

# Data Integration Initiative: Planning Document

## 1 PREFACE

---

During 2017, CODATA initiated and led a discussion with data science groups and international scientific unions and associations about the timeliness of a major initiative on interdisciplinary data integration. Meetings at the ICSU HQ in Paris in June 2017 and at the Royal Society of London in November 2017 produced a report and communiqué<sup>1</sup> supporting a long-term initiative and outlining some of the essential issues to be addressed. The key priorities for this initiative are to address data integration in support of major global challenges and to develop relevant data capacities across all the disciplines of science.

An *ad hoc* steering group was created to plan how these should be carried forward, comprising: **CODATA**: Geoffrey Boulton – President; Simon Hodson – Executive Director. **ICSU**: Heide Hackmann – Executive Director. **Application Domain leaders**: Laura Merson - Infectious Disease Outbreaks; Virginia Murray – Disaster Risk Reduction; Stephen Passmore – Resilient Cities. **Data Scientists**: Simon Cox – CSIRO; Lesley Wyborn – ANU; Bob Hanisch – NIST; Phil Archer – Consultant.

Supporting the steering group in making contributions to the initiative are: Gisbert Glaser - ICSU; Katsia Paulavets - ICSU; Bill Michener – DataONE; Kevin Blanchard - PHE; John Broome – CODATA.

The formal governance of the initiative is yet to be determined.

This planning paper is an outcome of a meeting of the steering and supporting group on 19 January 2018. It is designed as a first scoping of the purpose, structure and roadmap for the initiative, and will shortly be made available to the community of practice represented by the attendees and invitees of the 2017 meetings. Its primary use is as a live document for planning purposes. It is not an early draft of a bid for support or funding, though it is likely to be a source text for such.

---

<sup>1</sup> URLs to be inserted for report and communicate

# Purpose, Structure, and Roadmap for a Major Data Integration Initiative

## 2 BACKGROUND

---

### 2.1 CHALLENGE AND PURPOSE

The digital revolution of the past two decades offers profound opportunities for science to discover hitherto unsuspected patterns and relationships in nature and society, on scales from the molecular to the cosmic, and all in areas of human concern, from cultural artefacts and local health systems to global sustainability.

There is a major, largely unrealised potential to merge and integrate the data from different **disciplines** of science in order to reveal deep patterns in the multi-faceted complexity that underlies most of the **domains of application** that are intrinsic to the major global challenges that confront humanity. The challenge is that varying and incompatible data standards have been used across the different disciplines, along with inadequate definition of the vocabularies needed to categorise them. The result is that integration of diverse data can generally only be achieved within and between closely allied fields.

Characterising, understanding, and dealing with the complexity inherent in major global challenges will be integral to the mission of the new International Science Council that will come into being in the first week of July 2018; the first meeting of whose Governing Board will take place in late September 2018.

### 2.2 ACTIONS

We plan to identify, promote and implement a programme of work that will substantially increase the capacity of the international scientific community to achieve rigorous, interdisciplinary integration of data to support work on major global challenges as a matter of routine. This will be a long-term, decadal initiative that has the potential to fundamentally enhance the capacity of science in the 21st century.

The communiqué of November 2017 expressed the agreement of meeting participants to work together with the broader research community to:

- develop and apply solutions for interdisciplinary data integration;
- pursue this through data integration for major global challenges that can also act as exemplars of its interdisciplinary potential;
- support, in parallel, the development of capacities to realise the potential of modern data resources across all the disciplines of science; and

- recognise that in many disciplines, foundational work is required to develop specific vocabularies ontologies and provenance tracking systems that are needed to enhance data discovery, use, interoperability and integration.

## 2.3 PRIORITIES

It was agreed to work to launch two major strands of work:

**Strand 1** will involve projects in the application domains of three global challenges: **infectious disease outbreaks** (1A), **disaster risk reduction** (1B), and **resilient cities** (1C). They have been chosen as major issues where relevant data exists and is accessible, where data integration is a tractable objective, and where there are existing communities of practice that are willing to collaborate. They are designed to have practical outputs of value to policymakers and users, to develop technical approaches and methods that have generic value, and to be persuasive demonstrators to the broader scientific community of the value of the approach. Discussions are currently taking place on the possibility of adding a further pilot project on **agriculture**.

**Strand 2** will seek to support those disciplines of science that have not yet developed the standards (vocabularies, ontologies, etc) that are necessary for effective data integration. These are necessary for those disciplines to efficiently utilise modern data resources, and if the broader potentials of disciplinary knowledge to address major interdisciplinary challenges are to be realised. Formalisation of the discipline-specific vocabularies is an essential pre-requisite for integration of data from different disciplines.

The work of Strand 2 underpins the approach taken in Strand 1 through a process that becomes progressively more straightforward, even trivial, through repetition in further projects, thereby building a more solid foundation for interdisciplinary, data-driven science.

# 3 STRAND 1 – INTERDISCIPLINARY DATA INTEGRATION FOR GLOBAL CHALLENGES

---

## 3.1 APPROACH

The projects in Strand 1 were chosen from a long list of 15 global challenges after careful prior analysis followed by discussion at the November 2017 community meeting that they satisfied three primary criteria that:

- important results could be obtained of value to stakeholder communities;
- communities of practice exist that would welcome and would in principle collaborate with a data integration project;
- much of the necessary data exists and is accessible.

The purpose of the projects in strand 1 is:

- to demonstrate the power of data integration to provide practical benefit to the application domains;
- to develop procedures and methods of data integration that are generic, with the potential to be applied in a wider range of application domains;
- to identify the need for further data that will enhance the potential of data integration to achieve major new insights in the chosen domains;

- to develop and demonstrate mechanisms to support the reporting by UN member states of the global targets and indicators identified in the 2015 UN Landmark Agreements including the Sendai Framework for Disaster Risk Reduction 2015-2030 and the Agenda 2030 Sustainable Development Goals.

These projects are expected to make a powerful business case for data collection, data infrastructure, data processing and integration.

It is recognised that these are major long-term tasks, and that a staged approach will be necessary. We envisage three stages:

**Stage 1** – pilot projects to produce useable and applicable results through the integration of available data in tractable ways that yield demonstrable benefit to policymakers and users, and to act as demonstrations of the value of the basic approach to scientists. They should be tightly circumscribed to ensure deliverable and useful results in a limited timeframe. They should develop methods and approaches that can be applied more widely.

