Open Science Conference
Back-to-back Belgium-European Commission
Palace of the Academies, Brussels, Belgium
21 November 2018

How can we build capacity for an Open Science and FAIR Data ecosystem?

Simon Hodson, Executive Director, CODATA
www.codata.org

International Science Council

Revision of the SI Units and the CODATA Fundamental Physical Constants

- Major revision of the SI Units agreed on 16 November 2018.
- The kilogram, ampere, kelvin and mole will now be based, respectively on the Planck constant h, the elementary charge e, the Boltzmann constant k, and the Avogadro constant NA.
- See http://bit.ly/codata-fundamental-constants and http://iopscience.iop.org/article/10.1088/1681-7575/aa950a/pdf

CODATA Prospectus:

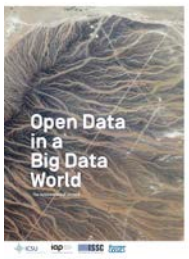
https://doi.org/10.5281/zenodo.1167846

Principles, Policies and Practice

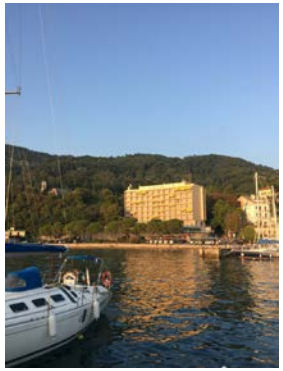
DC1 Data Citation Principles

INTERNATIONAL COUNCIL FOR SCIENCE

OECD



Capacity Building

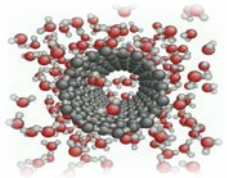


UNESCO

GOFAIR

The Abdus Salam International Centre for Theoretical Physics

Frontiers of Data Science



ubiquity press open scholarship

Data Science Journal

CODATA 2017, Saint Petersburg 8-13 Oct 2017

ICSU WORLD DATA SYSTEM

RDA

ASSAf

Major CODATA Initiatives

- Data Policy Committee: http://www.codata.org/strategic-initiatives/international-data-policy-committee
- Data Interoperability and Integration for Interdisciplinary Research: http://dataintegration.codata.org
- African Open Science Platform: http://www.codata.org/strategic-initiatives/african-open-science and https://doi.org/10.5281/zenodo.1407488
- CODATA-RDA Schools of Research Data Science: http://www.codata.org/working-groups/research-data-science-summer-schools

CODATA RDA
School of Research Data Science
https://vimeo.com/299263596

CODATA-RDA School of Research Data Science

Trieste, 2018

Convenors and Organisers

Partners

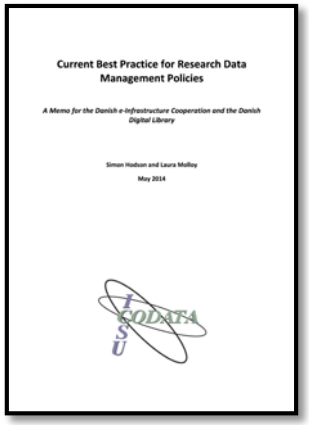
05:28

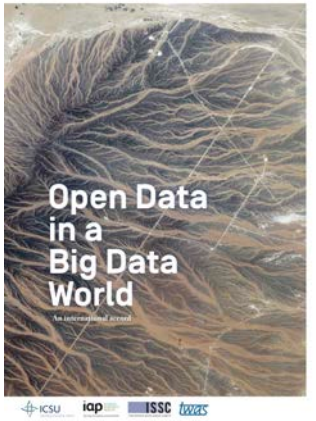
New CODATA
Executive Committee and President

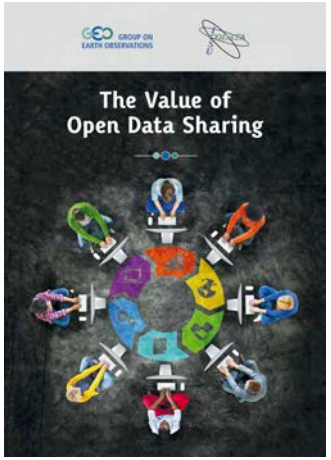
- CODATA General Assembly Nov 2018, new CODATA President is Barend Mons.
- Replaces Geoffrey Boulton.
- New Executive Committee enhances expertise and retains global geographic reach.
- Refocussing of CODATA activity.

CODATA Data Policy Reports Since 2016

- Current Best Practice for Research Data Management Policies (for Danish e-Infrastructure Cooperation, 2015) http://dx.doi.org/10.5281/zenodo.27872
- The Science International Accord on Open Data in a Big Data World (for ICSU, 2015) http://www.science-international.org/#accord
- The Value of Open Data Sharing (for GEO, 2015) http://dx.doi.org/10.5281/zenodo.33830
- Guidelines for the Legal Interoperability of Research Data (with RDA, 2016) https://doi.org/10.5281/zenodo.162241
- Business models for sustainable research data repositories (with OECD, 2017) https://doi.org/10.1787/302b12bb-en







Why Open Science / FAIR Data?

