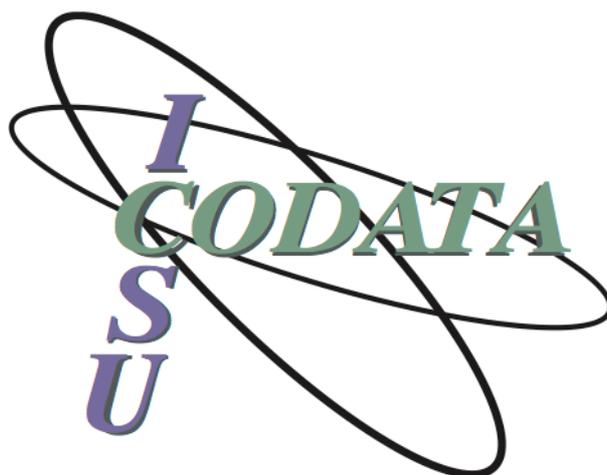


# Mobilising the Data Revolution: the CODATA strategy



*May 2015*

Prepared by the Officers and Executive Committee of  
CODATA, the ICSU Committee on Data for Science and Technology



*Members of the newly elected CODATA Executive Committee at the Indian National Science Academy, New Delhi, following the General Assembly in November 2014.*

From left to right: Simon Hodson, Executive Director; Anil Kumar, India; Paul Laughton, South Africa; Mary Zborowski, Canada; Sara Graves, Secretary General, USA; John Broome, Treasurer, Canada; Geoffrey Boulton, President, UK; Niv Ahitiv, Vice-President, Israel; Bonnie Carroll, USA; Mark Thorley, UK; Der-Tsai Lee, China: Taipei; Alena Rybkina, Russia.

Not present: Huadong Guo, Past-President, China; Takashi Gojobori, Vice-President, Japan; Jianhui Li, China; Sarah Callaghan, UK.

## Table of Contents

<b>Preface</b> .....	<b>1</b>
<b>A: Context and Vision</b> .....	<b>1</b>
The data revolution .....	1
Why it matters for science .....	1
Rising to the Challenge .....	2
<b>B: CODATA, ICSU and International Collaboration</b> .....	<b>3</b>
<b>C: National Data Systems and International Programmes</b> .....	<b>3</b>
<b>D: CODATA's Strategic Response</b> .....	<b>5</b>
<b>E: Priority 1. Data Principles and Practices</b> .....	<b>6</b>
Purpose .....	6
Outputs .....	6
<i>a) An assessment tool</i> .....	6
<i>b) The data agenda for international science</i> .....	6
<i>c) Other planned/proposed activities - 2015/16</i> .....	7
Outcomes .....	8
<b>F: Priority 2 – Frontiers of Data Science</b> .....	<b>8</b>
Rationale .....	8
Objectives .....	9
Outputs .....	9
<i>a) International Data Week 2016</i> .....	9
<i>b) The Data Science Journal</i> .....	9
<i>c) Uniform Description System for Nanomaterials</i> .....	9
<i>d) Inter-Union Work on Data Standards and Ontologies</i> .....	9
<i>e) Integrating Geospatial Data on the Web</i> .....	10
<i>f) Big Data and Data Integration for International Science</i> .....	10
Outcomes .....	10
<b>G: Priority 3 – Capacity Building</b> .....	<b>11</b>
Rationale .....	11
Outputs .....	11
<i>a) Data training workshops</i> .....	11
<i>b) An initiative in Sub-Saharan Africa</i> .....	12
<i>c) A generic approach to LMICs</i> .....	12
Outcomes .....	12
<b>H: Roles of Committees, Task Groups and Work Groups</b> .....	<b>13</b>
Executive Committee .....	13
Other Strategic Committees .....	13
Task Groups .....	13
Working Groups .....	13
<b>I: Budget and Business Planning</b> .....	<b>14</b>
<b>J: Membership</b> .....	<b>15</b>
<b>K: Outreach, Communications and Reporting</b> .....	<b>15</b>

## Preface

*CODATA's second strategic plan covers the period 2013-2018. Working within this context, a new Executive Director, appointed in August 2013, has had a significant impact by building on existing activities and developing new and influential activities. A new President and new Executive Committee, elected in 2014, have used this record of recent achievement, and the comments of the 2014 ICSU review of CODATA, to sharpen the strategic focus of CODATA's work, with a clear view of the way that the small secretariat but large CODATA community can use their resources in the most productive way. Three major strategic programmes are identified that will form clear priorities for CODATA to the end of the planning period in 2018. They are: a) data principles and practice, b) frontiers of data science and c) capacity building. These priorities are central to exploiting the scientific potential of the data revolution. Their credibility as achievable objectives is demonstrated by CODATA's impressive record of recent achievement. This document argues the relevance of these priorities, sets objectives, identifies planned outputs and indicates the intention to review outcomes at the end of the planning period.*

## A: Context and Vision

### The data revolution

1. Recent decades have seen an unprecedented explosion in the human capacity to acquire, store and manipulate data and information and to instantaneously communicate them globally, irrespective of location. It is a world historical event involving a revolution in knowledge creation, communication and utilisation as profound as and more pervasive than that associated with Gutenberg's invention of the printing press.
2. This revolution has already produced fundamental changes in human, social and economic behaviour, but it is a revolution that has not yet run its course. Although its implications and benefits are yet to be fully understood or realized, it is clear that it has profound implications for science by opening up powerful new opportunities for discovery whilst challenging many of the ingrained habits of researchers and their institutions.