**Stage 2** – more ambitious projects that build upon demonstrably productive approaches to data integration in the domain, that persuade stakeholders, funders, domain experts and the broader scientific community of the potential benefit of longer term efforts, and that encourage application of data integration approaches to further application domains.

**Stage 3** – the three chosen projects are also ones that have the potential for further analysis of the links that exist between their domains. There is conceptual coupling between disease outbreaks, urban resilience and disaster risk reduction which offer the potential for deeper understanding of the interactions between these domains, in ways that build on the outcomes of Stage 2.

## 3.2 STRUCTURE, COLLABORATION, AND COORDINATION

The success of Strand 1 will depend upon collaboration and coordination of efforts between a number of players as illustrated in figure 1. They are:

### ***Domain & data scientists***

The role of Domain Scientists is to identify the principle challenges of their respective domains. They will need to work together with Data Scientists to establish and to apply the methods by which integration can be achieved.

### ***Other collaborating experts and groups***

There are many groups that have made significant strides in data integration in cognate areas, whose tools and approaches will be of value. For example, close collaboration will be sought with the Research Data Alliance (RDA) where data interoperability has been a major priority. It is important that we avoid unnecessary duplication, ignore existing work of relevance and avoid telling others what to do. Our approach should be to build on existing work wherever possible and appropriate and to provide facilitating and helpful processes.

### ***Data and data-service providers***

Access to relevant data holdings is a necessary pre-requisite for success, and strong links and agreements with them will be necessary. Collaboration in this with bodies such as ICSU-WDS, UNISDR, IRDR, GEO, GBIF, OGC and CGI will be important. There is also considerable experience in bodies such as ELIXIR in providing data services to disciplinary communities, with experience and methods that will be of considerable value.

### ***Stakeholders***

It is important that each project should engage with relevant policymakers and users in an iterative conversation to ensure that the outputs of the projects are useable and of value to stakeholders. It must not be just a provider-driven process. Stakeholder enthusiasm and commitment will be vital to success. Such productive relationships would be enhanced if the call (September 2017) from the UN Sustainable Development Solutions Network for each country to have a 'Chief Data Officer' ("Counting on the World") were to be satisfied.

### ***Funders***

Early discussions with international and national funders will be necessary both in understanding and agreeing how to approach the staged elements described in 3.1, how to address their long-term perspectives, and how to provide immediate funding of the necessary planning and coordination involved in preparing substantive bids for funding and support.

### ***Sponsors***

Success of the ambitious, long-term aspirations of this initiative will strongly depend upon support and guidance from bodies concerned with the long-term development of global science. The International Science Council formed by the union of ICSU and ISSC, to be launched in early July 2018, is precisely such a body. A major target for a well-conceived proposal for support will be the first meeting of the Council's Governing Board in late-September 2018. Advice from the current Executive Directors of ICSU and ISSC about how best to approach the new Council should be sought as a matter of urgency. An early meeting with UNESCO will also be important given their considerable convening power. The three chosen projects (see 2.3) are also of great importance to the UN Sustainable Development Goals and the Sendai Framework for Disaster Risk Reduction. Consequently, approaches should be made to the UN STI Forum, possibly through presentations and discussions at the meeting to be held at UN HQ on 5-6 June 2018. Presentations should also be made at the UN World Data Forum Dubai, UAE, 22-24 October, 2018, to raise awareness. The deadline for session proposals is 31 January.

### ***Governance***

The nature of governance for the initiative is currently under discussion.

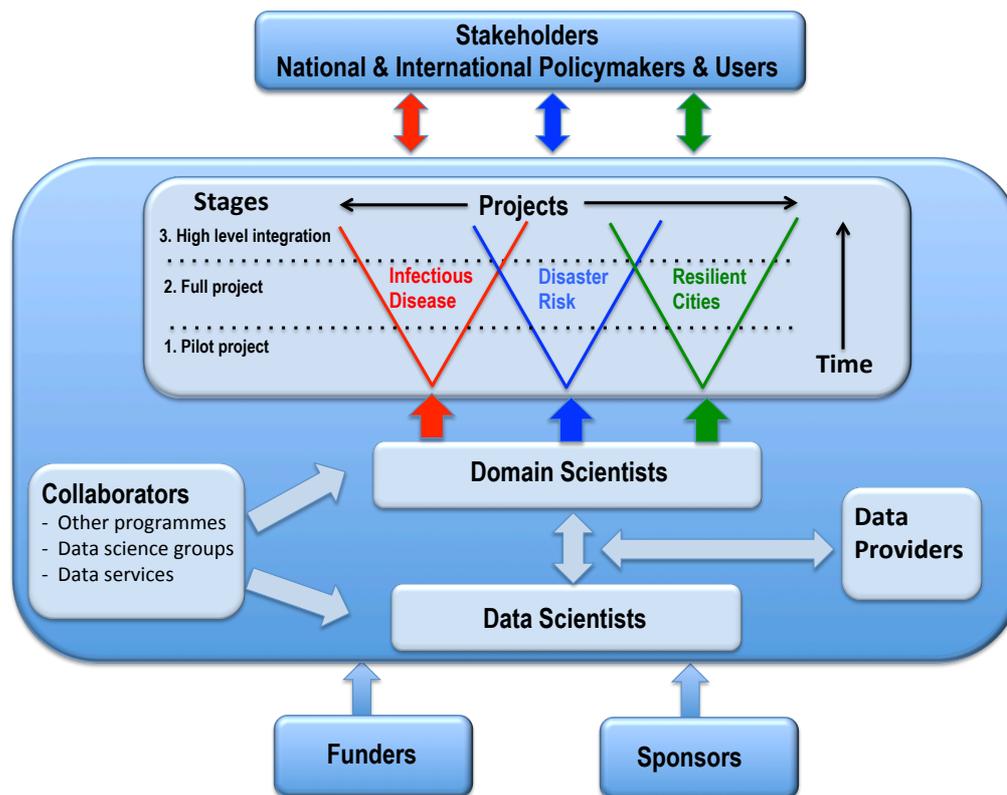


Figure 1. Structure of Strand 1 & relationships between key players

### 3.3 PILOT PROJECT 1A – INFECTIOUS DISEASE

Project Leader – Laura Merson (Centre for Tropical Medicine, Oxford University)

#### **Vision**

Misjudgements in the prediction, detection and response to infectious diseases outbreaks mean that epidemics and pandemics remain among humanity’s greatest threats. Data that characterise many of the factors influencing the progression of an outbreak are available, but remain isolated in siloes within the various domain- specific communities, often with their own domain-specific formats, vocabularies and ontologies. Integration of the varied data resources about disease vectors and transmission routes has great potential to identify underlying relationships that determine the trajectory of an outbreak. Each of these are vital to maximise the effectiveness of a response. Those working to prevent infectious disease outbreaks from reaching epidemic or pandemic levels are frequently frustrated that access to and analysis of currently disjoint data was not available at the time when such access could have saved tens of thousands of lives .