- **Good scientific practice depends on communicating the evidence.**
  - Open research data are essential for reproducibility, self-correction.
  - Academic publishing has not kept up with age of digital data.
  - Danger of an replication / evidence / credibility gap.
  - Boulton: to fail to communicate the data that supports scientific assertions is malpractice
- **Open data practices have transformed certain areas of research.**
  - Genomics and related biomedical sciences; crystallography; astronomy; areas of earth systems science; various disciplines using remote sensing data…
  - **FAIR data helps use of data at scale, by machines, harnessing technological potential.**
  - Research data often have considerable potential for reuse, reinterpretation, use in different studies.
- **Open data foster innovation and accelerate scientific discovery through reuse of data within and outside the academic system.**
  - Research data produced by publicly funded research are a public asset.

Policy Push for Open Research Data

- The three Bs (Budapest, Berlin and Bethesda) and Open Access, 2002-3
- OECD Principles and Guidelines on Access to Research Data, 2004, 2007
- UK Funder Data Policies, from 2001, but accelerates from 2009
- NSF Data Management Plan Requirements, 2010
- Royal Society Report 'Science as an Open Enterprise', 2012
- OSTP Memo 'Increasing Access to the Results of Federally Funded Scientific Research', Feb 2013
- G8 Science Ministers Statement, June 2013
- G8 Open Data Charter and Technical Appendix, June 2013
- EC H2020 Open Data Policy Pilot, 2014; Adoption of FAIR Data Principles, 2017.
- Science International Accord on Open Data in a Big Data World, Dec 2015: http://bit.ly/opendata-bigdata

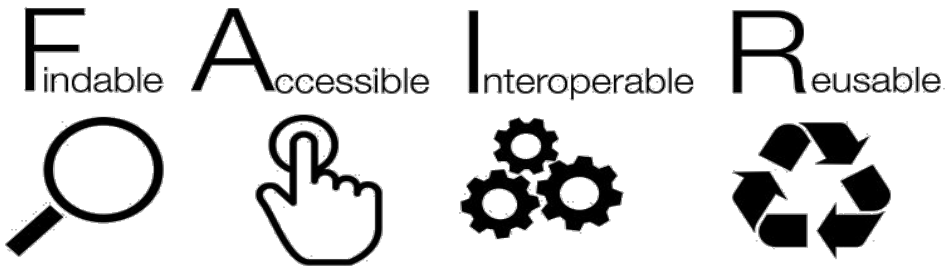
Attributes that give value to research data

- Builds on previous definitions…
- OECD Statement of Principles and Guidelines for Access Research Data: include a number of principles including accessibility, interoperability, quality, legal transparency, sustainability…
- Royal Society, 2012, *Science as an Open Enterprise*, Intelligent Openness: **accessible, intelligible, assessable, usable**.
- G8 Science Ministers' Statement, 2013, 'Open scientific research data should be easily **discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards**.'

- FAIR Data now at the heart of H2020 policy, European Open Science Cloud etc.
  - **Under the revised version of the 2017 work programme, the Open Research Data pilot has been extended to cover all the thematic areas of Horizon 2020.**
- Current EC Guidance at http://bit.ly/EC_H2020_RDM_Guidance and http://bit.ly/EC_H2020_OpenData_Infographic

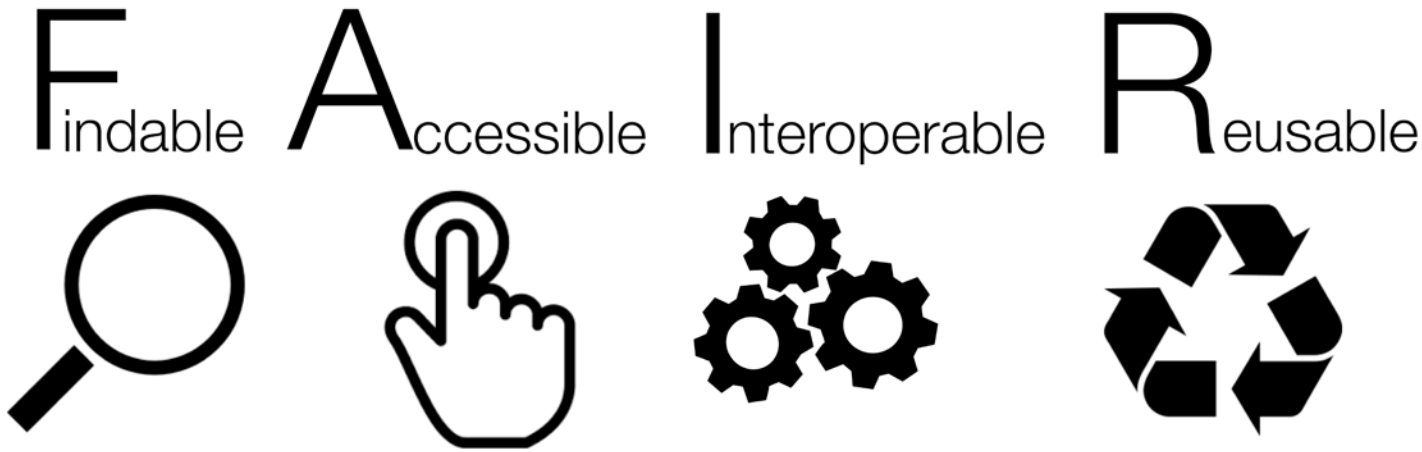
What is FAIR?