### Why it matters for science

3. Science has rapidly moved from an era of scarcity, in which, with some exceptions, data has been small in volume and sparse in distribution, with statistical techniques optimised to extract information from such limited data, to one of abundance, in which an unprecedented storm of data offers major opportunities and profound challenges for science. Many of these opportunities and challenges arise from so-called "big data" for which classical statistical approaches are often inadequate. The data are "big" because of the volume that systems must ingest, process and disseminate; because of their diversity and complexity; and because of the rate at which data streams in or out of the systems that handle them. Terabyte-sized data sets – and data collections measured in multiple Petabytes – are now common in Earth and space sciences, physics and genomics etc. Data from diverse sources in particular pose problems of integration that require novel solutions.
4. These developments have created major new opportunities for science:
  - to identify patterns and processes in phenomena that have hitherto been beyond our capacity to resolve;
  - to integrate data reflecting a wide variety of coupled processes to obtain much deeper understanding of relationships than has hitherto been possible;
  - to improve forecasts of system behaviour by integrating data acquisition and modeling;
  - to make cognate datasets and data-integration tools readily available and useable by individual researchers from the rapidly growing number of open databases;
  - to permit re-use, re-combination and re-purposing of data in ways that make data of perennial, cumulative value, rather than being lost from generation to generation;
  - to exploit the opportunities created by inter-communicating, automated sensors (the "internet of things") in exploring complex phenomena and unraveling complexity.

5. The exploitation of these opportunities – and the application of the new techniques of ‘big data science’ – is essential if we are to address the major, fundamentally interdisciplinary, scientific and social grand challenges emerging from the changing planetary environment.
6. At the same time, however, science has been sleep walking into a crisis as a consequence of the difficulty of adhering to a fundamental principle that has been the engine of scientific progress in the last three centuries. That is the principle of self-correction, that the evidence (the data) underpinning published concepts should be concurrently available to permit others to scrutinize the logic of the data-concept relationship, attempt replication of the observations or experiments, and thereby support or invalidate the concept. Observing this principle with large and complex datasets is onerous, not only requiring data to be intelligently open<sup>1</sup> in an electronic database, but to include the metadata (data about data) that permits it to be appraised and re-used. This needs to become a non-negotiable criterion for a published scientific article if the vital principle of self-correction is to survive, but its implementation will depend upon software tools and management systems that ease the process of intelligent data deposition – as well as cultural shifts in the way scientific contribution is recognized and rewarded.

## Rising to the Challenge

7. Open sharing of data and their availability for re-use and re-purposing are essential priorities if the opportunities listed in paragraph 4 are to be seized. The benefits to individual researchers of harvesting data from a wide variety of open sources offers them and the scientific enterprise a far more effective and efficient use of the data corpus than depending exclusively on self generated data, and particularly if that data remains closed to wider integration.
8. Such processes are however inimical to many ingrained habits, not only for individual researchers and their institutions, but also for those who fund, publish and evaluate research and those that seek commercial benefit from it. Fundamental changes in policy, process and motivation are required if those habits are to be changed and the opportunity grasped. Such changes are required at national levels where policies that determine the motivations of research institutions are set and at institutional level where support and motivation is provided for researchers and how they work. National learned societies and international unions both reflect and determine the principles and priorities of their disciplines. Some disciplines have led the way in taking up the data challenge whilst some are struggling to evaluate whether and how best to exploit it in their own context.
9. There are however boundaries to openness, namely legitimate commercial interests, privacy, safety and security. They are difficult boundaries to delimit with precision, and need careful consideration. They must not be used to justify blanket exceptions to openness.
10. There are also deeper issues of the social setting within which science is done. An open data environment offers novel possibilities for commercial innovation, for greater involvement of a wider range of stakeholders and citizens in co-production of knowledge, and for a deeper democratic engagement with the ways that scientific knowledge is created and used. These broader horizons, of integrating data from a wide variety of

---

<sup>1</sup> Openness of itself is of no value unless it is “intelligently open” – that it is discoverable, accessible, intelligible, assessable and useable.

disciplines and transforming science into a more public enterprise rather than one conducted behind closed laboratory doors add a further, more profound reason for responding energetically to the opportunities offered by the data revolution. They enhance the possibility of producing globally integrated science solutions and therefore, as flagged in paragraph 5, above, of responding globally to the major problems facing the planet and its inhabitants through transformations in behaviour and approach for which neither the scientific community nor the general public are well prepared. Such programmes are currently typified within ICSU by *Future Earth*, *Integrated Research on Disaster Risk*, and *Health and Wellbeing in the Changing Urban Environment*, and beyond by the *Science and Technology Alliance for Global Sustainability*, which need to ensure that their data strategies are relevant to these wider horizons.