Regrets that data existed but was not available to the right people at the right time is often obvious with hindsight. Our vision is to use foresight to reduce suffering caused by emerging infections, improve global health security and contribute to global health equity.

## **Objectives**

- To develop tools that enable exploration and understanding of the cross-sections and correlations among biological, environmental, socio-economic, and behavioural factors that shape epidemics.
- To anticipate and mitigate the impact of emerging infection outbreaks by improving:
  - *Preparedness*: To apply new tools to augment the analysis of multi-disciplinary risk factors that define vulnerabilities to emerging infection outbreaks.
  - *Detection*: To enhance the scope, granularity and timeliness of infectious disease surveillance.
  - *Response*: To develop methods to assimilate communications data, social media data and public information to triangulate the status and trajectory of transmission chains. Further, to integrate real-time data on health, environmental and human resources in support of local and international responders.

## **Process**

We will begin by engaging health authorities from high-risk countries to understand local priorities and needs for data integration. Availability of datasets from industry, the research community, national public health surveillance, climate and environmental monitoring systems, health systems administration, social media feeds, and animal health services will then be sought in order to understand how their integration can fill critical knowledge gaps across disciplines. Reports and lessons learned from previous infectious disease outbreaks have identified clinical, genomic, demographic, pathogen and vector surveillance, communications, land-use, health administration, and environmental data as powerful inputs to support planning and operationalising outbreak response. We can anticipate data in numerous formats such as tabular data in spreadsheets, CSV, TSV, and/or plain text, geospatial point-wise data, geographic data, and a variety of XML and JSON dialects. For the domains of interest, available ontologies will be sourced and compared to determine methods for integration and interchange. Tools such as Fair Sharing and Linked Open Vocabularies are particularly useful for this task but where necessary, high level ontologies will be created to allow data integration.

## **Examples of high impact domains for integration**

- Combining human, animal and vector surveillance systems with measurements of relevant climate, environmental and ecological factors to strengthen risk assessments and improve transmission forecasts. Abundance of disease vectors and animal-reservoirs will be better understood when data on flooding, drought, land use changes, habitat loss, deforestation and warming are explored in aggregate.
- Linking human and pathogen genomic libraries with clinical data from individual patients to bring clinical significance to widely-shared gene sequences. Development of a system to link publicly catalogued gene sequences to internationally recognised clinical data interchange standards, notably CDISC and HL7, would amplify the scientific potency of human and pathogen genetics by giving them clinical meaning.
- Transmission networks can only be deciphered when maps of socio-economic, health and behavioural factors are superimposed. This can be achieved when data such as water access, vaccination coverage, and burial practices can be integrated.

## **Stakeholders**

- National public health institutes and ministries of health are the primary targets and stakeholders of this new multi-disciplinary application of data. Initial selection of domains will be led by the priorities of these agencies and the communities they serve.
- Data producers of interest include: Ministries of Health, public health agencies, environmental and agricultural agencies (including hydrologists, meteorological agencies etc.) NGO first responders, social media platforms, the public, researchers
- Time series of satellite imagery are also likely to be of high value.
- Data consumers of interest include: Ministries of Health, Health agencies, the public, NGO first responders, vaccine developers, academics, CDCs, WHO

#### ***Challenges/barriers***

- Due to concerns about privacy and ownership, clinical data remains one of the more challenging data domains to access.
- Genuine engagement of public health agencies and ministries of health in outbreak-affected countries is critical to define relevant priorities and approaches to tractable solutions.
- Sources of data are fragmented across a wide range of actors, repositories and governed access systems.
- The geographic, epidemiologic and socio-economic diversity of outbreaks may limit the extrapolation of tools developed on pilot projects.

### **3.4 PILOT PROJECT 1B – DISASTER RISK REDUCTION**

Project Leaders – Virginia Murray (Consultant in Global Disaster Risk Reduction, Public Health England), Kevin Blanchard (Senior Environmental Scientist, in Global Disaster Risk Reduction, Public Health England), Helen Green (Public Health Registrar, Public Health England)

#### ***Vision***

Disasters can significantly set back progress towards sustainable development and many are exacerbated by climate change. Evidence indicates that exposure to risk of persons and assets in all countries has increased faster than vulnerability has decreased. There are new risks and a steady rise over time in disaster related losses, with a significant economic, social, health, cultural and environmental impact in the short, medium and long term, especially at the local and community levels. Recurring small-scale disasters and slow-onset disasters in particular affect communities, households and small and medium-sized enterprises, and constitute a high percentage of all losses. All countries, especially developing countries where the mortality and economic losses from disasters are disproportionately higher, are faced with increasing levels of possible hidden costs and challenges in order to meet financial and other obligations.

Disaster risk reduction requires a multi-hazard approach and inclusive risk-informed decision-making based on the open exchange and dissemination of disaggregated data (including by sex, age and disability). We require easily accessible, up-to-date, comprehensible, science-based, non-sensitive risk information, complemented by traditional knowledge, as data on disaster impacts have been poorly documented so it is difficult to manage what cannot be measured. Furthermore, data that characterise many of the factors that influence this knowledge are available, but remain in siloes within the various domain-specific communities, formats and ontologies that created them.

The UN 2015 landmark agreement, the Sendai Framework on Disaster Risk Reduction 2015-2030, identified seven global targets. On 2 February 2017, the United Nations General Assembly endorsed the Report of the Open-ended Intergovernmental Expert Working Group on Indicators and

Terminology Related to Disaster Risk Reduction, and the recommendations for 38 indicators within the seven global targets relating to disaster risk reduction contained therein.

The Sendai Framework calls for enhancing the development and dissemination of science-based methodologies and tools to record and share information concerning disaster losses. The recorded disaggregated data and statistics then populate the indicators and subsequently monitor progress against defined targets. The United Nations International Strategy for Disaster Reduction (UNISDR) began the first cycle of monitoring using the online Sendai Framework Monitor in January 2018, and will exceptionally cover the two biennia 2015-2016 and 2017-2018. Our vision is to support UNISDR and partners to assist UN member states in determining and accessing data required for reporting against the global targets and indicators.