A set of principles that describe the attributes data need to have to enable and enhance reuse, by humans and machines



Emerging Policy Consensus? FAIR Data

- **FAIR Data** (see original guiding principles at https://www.force11.org/node/6062
  - **Findable:** have sufficiently rich metadata and a unique and persistent identifier.
  - **Accessible:** retrievable by humans and machines through a standard protocol; open and free by default; authentication and authorization where necessary.
  - **Interoperable:** metadata use a 'formal, accessible, shared, and broadly applicable language for knowledge representation'.
  - **Reusable:** metadata provide rich and accurate information; clear usage license; detailed provenance.



Turning FAIR into Reality

- **Report and Action Plan: Takes a holistic** approach to lay out what needs to be done to make FAIR a reality, **in general and for EOSC**
- **Addresses the following key areas:** Concepts for FAIR, the FAIR data culture, the FAIR data ecosystem, skills, incentives and metrics, investment and sustainability.
- **Recommendations and Actions:** 27 clear recommendations, structured by these topics, are supported by precise actions for stakeholders.
- **Turning FAIR into Reality:** Report and Action Plan: **https://doi.org/10.2777/1524**

FAIR Data EG: membership

Simon Hodson, CODATA
Chair of FAIR Data EG

Rūta Petrauskaité, Vytautas Magnus University

Peter Wittenburg, Max Planck Computing & Data Facility

Sarah Jones, Digital Curation Centre (DCC), Rapporteur

Daniel Mietchen, Data Science Institute, University of Virginia

Françoise Genova, Observatoire Astronomique de Strasbourg

Leif Laaksonen, CSC-IT Centre for Science

Natalie Harrower, Digital Repository of Ireland – year 2 only

Sandra Collins, National Library of Ireland – year 1 only

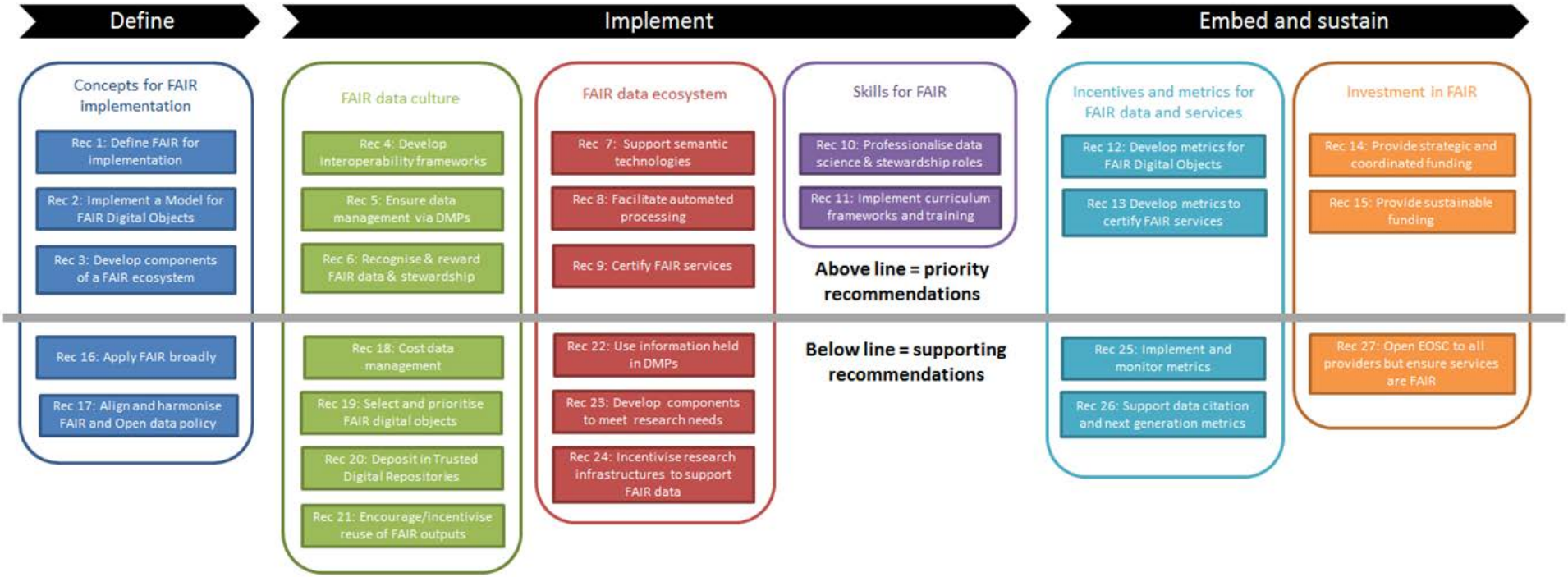
European Commission Expert Group on FAIR Data: Objectives