## B: CODATA, ICSU and International Collaboration

11. Creating the policies and processes for data-intensive science that will enable effective and equitable international collaboration is vital both to scientific progress and to maximising national and regional benefits. In our view, effective exploitation of data is the single most important international issue of “policy for science” (as distinct from “science for policy”). In this context, the 2014 General Assembly of the International Council for Science (ICSU) approved a statement on *Open Access to Scientific Data* and endorsed the *OECD Principles and Guidelines for Open Access to Research Data from Public Funding*. CODATA, the body set up by ICSU to promote access to data for the scientific community, has both expertise and structure to promote policies and practices for the implementation of an open data regime and effective exploitation of “big data”. CODATA will therefore engage with ICSU in advising it of appropriate responses to the data challenge, and bid for ICSU funding to support it in these activities.
12. The international landscape of bodies promoting and implementing open data policies and practices is a busy one (RDA, WDS, ICSTI, DataCite etc). If we are to avoid confusion and maximize the impact of their work, communication, coordination, collaboration and the avoidance of excessive overlap, are priorities. Joint moves to do this are currently under way, as exemplified in agreement between CODATA, RDA and WDS to organize jointly the International Data Week 2016. A potentially powerful and effective step beyond this is the possibility of fusing in some form their different activities in a single overarching body, whilst recognising their individually distinctive roles. CODATA will continue to promote coordination of effort with these and other bodies. Although the CODATA community is a large one, the secretariat resources required to mobilise and orchestrate the efforts described in C & D below are limited. Similar problems apply to the other groups, a disadvantage that coordination of effort can minimize.

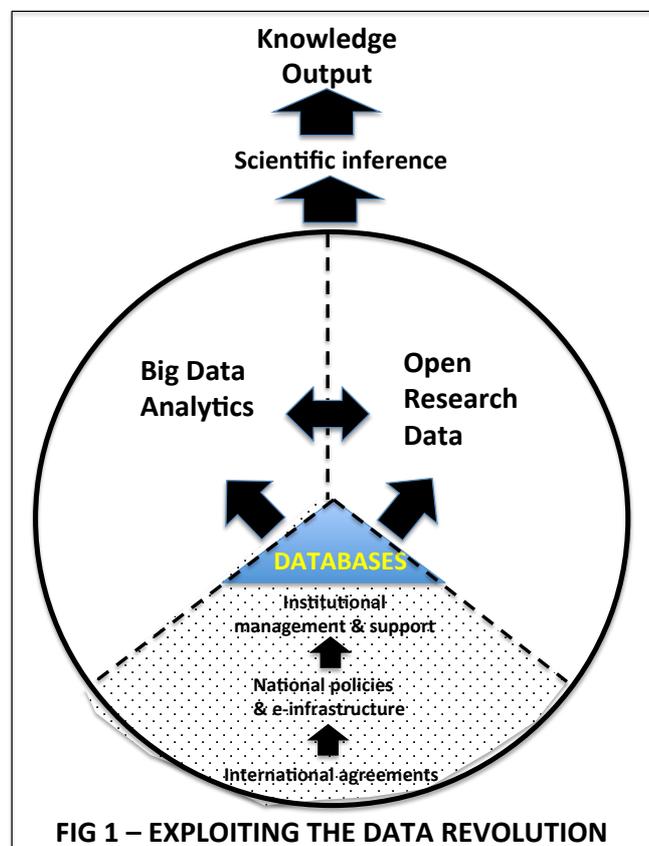
## C: National Data Systems and International Programmes

13. Although science is an intrinsically international enterprise, and although international programmes that address global challenges are increasingly important, researchers work within national systems of organization, funding and priority setting. The extent to which scientists are enabled to exploit the potential of the data revolution depends upon the way in which national and institutional systems support them in doing so. Many national science systems recognize the opportunities in paragraph 4 and are creating strategies to

enhance relevant national capacities. By the same token, international agreements on standards and practices are crucially important: witness the advances made in genomics through community agreements. The orthogonal structure of ICSU and of CODATA, with national members as one dimension and the scientific unions representing international science at disciplinary level as the other dimension, offers, in principle, an opportunity to address data challenges through both national and international dimensions.