### ***Objectives***

- To improve the knowledge of the data resources and capacities within ISC, CODATA, LODGD, IRDR, their collaborators and the Scientific Unions that could facilitate partnership with UN member states and their newly required reporting of Sendai Framework data using the evolving 'digital revolution' skills and concepts.
- To assist in identifying and, where necessary developing, minimum standards, definitions and metadata for disaster-related data, statistics and analysis with the engagement of national government focal points, national disaster risk reduction offices, national statistical offices, the UN Department of Economic and Social Affairs and other national and local relevant partners.
- To assist in providing suitable research approaches and developing methodologies for the measurement of indicators and the processing and storage of statistical data with relevant technical partners.
- To improve the networking, knowledge sharing and capacity development of UN member state scientists and partners to help support local and national delivery of data collection, analysis and reporting
- To contribute to the reporting process for the global targets and indicators by the UN member states as required and assisting in identifying gaps in data knowledge to support mechanisms to address these gaps.

### ***Process***

We will begin by partnering with colleagues engaged with the development and implementation of the UNISDR [Technical Guidance for Monitoring and Reporting on Progress in Achieving the Global Targets of the Sendai Framework for Disaster Risk Reduction](#) (January 2018) and engaging data networks and authorities from UN member states to understand local priorities and needs for data integration. This will allow us to articulate the level of current data capacity and understanding within UN member states. We will also partner with ICS Scientific Unions and other organisations to determine their resources that could be used to support data availability and compatibility using where possible available, easily accessible, up-to-date, comprehensible datasets from national and local public surveillance, the research community, and climate and environmental monitoring systems. Public health system administrations, industry, social media feeds, and other services will then be sought to explore how to fill critical knowledge gaps across disciplines. For the domains of interest, available ontologies will be sourced and compared to determine methods for integration and interchange.

### ***Examples of the impact of integrating data***

Integrated data is required to show progress against the Sendai Framework's call for investment in developing, maintaining and strengthening people-centred multi-hazard, multi-sectoral forecasting and early warning systems, disaster risk and emergency communications mechanisms, social technologies and hazard-monitoring telecommunications systems. Such data is necessary to underpin systems that facilitate a participatory process; to tailor them to the needs of users, including social and cultural requirements (in particular gender); and to promote the application of simple and low-cost early warning equipment and facilities; and then broaden release channels for disaster early warning information.

Integrated data can:

- Promote and enhance access to, sharing and use of non-sensitive data and information, as well as communications, geospatial and space-based technologies, and related services. This is achieved through international cooperation including technology transfer.
- Maintain and strengthen in situ and remotely-sensed earth and climate observations.
- Strengthen the utilisation of media, including social media, traditional media, big data and mobile phone networks.
- Support national measures for successful disaster risk communication in accordance with national laws and culture.

Ebola is a recent disaster that demonstrates the need to combine human, animal and vector surveillance systems with measurements of relevant climate, environmental and ecological factors to strengthen risk assessments and improve transmission forecasts. Abundance of disease vectors and animal-reservoirs will be better understood when data on flooding, drought, land use changes, habitat loss, deforestation and warming are explored in aggregate.

### ***Stakeholders***

- National Focal Points for Disaster Risk Reduction are the primary targets and stakeholders of this new multi-disciplinary application of data for data reporting, including Civil Contingencies Secretariats or National Disaster Management Agencies, national public statistical institutes and relevant ministries including Ministry of Finance/Treasuries. Initial selection of domains will be led by the priorities of these agencies and the communities they serve.
- Data producers of interest include: Government Ministries with data such as Ministries of Health, statistical agencies, Met Services, NGO first responders, the public, researchers linked to ISC Scientific Unions and other relevant organisations.
- Data consumers of interest include: UN member states focal points for data collection for Sendai monitor including national focal points for disaster risk reduction such as civil contingencies secretariats or national disaster management agencies, national statistical offices, ministries of health, the public, NGO first responders, academic partners and UN partners.

### ***Challenges/barriers***

- Opportunity to convert disaster data to useful, usable and used knowledge to provide evidence to inform policy and practice to reduce impacts.
- Recognising that data on disasters is unlikely to be complete because of its complexity, there will be requirements for ethical statistical and Bayesian hierarchical modelling and other techniques.
- Definitions and standards for disaster-related data and its use to identify the measures required by the global targets and indicators.

- Scale, types and duration of disasters needs to be addressed across UN member states if possible.
- Fragmentation of disaster data sources which are widely siloed in many UN member states.
- Learning how to implement and maximise the potentials of the ‘digital revolution.’

### **3.5 PILOT PROJECT 1C – RESILIENT CITIES**

Project Leader – Stephen Passmore (Technology Director, Ecological Sequestration Trust & Resilience Brokers).

#### ***Vision***

The 21st Century is the century of urbanism: cities are now humans’ dominant habitat. By 2030, urban areas are projected to house 60% of people globally and one in every three people will live in cities with at least half a million inhabitants. Cities already contribute 70% of global GDP while consuming over 60% of the global energy, producing over 70% of global waste and over 70% of greenhouse gas emissions.

Urban areas are complex adaptive systems that include diverse institutional structures, and that are structured to satisfy the human needs for water, shelter, energy and connectivity. The complexity of cities and urban settlements is often thought of as a system of systems, interwoven by transport, water, energy, food, building, health, education, emergency, ecological and cultural systems. With the majority of humanity now residing in cities, they are now the focus of unprecedented global commitments for tackling the major humanitarian challenges of our age including climate change, disaster risk reduction and sustainable development. Indeed, the battle for sustainable development is said to be likely won or lost in cities.

Conventional urban governance frameworks have largely approached infrastructure, service and utility provision and protection from specific threats, and tend to manifest in silo-based and hierarchical governance practices that frequently overlook the organisational and socio-political factors shaping local adaptive capacity. Further urban governance structures are often situated within rigid spatial boundaries set by political demarcation. Indeed, data that characterise many of the interventions in urban areas are available within the various domain-specific communities, often with their own silo-specific formats, vocabularies and ontologies. Given the increased complexity of the urban operating environment, recent governance concepts (e.g. the Sendai Framework, SDG11, 100 Resilient Cities, World Urban Forum 9) call for better integration in order to account for the contextual and interdependent nature of the complex urban system.

