleaning, manipulation and harmonisation. It will maximise the amount of machine-actionable FAIR data available for analysis and linking.

The programme will play a catalysing role, prioritising its activity to enable and coordinate the development and adoption of standard technical specifications and vocabularies and associated technology approaches, while working with partners to ensure that the necessary capacity for their effective use exists, and has a tangible and growing impact on the data available.

Under the oversight of CODATA (and the programme governance) a **coordinating programme office** and a **cohort of experts** will play a facilitating role to build consensus and convergence. The development and implementation of technical and semantic solutions, where needed, will be done by **programme nodes and partners**.

CODATA is now charged to put in place the core funding, capacity and partnerships in order to launch a Decadal Programme at the ISC General Assembly in Oman in October 2021.

As part of this process of partnership building, CODATA invites discussions with IIASA to explore the potential for collaboration on this initiative. Systems analysis and the cross-domain, data-related activities in which IIASA is a world-leader, will provide excellent case studies and would likely benefit from the approach outlined.