1. **To develop recommendations on what needs to be done to turn the FAIR data principles into reality (EC, member states, international level).**
2. **Develop the FAIR Data Action Plan, by proposing a list of concrete actions as part of its Final Report.**
3. Extensive consultation on report framework and on interim report published early June 2018.
4. **Launch and disseminate FAIR Data Action Plan and support Commission communication in November 2018**

http://tinyurl.com/FAIR-EG

Structure of the Report and Action Plan

1. Concepts: why FAIR?
2. Creating a culture of FAIR data
3. Creating a technical ecosystem for FAIR data
4. Skills and capacity building
5. Measuring change
6. Funding and sustaining FAIR data
7. FAIR action plan


Define
Implement
Embed and sustain
Concepts for FAIR implementation
Rec 1: Define FAIR for implementation
Rec 2: Implement a Model for FAIR Digital Objects
Rec 3: Develop components of a FAIR ecosystem
Rec 16: Apply FAIR broadly
Rec 17: Align and harmonise FAIR and Open data policy
FAIR data culture
Rec 4: Develop Interoperability frameworks
Rec 5: Ensure data management via DMPs
Rec 6: Recognise & reward FAIR data & stewardship
Rec 18: Cost data management
Rec 19: Select and prioritise FAIR digital objects
Rec 20: Deposit in Trusted Digital Repositories
Rec 21: Encourage/incentivise reuse of FAIR outputs
FAIR data ecosystem
Rec 7: Support semantic technologies
Rec 8: Facilitate automated processing
Rec 9: Certify FAIR services
Rec 22: Use information held in DMPs
Rec 23: Develop components to meet research needs
Rec 24: Incentivise research infrastructures to support FAIR data
Skills for FAIR
Rec 10: Professionalise data science & stewardship roles
Rec 11: Implement curriculum frameworks and training
Above line = priority recommendations
Below line = supporting recommendations
Incentives and metrics for FAIR data and services
Rec 12: Develop metrics for FAIR Digital Objects
Rec 13 Develop metrics to certify FAIR services
Rec 25: Implement and monitor metrics
Rec 26: Support data citation and next generation metrics
Investment in FAIR
Rec 14: Provide strategic and coordinated funding
Rec 15: Provide sustainable funding
Rec 27: Open EOSC to all providers but ensure services are FAIR

Key Concepts and Messages

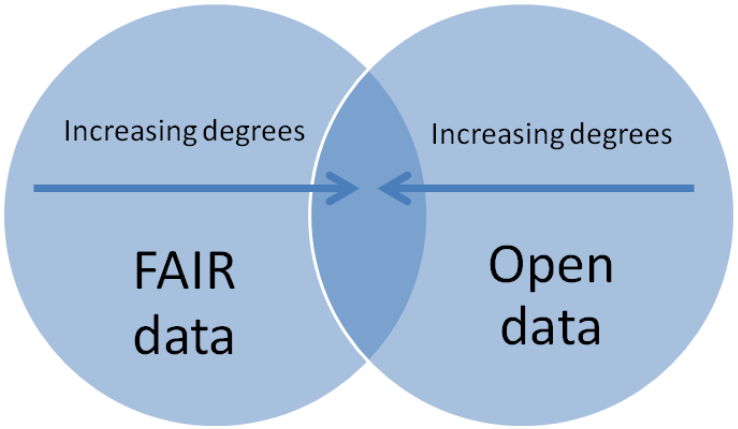
1. FAIR and Open and other supporting policies
2. Model for FAIR Digital Objects
3. FAIR ecosystem
4. Interoperability frameworks
5. Skills
6. Metrics and Incentives
7. Investment and Sustainability

FAIR and Open

- Concepts of FAIR and Open should not be conflated. Data can be FAIR or Open, both or neither
- The greatest potential reuse and value comes when data are both FAIR and Open
- Even internal or restricted data will benefit from being FAIR, and there are legitimate reasons for restriction which vary by discipline
- *Align and harmonise FAIR and Open data policy* to ensure that publicly-funded research data are made FAIR and Open, except for legitimate restrictions