Within this context the main benefit of these new integrated governance models relates to their ability to draw into the discussion previously unconnected and new stakeholders. By facilitating shared understanding and fostering enhanced connectedness and reorganisation, this process by itself creates resilience. However, there still is an implementation gap in operationalising urban resilience, related to challenges with mainstreaming resilience into everyday urban governance.

At the core of this observation is the argument that data collection and interpretation mediates, incentivises and provides evidence of resilient governance practices. They mediate because they highlight what and who “matters” within the urban system. The mediating role of sound data practices for resilient governance is further evident insofar as they lead to reflective processes and contribute to the development of shared understanding. By providing an assessment framework,

data practices incentivise resilient governance. They are essential for reducing the resilience implementation gap, as most of the recent frameworks call for integration, yet assessment indicators do not appear to measure the degree to which interdependencies are addressed.

Integrated data approaches and new combinations of sciences can bring data and evidence to support decision making in city-region governance structures to achieve healthier, more resilient and sustainable development paths. From molecular to human to planetary, data and interdisciplinary science can achieve better outcomes for humanity when applied to city-region solutions.

### ***Objectives***

- To assist in enhancing standards and definitions for metadata, research approaches and methodologies for the measurement of indicators at the city-region level . This is aligned to SDG11, the Sendai Framework's ten essentials and UN Member States' reporting requirements, and informed by existing city-region data frameworks.
- To develop tools that enable data infrastructure/ecosystem assessment and enhancement in city-region pilots, together with capacity building to realise the potential of modern data resources and design and track progress against global humanitarian challenges such as Sustainable Development, Climate Change, Planetary Health and Disaster Risk Reduction.
- To demonstrate the degree to which the representativeness of research approaches, data collection, processing and disseminating used for municipal decision-making lead to enhanced urban resilience.
- To identify the degree of interdependence in data analysis frameworks used for monitoring and evaluation of urban services
- To determine the extent to which data has been sourced in a participatory manner.
- To assist city-regions to utilise integrated data and interdisciplinary research to underpin 'bankable' project development which attracts investment and demonstrates the business case for data infrastructure.

### ***Process***

Establish international working group to review, define and implement a work programme to deliver objective one and scope for objective two (above). Identify two or three city-regions to act as exemplars for data, standards and interdisciplinary integration, ideally with Resilience Offices, which are representative of the diversity of city-regions globally (mega city, rapidly growing city, coastal city). Establish data collaborative laboratories, 'collaboratories,' in each city-region consisting of representatives from national and local government, local academic and private sectors and civil society groups. Select context specific urban challenges (water and sanitation, transport etc.), and/or infrastructure and policy intervention solutions. Map interdependencies and constituents (urban stakeholders – municipal governance sectors, service providers, NGOs, the urban communities) and analyse the municipal data system in the context of the interdependencies and degree of integration. Work with the ICS scientific unions and other organisations to determine resources that could be used to support data availability and compatibility. Establish tools to support city-region governance based around a data specification to include appropriate disaggregation (gender, age, socio-economic), scalability from neighbourhood to city-region, geospatial data of land use change, infrastructure and ecological systems and movement patterns, temporal information including historical and 'smart' short term, pervasive sensing data infrastructure, environmental, climate and weather, supply-chain and city-region level imports and exports, Hazard and risk and markets including jobs, skills, sectors, housing, buildings and infrastructures, human health and wellbeing.

### ***Stakeholders***

- International Science Unions, GPSDD, UNISDR, UNSDSN, UNFCCC, WHO, International City networks C40 and ICLEI.
- Data producers Group on Earth Observation (GEO), government ministries, municipal governance institutions including service providers, statistical services and Meteorological offices, data observatories, scientific unions, local academic institutions, Urban Civil Society (Slum Dwellers International) and other relevant organisations.
- Data consumers of interest include municipal governments, government ministries, citizens groups, private sector developers and businesses, investors, insurers, credit ratings agencies

#### **Challenges/barriers**

- Complexity of interdependencies and ability to represent them to understand whether data accounts for them or not.
- Need to develop a holistic view of the urban data system (may only be partial).
- Extrapolation of tools from two or three city exemplars to all city-regions globally.

### **3.6 COMMON DATA SETS AND HIGH LEVEL INTEGRATION (STAGE 3)**

It is notable that the three pilots all expect to make use of very similar data, albeit in different ways. Geospatial data is foundational to all of them but the very general term 'geospatial data' includes everything from the locations of buildings and land use, through to water catchments, air and water quality, geology, climate and more. OGC standards are central to all these domains and are designed to interoperate. Away from the environmental sciences, statisticians routinely relate their data to regions on a map, as do a variety of social scientists. All three pilots anticipate making use of social and professional media sources where, again, location is almost always to be found within the data. Both animal and human health information, as well as demographics, typically relate to location.

With location as a common factor in almost all of the data expected to be used in the pilots, integration might appear to be easy. However, this is rarely the case as there are so many ways of referring to it. Trivial differences exist between recording a location as, for example, 'USA', 'US' and 'America' (and all three may readily appear in the same dataset) through to much more substantial differences. For example, national mapping agencies – and therefore maps produced by local and national government departments - will use a local coordinate reference system. This is not the same as latitude and longitude used by, for example, social media, and the conversion between the two requires computation using spherical geometry. Between the two extremes are many variations in the size of area covered, areas that overlap and so on.

By solving the integration problem for disparate datasets relevant to the three pilots, the project will also create a firm basis for further exploration of any discipline where human actions and natural forces impinge on each other, and provide a sound basis for stage 3 of high level integration (see diagram).

### **3.7 DATA SCIENCE SUPPORT FOR THE PILOTS**

For the initial pilots to have maximum chance of success we must ensure that they not only address a significant issue, but that they are also feasible from a technical and resource perspective. Data science expertise and experience is needed to assess and determine:

- Availability and accessibility of data sources needed to address the research question. There are technical issues (around availability, compatibility or transparency of formats and data structures), as well as legal or policy and social issues (are the data cleaned of Personally Identifiable Information (PII), or properly anonymized? Are the custodians willing to allow

re-use and are there any restrictions on re-use? Are there non-disclosure agreements that need to be signed? If so, who has the authority to sign them?)\*.