14. Figure 1 shows an idealized operational structure for the exploitation of the data revolution. The shaded area represents the support systems for data-intensive research. It comprises:

- *International agreements, practices or standards* codified in the context of subject specific conventions (e.g. genomics), international collaborations (e.g. GEO) or standards (e.g. crystallography). In some research disciplines international agreements, policies or standards have been of foundational importance, yet there is urgent need for greater coordination particularly in addressing the Grand Challenges at the heart of major international programmes.
- *National policies and practices* that fund and incentivise research, and thereby determine researcher behaviour, the physical and software infrastructure that support data-intensive research and the specialist technical groups that develop support systems and tools.
- *Institutional policies and practices* that determine how a research institution will be responsible for the data that its researchers generate, organize management systems for data curation, make those data openly available for re-use and provide the support for researchers in accessing and utilising large and/or complex datasets.



15. The top-right hand segment relates to the ways in which open data can be most effectively utilised by researchers, including data from other sources. The ways in which data-intensive research is undertaken will often depend upon the nature of the research discipline and their modes of collaboration developing in those disciplines (including discipline-specific databases and tools). **Achieving Open Research Data depends upon developing and promoting effective data policies and practices, supported by data education and capacity building.**
16. The top-left hand segment addresses the major analytical, often problematic issues that arise in dealing with big data, and which are the subject of very active research (see paragraphs 31-34). Leadership in developing the “big data” concept and exploiting it in practice has not primarily lain with the publicly funded research community but through the work of the ICT industry, private consultancies and the commercial companies that implement their systems in business practice. It is not only crucial that the publicly funded science community learns about process from the private sector, but also seeks to collaborate with it in exploiting the potential of open data across the public/private interface. **Improving Big Data Analytics, depends on addressing major challenges at the frontiers of data science.**
17. The rationale for the open data regime summarized in figure 1 is to maximize scientific understanding (figure 1 – top) from the vast and increasing data volumes that technology makes available. A current priority in this regard is to ensure that the large international programmes that address global challenges develop **effective and holistic data strategies**. Underpinning strategies, must be based on: a) the development and implementation of effective data policies and the promotion of good practices; b) measures to address the key issue of making valid inferences from linked data) and, c) the promotion and acquisition of up-to-date data skills.
18. The imperatives described above, and the need to promote efficient and effective data strategies, determine CODATA’s own priorities and the activities through which they will be realized to the end of the current planning period in 2018.

## D: CODATA’s Strategic Response

19. The following three sections outline CODATA’s strategic priorities in contributing to the developments needed if the international science community and its national components are to effectively exploit the data revolution. CODATA’s strategy has three components: data policy, data science and data education:
  - 1) Data Policy: Supporting implementation of **data principles and practices**
  - 2) Data Science: Addressing the **frontiers of data science**
  - 3) Data Education: **Capacity building** (particularly in low and middle income countries - LMICs)

Each priority is expressed in terms of their: a) purpose and rationale, b) planned outputs (things we can control) and c) hoped for outcomes and impacts (things that outputs are designed to influence/create).

## E: Priority 1. Data Principles and Practices

### Purpose

20. A major purpose of CODATA is both to advocate the vision for data intensive science as set out in section A and to promote implementation of the policies and practices that will realise it in national science systems. Consequently, in this priority area, CODATA will focus its efforts on three targets:

- Options for the **national policies and actions** required to support and incentivise best practice for research data.
- **Institutional options** (universities and institutes) for the management and support of good practice in exploiting data resources, and in matching their role to an international data ecology.
- Best practice options for **international programmes or large disciplinary/interdisciplinary consortia** (particularly in major global programmes such as *Future Earth*, *Integrated Research on Disaster Risk*, *Global Health* or ‘*Smart Cities*’).

### Outputs

#### a) An assessment tool

21. We will utilise existing work of CODATA together with that of its partners to produce a register of good practice based on the successful examples that have been developed by collaborating bodies in pursuit of their own goals. This will be complemented by **assessment tools** designed to evaluate the extent to which best practice is achieved, and directed to the specific needs of national authorities, scientific institutions (universities and institutes) and major data-intensive programmes. They will set out how current best practice should be applied in each category and provide a check-list for appropriate policies and best practice, thereby providing a basis for self-evaluation. The specifications for this task are being undertaken by the CODATA Data Policy Committee, and should be completed by July 2015. Funding is being sought for a researcher to undertake detailed work on the project, overseen by the Data Policy Committee. A draft available for consultation should be ready by early 2016 for delivery later in that year.
22. The register and tool will be perennially up-dated and will recognize that the details of implementation will vary between different national jurisdictions and different areas of science. This work will be sensitive to and linked with the interests of the two orthogonal components of ICSU membership, its national members and its member unions (as noted above in paras 13 and 14).
23. The approach above not only exploits work that CODATA has done, but also relies heavily on links with other expert bodies, in which CODATA “harvests”, integrates and disseminates best practice approaches to data issues. This will also provide a resource for Priority 3 on capacity building (Section G, below).

#### b) The data agenda for international science

24. The Data Policy Committee will also produce a short document presenting and exemplifying the data agenda for international science based on the principles and

priorities set out in the present strategy document. This document is in preparation and should be available for circulation by autumn 2015. It is designed to address a series of specific communities (researchers, research institutions, funders of research, those responsible for national policies, and publishers) through a distinctive and creative format that sets out the priorities and benefits of an open data regime for each group.