FAIR and Open

- **‘As Open as possible, as closed as necessary'**
- By default, data created by publicly funded research projects, initiatives and infrastructures should be to made available as soon as possible.
- Policies could allow for (short) embargo periods to facilitate the right of first use for creators
- Guidance should be provided to researchers to help find optimal balance between sharing and privacy



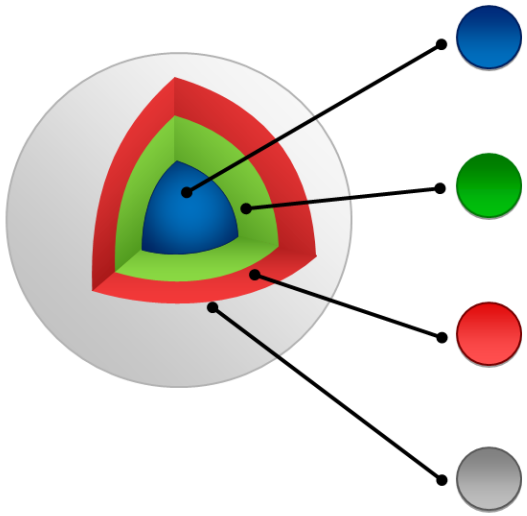
Concepts Implied by the Principles

Making FAIR a reality depends on additional concepts that are implied by the principles, including:

- The timeliness of sharing
- Data appraisal and selection
- Long-term preservation and stewardship
- Assessability – to assess quality, accuracy, reliability
- Legal interoperability – licenses, automated

FAIR Digital Objects

- Implementing FAIR requires a model for FAIR digital objects
- Digital objects can include data, software, and other research resources
- Universal use of appropriate PIDs
- Use of common (ideally open) formats; data accompanied by code
- Rich metadata and clear licensing enables broadest reuse



DIGITAL OBJECT
Data, code and other research resources
At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS
Persistent and unique identifiers (PIDs)
Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

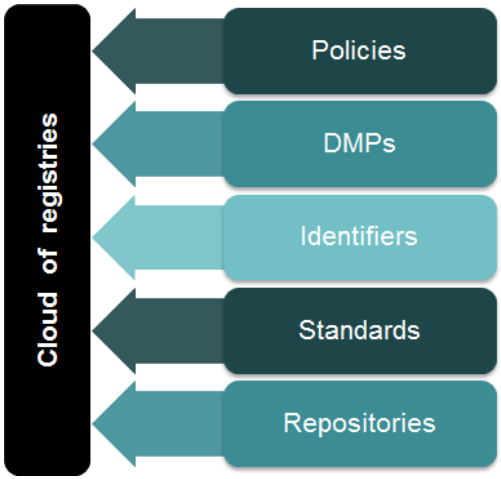
STANDARDS & CODE
Open, documented formats
Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA
Contextual documentation
In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

The FAIR Ecosystem

- The realisation of FAIR relies on an ecosystem of components
- Essential are:

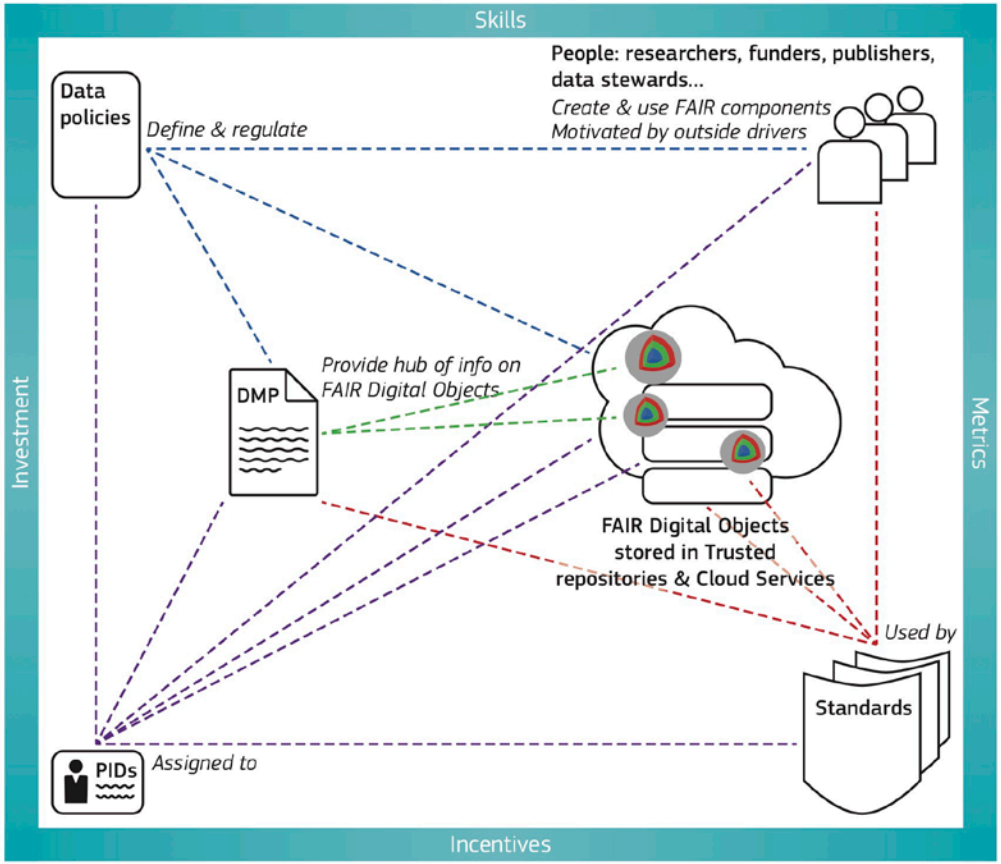
  Policies
  Data Management Plans
  Identifiers
  Standards
  Repositories