- What is the state of vocabularies, data dictionaries, ontologies – anything that can be used to understand the semantics (meaning), not just the syntax of the data? Are the data labelled? Is the metadata consistent and complete? Are the data in consistent units, or units that are straightforward to convert into a common system?
- What are the foundations of the vocabularies – are they compatible across multiple domains? Is it necessary and feasible to create a more abstract semantic layer that encompasses multiple domains? Are the vocabularies in use governed appropriately, and available from controlled and persistent name spaces, and under suitable change control management? Can we identify a suitable authority to govern each vocabulary? Is the authority able and willing to make the vocabularies available in suitable forms, or do we need to offer technical assistance?
- Are the data resources of tractable size and complexity? Whilst in most cases TB-scale data is likely to be tractable, we should avoid starting with a project that requires highly complex linked databases.
- What tools are necessary and available in order to support the pilot study goals and how are they licenced? It would appear, for example, that each of the three research challenges will require use of geospatial information.
- What computational and data management platforms will be needed and are available to carry out the pilot studies?
- Which technical staff will actually work on the pilot projects, and how will they communicate with the domain science leaders? What level of programming support and expertise will be required, or can the pilot studies be carried out with COTS or open source software?

For each pilot a requirements and capabilities matrix can be developed, and if a pilot study appears to demand requirements that are out of scope, the data science experts and the domain science leaders will have to negotiate a downsizing or reprioritisation of pilot project components.

## **4 STRAND 2 – EFFICIENT USE OF DATA RESOURCES ACROSS THE DISCIPLINES OF SCIENCE**

---

### **4.1 THE CURRENT STATE OF DATA SCIENCE AND DATA STANDARDS ACROSS THE DISCIPLINES**

Open Data processes and the provision of supporting services have tended to be developed most readily in those disciplines where well-defined data streams are created by high throughput instruments, such as in astronomy, bio-informatics, crystallography, earth observation, high energy physics etc. or in the social sciences where consistent time-series records have been kept of policy-relevant parameters in education, health or production systems. However, vast amounts of research data are acquired by individual investigators or small teams known as the ‘long-tail communities’ who rarely have access to community-based data infrastructures that will ensure persistent access, quality control, and standardisation. A major problem for many disciplines therefore lies in harmonising and extracting meaningful features from such varieties of data sources.

Two interrelated problems arise from this. Some disciplines have not developed the protocols necessary to realise the potential for discovery from the data resources available to them because of the difficulty of coping with complex, mixed data sources. Many disciplines that are developing their own systems tend to do so entirely within a disciplinary frame, leading to a worrying level of duplication/ incompatibility between the plethora of data and information standards (including vocabularies, ontologies) within and between them. There are thus two challenges for the disciplines:

- to develop and implement the standards that are crucial to effective data discovery and utilisation where this has not yet been done;
- to develop means whereby established standards and those to be developed by disciplines new to the task, can interoperate with data from other disciplines.

Science is an international enterprise, where the adoption of common standards and protocols for disciplinary or application-domain use should in principle be most effective through international agreement by international representative bodies. Although the international scientific unions of ICSU and the associations of ISSC might seem to be ideal vehicles for this role, in practice there has been little connection between them and those groups that are developing best practices in data and informatics infrastructures and services at disciplinary or sub-disciplinary level, such as Elixir in bioinformatics, the Database of Crystal Structures, the Earth Science Information Partners and by international cross-disciplinary bodies such as the Research Data Alliance, and national efforts such as the Australian National Data Service, EU Horizon 2020 projects, etc.

Many scientific unions have been very active in coordinating and agreeing nomenclature and standards for description and analysis in their own disciplines. For example, the International Union of Geological Sciences' (IUGS) Commission on Stratigraphy precisely defines standards and procedures for measuring Earth history; the International Astronomical Union (IAU) defines precise nomenclature for astronomical features; the International Union of Pure and Applied Chemistry (IUPAC) defines nomenclature for chemical substances; the World Values Survey has been adopted and advocated by the International Union of Psychological Sciences as a standard, etc<sup>1</sup>.

Unfortunately, this logic, whereby scientific unions take responsibility as a legitimising body on behalf of their discipline for standard nomenclature has not, with notable exceptions (e.g. the International Union of Crystallography IUCr), been generally pursued into the digital data arena where standardised vocabularies and ontologies are vital tools in data discovery and interoperability. It is important that all relevant areas of research understand the level of data standardisation that is required in their discipline for effective discovery, access, and use of data, and the problems that need to be solved in using data in inter- and trans-disciplinary modes.

In principle we suggest that the science unions should review and, where appropriate, endorse what their constituency believes is best practice in data sharing and interoperability, as well as the authoritative vocabularies and standards that support this. We therefore argue that international scientific unions and associations that do not address the issue of data standards and should either:

- where there are non-union/association groups that have created such standards as a basis for data services, they should consider reviewing, and potentially adopting, such standards at union/association level;

or:

- mobilise or create their own standards body to commence this task.

Coordinating data standards is essential at an international level. It is an imperative that we should aspire to a state where we are able to discover and access data from across, and beyond, the individual sciences to enable data from all disciplines be used in transdisciplinary projects, particularly those that seek to tackle the most pressing environmental and societal issues facing humanity, such as in the ICSU Future Earth programme. Without a strategic effort to harmonise the standards and data infrastructures that are being developed it will be almost impossible to access the breadth and depth of scientific data, both today and in future, to provide transparent, evidence-based advice to governments on pathways to global sustainability.

Sharing data, information and services in the most efficient and accessible way, and utilising them to best effect in the creation of new knowledge, is dependent on the development and use of common practices for the discovery, access, sharing, interpretation and retention of these data. Many, if not most, of the decisions about what to store, what shared agreements or standards to apply, and what are the minimum required metadata lie, or should lie, with the relevant disciplines and the international scientific unions that help define the priorities, principles and needs of those disciplines. It is vital that they systematically concern themselves with raising awareness of, and promoting such standards.