25. Associated with this, there are a number of unresolved, problematic issues that are highly relevant to policies and practice and that will continue to be addressed by CODATA, in conjunction with other groups. They currently include: data privacy; intellectual property; responsibilities of data owners; public access to and public usability of data; approaches to the retention or disposal of data; data citation standards and practices.

### c) Other planned/proposed activities - 2015/16

26. *Workshop on Data Policy Development.* CODATA has proposed to the ICSU Europe secretariat an ICSU Europe-CODATA Workshop on Data Policy Development and Implementation. Its purpose is to engage ICSU Europe members better with the data policy agenda and issues around the implementation of ICSU's statement on *Open Access to Scientific Data* and policies being piloted in the EU Horizon 2020 Programme. The proposed date for the workshop is late November/early December 2015. This will also feed into output a).
27. *Data Task Force for Future Earth.* Since 2012, CODATA has worked with WDS to ensure that data issues are taken into consideration in the planning of and pursuit of the science objectives of *Future Earth*. The science objectives of *Future Earth* will be compromised unless they are supported by effective policies and practices for data and its integration. A result of this effort has been the creation of a Future Earth Data Task Force, to which CODATA will contribute fully.
28. *Regional Workshops on Data Strategies for Interdisciplinary and International Research into Smart Cities.* It is planned to support collaboration between CODATA members on a series of workshops on the above topic. These will concentrate on data policies, data management, data integration, the place of data in scientific reasoning and capacity building for these issues. Research around 'Smart Cities' is of great interest internationally and will be of importance for the Future Earth Programme. A concept note has been prepared and there is initial interest from CODATA contacts in India, Taiwan and Australia. The scope for workshops can be extended rapidly to include China, South Africa, Mexico, Brazil, Indonesia, Malaysia, Japan and others.
29. *CODATA Task and Work Groups.* There are several Groups whose work informs CODATA's policies and practices objective, and for which funding was agreed at the 2014 General Assembly. These are: Data Citation Standards and Practices; Interoperable Data Publications; Earth and Space Science Data Interoperability; Linked Open Data for Disaster Research; Science and the Management of Physical Objects in the Digital World; Data at Risk. As an example of the planned outputs of Task Groups, the Task Group on Data Citation Standards and Practices plans to hold 11 regional workshops to promote the implementation of good data citation practice by the time of the 2016 General Assembly. They will take place in China, Taiwan, Japan, India, Australia, South Africa, Israel, Brazil, UK, France and the USA. Additionally, there is a joint CODATA-RDA Working Group on Legal Interoperability of Research Data. CODATA is taking a leading role in the GEO Data Sharing Working Group and the GEO Data Management Principles Task Force.

## Outcomes

30. The outcomes from this component of the CODATA strategy are designed to help ensure that national and institutional policies and practices for data-intensive science reflect the best contemporary practice and that such practice is also routinely embedded in the planning and creation of major science programmes, at both international and national levels. A role for CODATA at the end of the planning period (2018) is to assess the extent to which these outcomes have been achieved globally.

## F: Priority 2 – Frontiers of Data Science

### Rationale

31. We distinguish two categories of issue at the frontiers of data science. The first is defined by fundamentals of data analytics and the need to ensure valid reasoning from data. The second relates to the creative development of tools and protocols that support the processes of intelligent openness and research data management.
32. Managing much data frequently poses non-trivial problems of organizing, visualizing, summarizing and navigating data collections, particularly where they are to be integrated from a wide variety of sources, for example in the increasing need to integrate data from the natural and social sciences. The outputs of machine analyses of data are frequently difficult to make available in forms that are easily accessible to human cognitive processes, and require “intelligent agents” and machine learning technologies to support data visualization. Data interoperability is a challenge that increases as more and more disparate data become available and as applications require data from more diverse disciplines.
33. Managing and making data available in a comprehensible and interoperable form is only one part of the fundamental agenda of data science, the complementary part being to ensure that it is used in a rigorous and valid fashion. It has been widely recognised that many of the statistical tools that have been routinely used in the past to analyse relatively small datasets are not valid when applied to the large, diverse, linked datasets that characterise the world of “Big Data”. The late Jim Gray commented: “when you see what scientists are doing, day in and day out, in data analysis, it is truly dreadful.” Major problems include understanding scale, sparse sensing, sampling, quantifying uncertainties, next generation semantic tools etc. These are related to fundamental priorities for data science of understanding the deep mathematical structures inherent in “Big Data”, and developing appropriate statistical tools for valid reasoning.
34. Leadership in developing the “Big Data” concept and exploiting it in practice has not primarily lain with the publicly funded research community but in the work of the ICT industry, private consultancies and the commercial companies that implement their systems in business practice. It is not only crucial that the publicly funded science community learns about process from the private sector, but also seeks to collaborate with it in exploiting the potential of open data across the public/private interface, of particular relevance to the social sciences. CODATA has begun discussions with senior private sector scientists about the potential to enhance this collaboration.