- Registries to catalogue each component of the ecosystem, and automated workflows between them.
- Begin by enhancing existing registries; build those for DMPs and IDs

The FAIR Ecosystem

- Ecosystem and its components should work for humans and machines
- Need to clearly define infrastructure components essential in specific contexts and fields
- Testbeds need to be used to evaluate, evolve, innovate the ecosystem



Interoperability Frameworks

Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research

Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure.

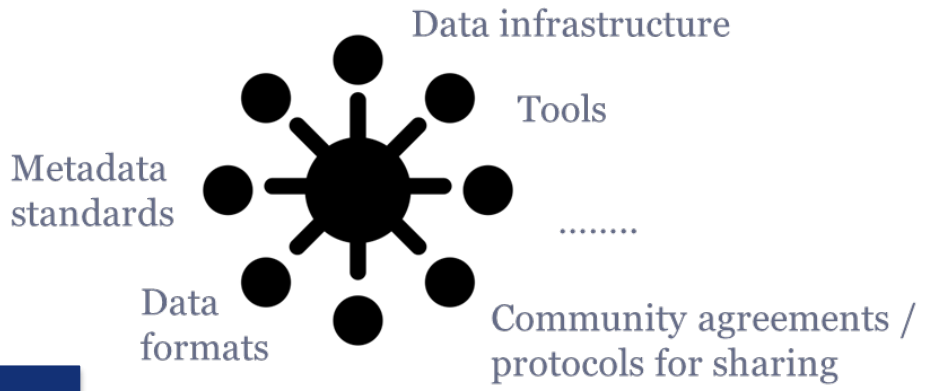
To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant. Intelligent crosswalks, brokering mechanisms and semantic technologies should all be explored to break down silos.

- Semantic technologies
- Metadata specifications
- Data formats
- Shared infrastructures
- Community agreements

Interoperability Frameworks

In Astronomy: The International Virtual Observatory Alliance (IVOA)

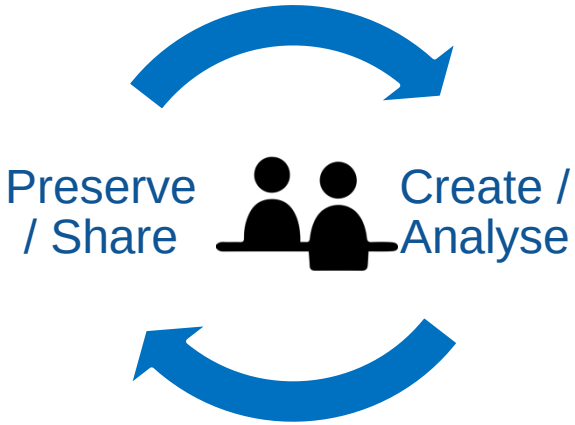
- Created in 2002; National initiatives from the 5 continents; Researchers and IT specialists
- An open and inclusive framework of standards and tools
- 100 “authorities” declared a resource in the Registry of Resources
- Customized by planetary sciences, astroparticle physics, the Virtual Atomic and Molecular Data Center, Registry concepts reused for the Materials registry (RDA WG)



Key drivers needed for change

Skills

- Two cohorts of professionals to support FAIR data:
  - data scientists embedded in research projects
  - data stewards who will ensure the curation of FAIR data
- Coordinate, systematise and accelerate the pedagogy
- Need for a major programme of train-the-trainer activity
- Support formal and informal learning, CPD
- Ensure researchers have foundational data skills



Metrics and Certification

- A set of metrics for FAIR Digital Objects should be developed and implemented, starting from the basic common core of descriptive metadata, PIDs and access.