The increasing numbers of disciplines, scientific unions and associations that have created “information communities” to engage with the digital challenge have established, often through trial and error, important lessons about what works and what does not, and their experience forms an important guide to later-comers:

- Collection of high-quality data is facilitated when there is prior agreement about data collection, data format and metadata standards.
- Easily understandable and user-friendly implementation mechanisms are key to the adoption of standards by the research community. Data and metadata standards that are not easily usable tend to be ignored. Web-based tools are ideally suited to this since they require no special software to be installed or learnt, and are inherently distributed.
- Traditional QA/QC, statistical and visualisation approaches do not typically scale to big, multidisciplinary data. New algorithms and approaches are often required.
- Vocabularies and lexicons used in one scientific discipline or domain are not universally understood, and significant effort and engagement are necessary to bridge disciplinary boundaries. Sharing across disciplines is however essential, and must be supported either through cross-disciplinary coordination, 'mappings' of concepts from one discipline to another, or by the development of core models that express the common elements of science data.
- Big science challenges often demand big data and such data can normally only be acquired and processed using machine-automated approaches. Successful use of big data depends on the degree to which data follow common structures and include the requisite structured metadata that can enable automated interpretation of the data.
- Groups that have had an engagement strategy with their community coupled with the offer of technical enhancements, such as those listed in [Appendix 1](#), have succeeded in attracting widespread use of their services.

## **4.2 MARSHALLING AND ENGAGING THE SCIENCE UNIONS**

Many disciplines and scientific unions and associations are either not engaged with data standardisation or have not developed adequate vocabularies and standards. In some cases where standards do exist, a “preferred authoritative standard” has not been clearly endorsed. In addition, existing standards often do not manage characteristics such as expressivity and data quality, which

are important for integration with other data types by non-experts. It is important that all relevant areas of research understand the level of data standardisation that is required in their discipline for effective discovery, access, and use of data, and the problems that need to be solved to use data in multi- and trans-disciplinary modes.

A CODATA Task Group on Coordinating Data Standards amongst Scientific Unions has been created to develop priorities and coordinating actions for Strand 2, in collaboration with ICSU. Its immediate priorities are:

- identify the level of activity in ICSU/ISSC Unions/Associations; including whether they have established either a commission on data and information, or identify the point of contact for information standards being developed/governed/endorsed by their Unions;
- raise awareness of standards endorsed by and/or being developed by Unions to assist in promoting authoritative standards and minimising duplication of effort;
- create web-accessible pages that provide links to repositories for data models, information standards, vocabularies, ontologies, etc., for each of the unions;
- determine a broad 'maturity model' for scientific standards adapted from the 5 star Open Data model (<http://5stardata.info>) and the American Geophysical Union's Data Maturity Framework (<http://dataservices.agu.org/dmm/>) that provides a guide to users as to the usability of the standards and a guide to developers about the overall maturity of their standards within the international scientific community and assist in ensuring 'fitness for purpose'.
- provide best practice examples for the development and application of the required standards and guidance on developing governance frameworks for the maintenance and revision of these standards, preferably by assisting linkages to key groups such as the international Research Data Alliance (RDA), as well as national efforts such as the Australian National Data Service (ANDS), the Earth Science Information Partners (ESIP), EU 2020 projects, etc.; and
- provide guidelines to the scientific community for the need to adhere to these standards and promote the benefits of adherence to standards to increase discovery and accessibility to data.

The CODATA Task Group has completed an initial analysis based mainly on website information and/or knowledge by group members of a particular union. No direct approach has yet been made to any union for information.

The initial analysis has shown that there are three categories of standards development:

- a) Unions that are fully engaged and are already leaders in developing the standards required for this project (e.g., IUCr, IUGS, IAU);
- b) Unions that accept that they need to be involved, but are unsure how to proceed, and do not currently have anything that could be utilised by this project; and
- c) Unions that are not engaged at all.

Note that some in category 2, although they do not have standards, could possibly participate by providing scientific insights and endorsement for standards relevant to their discipline that have been developed outside of their Union.

The next step is to approach the unions directly to:

- a) create a state of play on the level of engagement of each union, and where they exist their Commissions on Data and Information; and

- b) develop an authoritative, web accessible inventory of standards that are being developed and/or coordinated and/or endorsed by the unions.

As this is a direct approach to each union, the CODATA Task Group will need ICSU and ISSC will support in constructing and circulating a questionnaire to their unions and associations to establish the extent to which the above priorities are being addressed, and if not, how they might be.

CODATA will then follow up this enquiry to discuss how those unions and associations that wish to develop this aspect of their work and the extent to which packages of work designed to create appropriate standards, vocabularies and ontologies can be supported by the CODATA community.

It is hoped that information from this inventory will also provide information on the pattern of data research capacity across the disciplines.

The inventory will raise many issues. In particular it will raise the questions as to whether:

- a) the Unions/associations accept a fundamental responsibility to ensure that the vocabularies etc. in their disciplines enable the discipline to exploit modern digital resources?;
- b) we can encourage the Unions to work with the Research Data Alliance to improve, or develop, the required data infrastructures to enable interoperability both within and between the science unions?; and
- c) we can explore what role CODATA will have in coordinating this activity (given there are already joint CODATA/RDA Interest Groups and Working Groups).

### **4.3 MERGING THE CODATA TASK GROUP WITH THE 3 PILOT ACTIVITIES**

The following logical steps are suggested:

- 1) Each pilot activity will need to scope what data it needs to access and what tools it needs to process the data. This should be done independently of any knowledge of what is actually available. That is, the initial list of requirements should not be scoped around what is assumed to be easily available and what is 'safe'.
- 2) The list needs to be evaluated as to what is available and what is not and compared with the standards lists developed by the CODATA Task Group. It is highly unlikely that the required data standards are all located within the science unions, and there will need to be follow up to see if the required standards are accessible elsewhere.
- 3) The list of what is available will then inform the pilot projects of their potential scope and highlight where some activities will not be feasible in the immediate time frame.
- 4) The list of data that is not available, or the data is not in formats suitable for modern online digital analysis, will provide valuable information on which science unions could be motivated as a priority to for them to become involved in standards development, endorsement and/or coordination.

## **5 COMMUNITY OF INTEREST AND EXPERTISE**

---

The discussions and contributions to the development of the initiative have involved an aggregate of 70 scientists, including domain scientists from scientific unions, associations and other scientific bodies, and data scientists from a range of national and international bodies. This group now forms the core of a community of interest and expertise with the purposes of:

- contributing to the development of the initiative and advocating its priorities;

- commenting on priorities and approaches and providing advice and support;
- participating and collaborating in the activities of the long-term process as interests and responsibilities dictate.