## Objectives

35. A major ongoing objective for CODATA is to stimulate work on these issues, to be an agent for broadcasting results, solutions and best practice, and to embed these outcomes in the practices dealt with in the assessment tools described in paragraphs 21-23.

## Outputs

### a) International Data Week 2016

36. CODATA, in collaboration with the RDA and WDS, plans an International Data Week in September 2016, to be held in North America (USA or Canada). A central component of this will be a major international conference on the “Frontiers of Data Science”. This is a development of CODATA’s biennial SciDataCon meetings, the last one of which, held in Delhi in 2014, was organised jointly with WDS. The primary purpose of the conference is to address the issues described in paragraphs 31-34. It is hoped to involve the private sector in International Data Week 2016.

### b) The Data Science Journal

37. An open access, electronic journal has been produced by CODATA since 2001. The Journal is being re-launched in collaboration with a dynamic, researcher-led Open Access publisher, Ubiquity Press, with a new Editor-in-Chief, Dr Sarah Callaghan, and a new Editorial Board, with the ambition to become the leading journal for data science. A major part of the re-launched journal will focus on the frontiers of data science, with special issues dedicated to major themes, included papers from the 2016 “Frontiers” conference. It will also be a means of communicating access to materials relevant to the other two strategic priorities of CODATA. Both the RDA and WDS have agreed to advocate use of the Journal to their communities.

### c) Uniform Description System for Nanomaterials

38. CODATA received funding from ICSU for inter-Union work to develop a Uniform Description System for Nanomaterials. Version 1.0 of the UDS has been published for consultation and a final workshop will be organized in Summer 2015. The UDS is being further tested and validated in the context of a European FP7 Project, Further Nano Needs (FNN) and will be the focus of a major conference to be convened by the Project and CODATA in Paris in early 2016. The FNN Project will conclude in January 2018.

### d) Inter-Union Work on Data Standards and Ontologies

39. Building on the model used in the Nanomaterials work, CODATA is planning to convene an inter-Union activity on the development of ontologies and data standards across disciplines. A number of avenues to pursue this are available, including: a) reviving a previous proposal for inter-Union activity on ontologies; and, b) pursuing collaboration with IUBS and other bio-Unions on Integrative Biology. Any such work still needs to be scoped in further detail and may be included in the more strategic call for Task Groups described below.

### e) Integrating Geospatial Data on the Web

40. A proposal has been submitted to the British Embassy in Beijing to fund a pilot series of two symposia to bring together W3C, the standards body for web technologies, and OGC, the standards body for geospatial data, with the purpose of developing standards for geospatial data on the web. Current web systems permit a geospatial query of the type: “here is a location, what are its properties”. The purpose of the proposed work is to enable a question of the type: “give me a location(s) that has the following properties”, which has obvious and powerful potential in Earth observation and many other applications. The symposia will bring the two technical groups together with Earth observation experts to create precise specifications for the software that will be required to implement the function. Further external funding will be sought for this second stage. Engaging key international stakeholders, including GEO, the Group on Earth Observations, will play an important role in setting the world standard and the future direction for sharing Earth Observation data on the Web. If funded the first project will run for twelve months from April 2015.

### f) Big Data and Data Integration for International Science

41. In June 2014, CODATA convened a successful Symposium on Big Data for International Scientific Programmes which developed a set of recommendations and tasks for an International Working Group. This activity will be further scoped and taken forward. It might align with the work on Data Strategies for Smart Cities described above, be applicable to other major programmes such as Future Earth and contribute to the Assessment Tool for such programmes described in paragraphs 21-23.

### Outcomes

42. The outcomes that the above activities are designed firstly to stimulate the creation of research results and solutions that will support researchers in discovering, aggregating, and appropriately analyzing data, and also making their own data intelligently open and cited so that it can be used by others. Secondly, it is to ensure that these solutions are made available in useable form to researchers. An assessment of the extent to which this has been achieved could be part of the post-2018 review referred to in paragraph 30.



*Participants at the CODATA Workshop on Big Data for International Scientific Programmes, 8-9 June 2014, Beijing, China.*



*Participants at the CODATA Workshop on Open Data for Science and Sustainability in Developing Countries, Nairobi, 4-8 August 2014*

## G: Priority 3 – Capacity Building

### Rationale

43. As the importance of open data is increasingly recognized in national science systems, policies are being developed to provide the physical and organizational infrastructure, to develop competence in data science and to support the practices that are required to exploit data effectively and efficiently.
44. CODATA is committed to supporting these processes where it can, with a particular focus on deploying its community to support more effective data access and use in Low and Middle Income Countries (LMICs) where the science base is less well resourced. If the benefits of open data are as set out in paragraphs 3-6, it is crucial that these countries are assisted in adapting their processes to capitalize on these benefits rather than permitting another global “knowledge divide” to develop.