- Certification schemes are needed to assess all components of the ecosystem as FAIR services. Existing frameworks like CoreTrustSeal for repository certification should be used and adapted **rather than initiating new schemes based solely on FAIR, which is articulated for data rather than services.**

Assessing FAIR services

Many aspects of FAIR apply to services (findability, accessibility, use of standards…) but also important to assess:

- Appropriate policy is in place
- Robustness of business processes
- Expertise of current staff
- Value proposition / business model
- Succession plans
- Trustworthiness

From metrics to incentives

- Use metrics to measure practice but beware misuse
- Generate genuine incentives to promote FAIR practices – career progression for data sharing and curation, recognise all outputs of research, include in recruitment and project evaluation processes…
- Implement ‘next-generation’ metrics
- Automate reporting as far as possible

Investment and Sustainability

- Considerable evidence of the ROI for Open Science and FAIR.
- Provide strategic and coordinated funding to maintain the components of the FAIR ecosystem.
- Component / service providers need to demonstrate value proposition and robust business model.
- Ensure funding is sustainable, understand the dynamics of the income streams and funding models. No unfunded mandates.
- Open EOSC to all providers, but ensure services are FAIR

Data Integration and Interoperability Initiative

- **The interoperability of data in interdisciplinary grand challenge research programmes is one of the major challenges for global research.**
- Information on the initiative, workshops, position documents and the pilot at http://dataintegration.codata.org/
- Dagstuhl workshop in partnership with DDI, 1-5 October: http://bit.ly/codata-ddi-dagstuhl
- Vision of an international coordinating programme with an expanding series of pilots / case studies in interdisciplinary research areas.
- Proposal to new International Science Council for major 10 year programme.

DATA DOCUMENTATION INITIATIVE

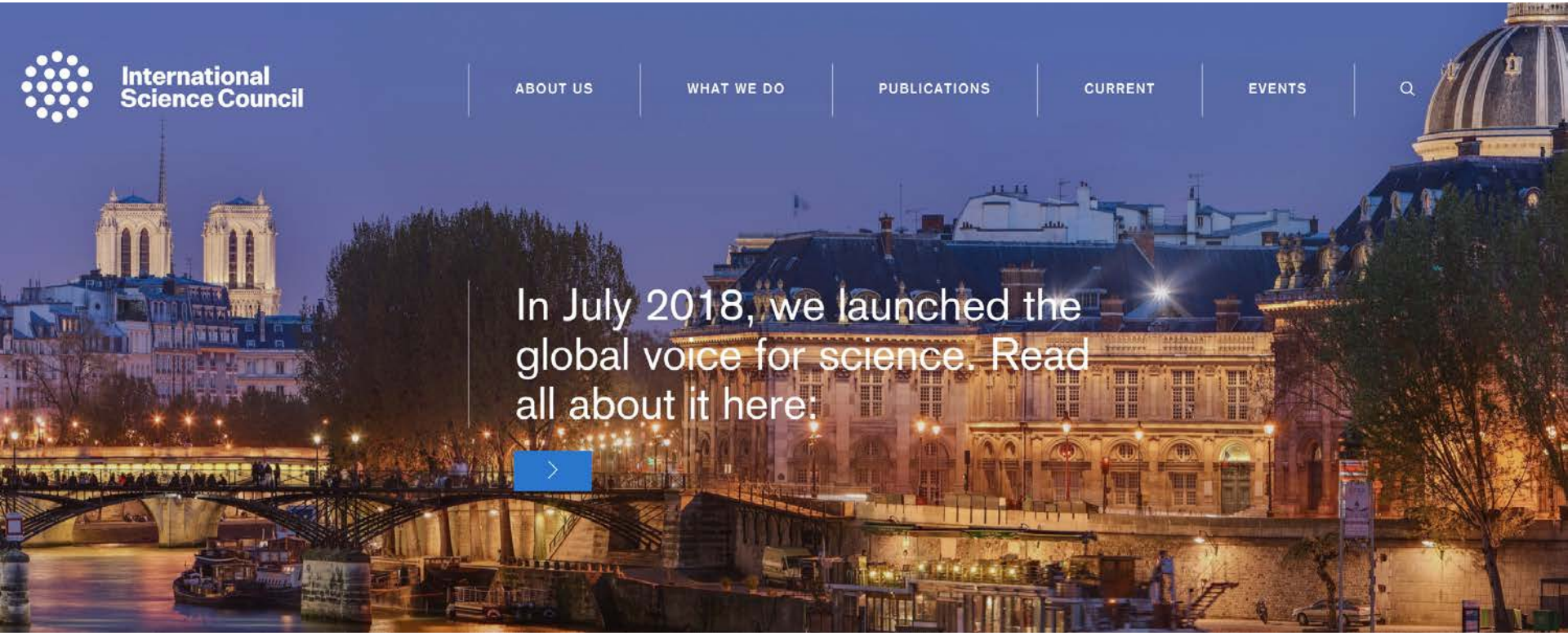
Workshop: Interoperability of Metadata Standards in Cross-Domain Science, Health, and Social Science Applications

Schloss Dagstuhl – Leibniz Center for Informatics, October 1-5, 2018 in Wadern, Germany

International Science Council

International Science Council

https://council.science/



- Formed by a merger of the International Council for Science and the International Social Science Council.
- Explicit mission for all the sciences and for interdisciplinary and transdisciplinary research.