1.	David	<b>Abreu</b>	CGIAR
2.	Suchith	<b>Anand</b>	Geo4All / Global Open Data for Agriculture and Nutrition (GODAN)
3.	Phil	<b>Archer</b>	World Wide Web Consortium (W3C)
4.	Tom	<b>Baker</b>	Dublin Core Metadata Initiative (DCMI) and Global Agricultural Concept Scheme (GACS)
5.	Yiming	<b>Bao</b>	Beijing Institute of Genomics, International Union of Biological Sciences (IUBS)
6.	Barbaros	<b>Gönençgil</b>	International Geographical Union (IGU)
7.	Franz-Josef	<b>Behr</b>	International Cartographic Association (ICA)
8.	Hugo	<b>Besemer</b>	RDA Interest Group on Agricultural Data (IGAD)
9.	Kevin	<b>Blanchard</b>	Public Health England (PHE) and UK Alliance for Disaster Research (UKADR)
10.	Niklas	<b>Blomberg</b>	European life-sciences Infrastructure for biological Information (ELIXIR)
11.	Timo	<b>Borst</b>	Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW)
12.	Geoffrey	<b>Boulton</b>	CODATA, FRS and University of Edinburgh
13.	John	<b>Broome</b>	CODATA
14.	David	<b>Carr</b>	Wellcome Trust
15.	Simon	<b>Coles</b>	International Union of Crystallography (IUCr)
16.	Simon	<b>Cox</b>	CSIRO and CODATA Task Group
17.	Vasa	<b>Curcin</b>	Medical Bioinformatics Group, King's College London
18.	Rachel	<b>Drysdale</b>	European life-sciences Infrastructure for biological Information (ELIXIR)
19.	Thomas	<b>Emery</b>	International Union for the Scientific Study of Population (IUSSP)
20.	Peter	<b>Fox</b>	Rensselaer Polytechnic Institute (RPI)
21.	Jeremy	<b>Frey</b>	International Union of Pure and Applied Chemistry (IUPAC)
22.	Patrick	<b>Garda</b>	Ministere de l'Enseignement Supérieur et de la Recherche
23.	Philippe	<b>Gaucher</b>	Ministere de l'Enseignement Supérieur et de la Recherche
24.	Catherine	<b>Geissler</b>	International Union of Nutritional Sciences (IUNS)
25.	Helen	<b>Glaves</b>	Oceans Data Interoperability (ODIP) BODC
26.	Heide	<b>Hackmann</b>	International Council for Science (ICSU)
27.	Bob	<b>Hanisch</b>	National Institute of Standards and Technology (NIST)
28.	John	<b>Helliwell</b>	International Union of Crystallography
29.	André	<b>Heughebaert</b>	Belgian Biodiversity Platform and Global Biodiversity Information Facility (GBIF)
30.	Tim	<b>Hirsch</b>	Global Biodiversity Information Facility (GBIF)
31.	Gobe	<b>Hobona</b>	Open Geospatial Consortium (OGC)
32.	Simon	<b>Hodson</b>	CODATA
33.	Markus	<b>Jobst</b>	International Cartographical Association (ICA)
34.	Andreas	<b>Kempf</b>	Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW)
35.	Dimitris	<b>Koureas</b>	Biodiversity Information Standards (TDWG)
36.	Bryan	<b>Lawrence</b>	National Centre for Atmospheric Science (NCAS)/WM
37.	James	<b>Malone</b>	European Bioinformatics Institute (EMBL-EBI), Open Biomedical

			Ontologies (OBO)
38.	Bob	<b>Mann</b>	International Astronomical Union (IAU) and Royal Observatory, Edinburgh
39.	Pam	<b>Maras</b>	International Union of Psychological Science (IUPsyS)
40.	Silvia	<b>Massaro</b>	International Union of Geodesy and Geophysics (IUGG)
41.	Peter	<b>McQuilton</b>	FAIRsharing and Oxford eResearch Centre (OeRC)
42.	Claire	<b>Melamed</b>	Global Partnership for Sustainable Development Data (GPSDD)
43.	Laura	<b>Merson</b>	Infectious Diseases Data Observatory
44.	Ellinor	<b>Michel</b>	International Union of Biological Sciences (IUBS)
45.	Bill	<b>Michener</b>	University of New Mexico, DataONE and Dryad
46.	Virginia	<b>Murray</b>	Integrated Research on Disaster Risk (IRDR) and Public Health England (PHE)
47.	Stephen	<b>Nortcliff</b>	International Union for Soil Sciences (IUSS)
48.	Isaac	<b>Nyamongo</b>	International Union of Anthropological and Ethnographic Sciences (IUAES)
49.	Marc	<b>Nyssen</b>	International Union for Physical and Engineering Sciences in Medicine (IUPESM)
50.	Neave	<b>O'Clery</b>	PEAKURBAN and University of Oxford
51.	Mike	<b>Oldham</b>	National Physical Laboratory (NPL)
52.	Tom	<b>Orrell</b>	Global Partnership for Sustainable Development Data (GPSDD)
53.	Stephen	<b>Passmore</b>	'Resilient Cities', Urban Health and Wellbeing Programme and 'Resilience Brokers'
54.	Andrea	<b>Perego</b>	European Commission
55.	Jean-Luc	<b>Peyron</b>	International Union of Forest Research Organizations (IUFRO)
56.	Gillian	<b>Petrokofsky</b>	International Union of Forest Research Organizations (IUFRO) and University of Oxford
57.	Steven	<b>Ramage</b>	Group on Earth Observations (GEO)
58.	François	<b>Robida</b>	International Union of Geological Sciences (IUGS)
59.	Marina	<b>Romanchikova</b>	National Physical Laboratory (NPL)
60.	John	<b>Rumble</b>	CODATA Nanomaterials WG
61.	Alena	<b>Rybkina</b>	International Union of Geodesy and Geophysics (IUGG)
62.	Susanna	<b>Sansone</b>	FAIRSharing and Oxford eResearch Centre (OeRC)
63.	Irene	<b>Schöfberger</b>	International Council for Science (ICSU)
64.	Ingo	<b>Simonis</b>	Open Geospatial Consortium (OGC)
65.	Jon	<b>Todd</b>	National History Museum (NHM)
66.	Maria	<b>Uhle</b>	National Science Foundation (NSF) and Belmont Forum
67.	Jean-Pierre	<b>Vilotte</b>	Institut de Physique du Globe de Paris (IPGP)
68.	Joachim	<b>Wackerow</b>	Leibniz Institute for the Social Sciences (GESIS) and Data Documentation Initiative (DDI)
69.	Sally	<b>Wyatt</b>	Huygens Institute
70.	Lesley	<b>Wyborn</b>	CODATA TG and Australian National University