### Outputs

#### a) Data training workshops

45. Through its Task Group on Preservation and Access to Scientific Data in Developing Countries (PASTD) CODATA has run a yearly series of workshops since 2002. In 2014, ramping up this activity with support and direction from the Secretariat, CODATA organised a Data Science Training Workshop in Beijing in collaboration with CAS, a Policy Workshop in Kenya in 2014 involving a wide range of international participants and tutors, and collaborated with the Indian Statistical Institute to convene a further Training

Workshop in Bangalore in March 2015. As an example of the sort of supporting processes that will be required, we are exploring the possibility of creating a certification process for those attending workshops and courses, as an important means of justifying the expense of attendance.

### **b) An initiative in Sub-Saharan Africa**

46. The strand of work in a) above is now being developed in a more strategic way to systematically target capacity building in data science and policy in LMICs. Our first target is to collaborate with academies and analogous bodies in sub-Saharan Africa and collaboration with ICSU in systematic, regional capacity building through the creation of a consortium of African national bodies. We are discussing with South African institutions the possibility that they might play a leadership role, with international bodies helping by contributing expertise (RDA, WDS, GEO and AfriGEOSS & others). The first steps in this will be discussion with stakeholders in RSA about how to build on the momentum developing behind the data needs of the Square Kilometre Array. These initiatives are currently in development and will be further refined through a visit of the President and Director to South Africa in June 2015.

### **c) A generic approach to LMICs**

47. Depending on the progress of the above initiative, we contemplate developing similar programmes elsewhere, particularly in South America and South and South-East Asia, in association with regional offices of ICSU. CODATA is collaborating with the RDA on a co-branded Working Group to develop a framework for Data Science Summer Schools. This includes development of a curriculum framework and certification process. A funding proposal has been submitted to The World Academy of Sciences and the International Centre for Theoretical Physics for a pilot school to be held in Trieste in Summer 2016. Again, the considerable data needs of the SKA will provide both an initial focus and a springboard for further activity.
48. The PASTD in Developing Countries Task Group will continue to be an important means by which CODATA pursues its capacity building agenda. Having developed a set of high level principles on Open Scientific Data for LMI Countries, the groups next task is to produce a set of implementation guidelines. The Group is considering collocation of a workshop with the GEO Plenary in Mexico City in November 2015 as an efficient means of taking this work forward.

### **Outcomes**

49. The purpose of this initiative is to catalyse and to collaborate with others in a process that will ensure that exploitation of the data revolution will not only benefit countries with a well-funded science base but also that they will be more widely distributed internationally through the development of national capabilities in those countries where investment in the science base is proportionately less. The extent to which this process is succeeding will be reviewed after 2018, the end of the current planning period.

## H: Roles of Committees, Task Groups and Work Groups

### Executive Committee

50. The CODATA Executive Committee (EC) is elected by the CODATA General Assembly and is responsible for planning strategy, agreeing the broad lines of its implementation, setting budgets and determining grant allocations.

### Other Strategic Committees

51. There are also committees working to promote the strategic objectives of CODATA and working with the Executive Committee, President and Director in ensuring that their tasks are undertaken in a creative and timely manner. The Data Policy Committee has been tasked to promote the *Data Principles and Practice* priority. Two new Committees are being created: a “Data Frontiers Committee” to promote the *Frontiers of Data Science* priority, and a Capacities Committee, which will also subsume the Preservation and Access to Data in Developing Countries Task Group, to promote the Capacity Building priority. Each committee is/will be chaired by a member of the Executive Committee.
52. A further committee, the Committee on Fundamental Physical Constants, maintains a long-term responsibility of CODATA in providing the scientific and technological communities with a self-consistent set of internationally recommended values of the basic constants and conversion factors in physics and chemistry based on current relevant data.

### Task Groups

53. Task groups address major data needs and policy issues. Proposals for projects are received from the scientific community and decisions whether to approve are made by the General Assembly. The CODATA Executive Committee may then assign funding as a means of leveraging greater resources from elsewhere. The projects agreed at the 2014 GA, with funds allocated at the March 2015 Executive Committee, are: Advanced Informatics for Microbiology; Anthropometry for Special Populations; Data at Risk; Data Citation Standards and Practices; Interoperable Data Publications; Earth and Space Science Data Interoperability; Global Roads Data Development; Linked Open Data for Disaster Research; Preservation and Access to Data in Developing Countries; Science and the Management of Physical Objects in the Digital Era.
54. Although these projects are generally successful in leveraging funds from elsewhere, the EC decided that it would focus funding for Task Groups in areas of its three strategic priorities, whilst maintaining a bottom-up, community bidding process. Future calls for proposals will therefore be selected on the basis of the strategic priorities, and bids for funding will be assessed using a prescribed set of criteria against which they will be ranked for funding.

### Working Groups

55. CODATA Working Groups are generally established to investigate short-term problems and to explore the need for CODATA action on specific issues. CODATA also works in collaboration with other organisations that make substantial contributions to such activities. Current work groups are: Early Career Data Professionals; Description of Nanomaterials; Legal Interoperability of Research Data (jointly with the RDA); Data Science Summer Schools (jointly with the RDA). CODATA also contributes to the Working Groups of other bodies: Data Sharing (GEO); Data Management Principles (GEO); Cost Recovery for Data Centres (RDA, WDS).

## I: Budget and Business Planning

56. The Treasurer's report for 2014 shows a balanced budget of €252k largely derived from members fees. However, this sum represents a core investment that is used to lever much greater value. It is estimated that activities to a value of about €1.3M were levered in 2014 alone, a leverage ratio of roughly 5:1. This figure is derived from 1) the additional sums obtained by Task and Working Groups; 2) the full value of events convened; and, travel and other support received by the Executive Director. It does not include a valuation of the in-kind support from the CODATA Executive Committee and Officers. As a specific example, in March 2015, CODATA convened a Training Workshop on Big Data with the collaboration of the Indian Statistical Institute. CODATA's own investment in the event totals c.€2.5K in travel and student support. The event as a whole leveraged an additional c.€55,550 in support, comprising international and local travel and accommodation for experts and students as well as sponsorship and local expenses.
57. The Executive Committee has committed CODATA to the ambitious strategy set out in section C. But it recognizes that notwithstanding the size and energy of the CODATA community, hitting these strategic targets will depend on the capacity of the officers and secretariat to plan and orchestrate the efforts of the community. The Committee has therefore agreed to make available additional resources from its reserve starting mid-2015 to expand secretariat support. Availability of additional secretariat support will permit the Executive Director to focus energy on accelerating strategic activities necessary to enhance CODATA's international profile, attract membership, and ultimately increase revenue. This investment in enhanced activity will also require a 3-year business plan for increasing income, with the target of a balanced budget by the end of the current planning period in 2018. Movement towards break-even will be monitored through intermediate targets for 2016 and 2017 as show in the table below. A business plan designed to plot the course towards break-even is currently being developed.
58. The Director and Treasurer will develop a business plan for increased income from on three potential sources:
- increased membership, including the possibility of business membership;
  - a proposal that a joint bid be made together with ICSU for support for delivering an ICSU open data strategy;
  - external funding bodies whose objectives converge with CODATA strategies, such that funding from them would substitute for sums currently committed from CODATA resources. (For example, capacity building in the "global south" (paras 43-49) is a priority for many governments, charities and trusts, CODATA has the capacity to act as a delivery agent for some of their objectives).
59. The current formulation of the budget shows the categories of direct expenditure. For strategic planning and for the external communication of value, it is important to have two additional formulations of the budget:
- a) to show the extent to which CODATA expenditure matches its strategic priorities by categorising expenditure by purpose.
  - b) to show the extent to which CODATA funding levers additional resources to support its science objectives.

## J: Membership

60. Increasing national membership is an important priority for CODATA for two reasons: to increase the impact and take up of its work, and to contribute to increased income. A paper should therefore be prepared, largely from material contained in this document, designed to make CODATA's value proposition to potential members.
61. The Director should develop a plan for the way in which approaches to new member countries should be made, exploiting the good offices of Executive Committee members where appropriate.

## K: Outreach, Communications and Reporting

62. The strategy outlined in this document is important and ambitious. If it is to have the impact that CODATA aspires for, CODATA needs to be better known to stakeholders by communicating evidence of its activities and demonstrating their importance. Such communication should be through:
  - further development of the website;
  - annual reporting to members, to Union members of ICSU and to national members of ICSU and to other parties interested in CODATA activities (in the form of a revived newsletter);
  - increasing synergy between CODATA national groups (from which we should seek annual reports) and CODATA "international" and the development of an overarching strategy that integrates both the international and national dimensions of CODATA work to a greater degree;
  - annual reporting from Task Groups and Work Groups;
  - working with ICSU union members promoting open data strategies in their disciplines;
  - increased participation by CODATA in events planned by others as a means of promoting CODATA priorities. Although this should be coordinated by the Secretariat, it must be the mission of all CODATA member organizations and individuals.
63. Financial and non-financial performance measurement will be used to ensure that CODATA is accountable to its members and stakeholders and as a tool to facilitate learning and improve performance. Accounting and information feedback systems will allow calculation of financial metrics such as donor dependency ratio, income utilization ratio, survival ratio, and financial leverage which will provide clarity in the usage of funds and the financial health of CODATA. Measures of non-financial performance will be developed to provide a balanced view of performance and support continuous improvement. Measures will be used to assess progress towards strategic objectives through task group, work group and other activities; to track the quality of relationships with the ICSU community, CODATA members, and partners; and to identify key areas for improvement.