International Science Council

Interdisciplinary Research and Use Beyond the Designated Community

- Major interdisciplinary research issues depend on the integration of data and information from different sources.
- Fundamental importance of agreed vocabularies and standards.
  - Fundamental to integration of social science, geospatial and other data
  - Essential to effective interface of science and monitoring (e.g. Sendai, SDGs, sustainable cities)
  - Recent CODATA work: LOD for Disaster Research, Nanomaterials Uniform Description System
- Huge opportunities but significant challenges.
- **Better exploitation of data resources for research is the epochal challenge of the 21st century.**
- With the merger of ICSU and ISSC to form the **International Science Council**, data integration should be a major initiative and the ISC should have a major role to play to encourage and accelerate these developments.

futureearth
research for global sustainability

IRDR
Integrated Research on Disaster Risk

URBAN HEALTH AND WELLBEING
A SYSTEMS APPROACH

International Science Council

CODATA WG on Description of Nanomaterials

Convene ISUs, International Stakeholders and data experts → Form Working Group → Draft Framework for Description of Nanomaterials → Refine/validate in FP7 Future Nano Needs Project

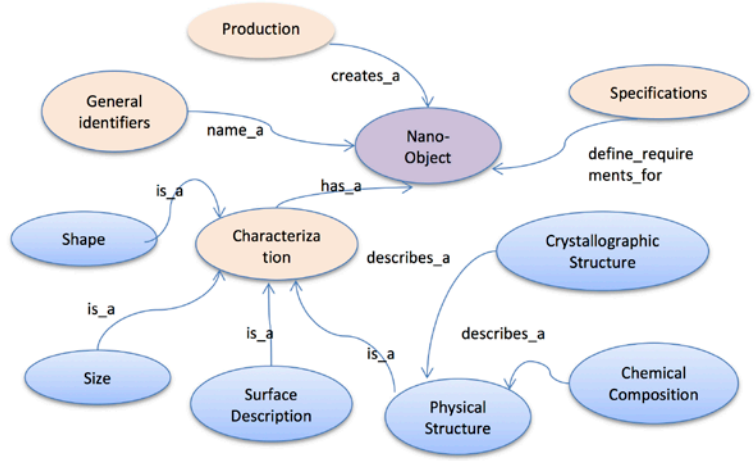


Figure 4. Information categories for describing an individual nano-object

CODATA WG on the Description of Nanomaterials:
http://www.codata.org/nanomaterials

Uniform Description System v.02, May 2016:
http://dx.doi.org/10.5281/zenodo.56720

Future Nano Needs Project:
http://www.futurenanoneeds.eu/

Initiative for Data Interoperability and Integration

- Series of workshops supported by CODATA, ISC.
- Paris, June 2017; London, November 2017; Beijing, August 2018
- Dagstuhl Workshop, October 2018
- **Kind support of CAST for pilots stage of the initiative: examining the core questions, data audit, preliminary recommendations and findings, proposal to ISC**

International Science Council

Initiative for Data Interoperability and Integration

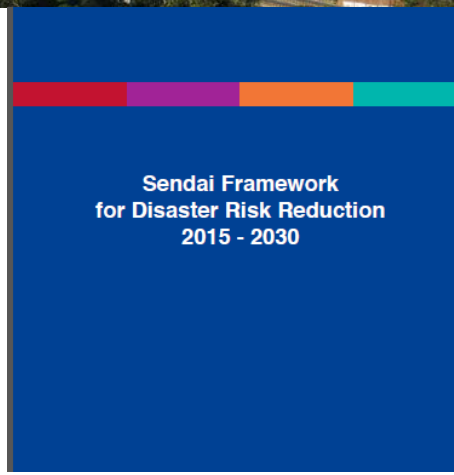
- **Strand 1.** This **addresses important application domains** (infectious disease, resilient cities, and disaster risk, with a possibility of adding agriculture): They have been chosen as major issues where relevant data exists and is accessible, where data integration is a tractable objective, and where there are existing communities of practice that are willing to collaborate.
  - **Demonstrate the tractable, achievable benefits of data interoperability.**

- **Strand 2** will seek to provide **data science support for the pilots** and by extension other disciplines of science that have not yet developed the standards (vocabularies, ontologies, etc) that are necessary for effective data integration. … Formalisation of the discipline-specific vocabularies is an essential pre-requisite for integration of data from different disciplines.
  - **Drawing generic lessons on the actions and support required to promote interoperability and data integration.**

International Science Council

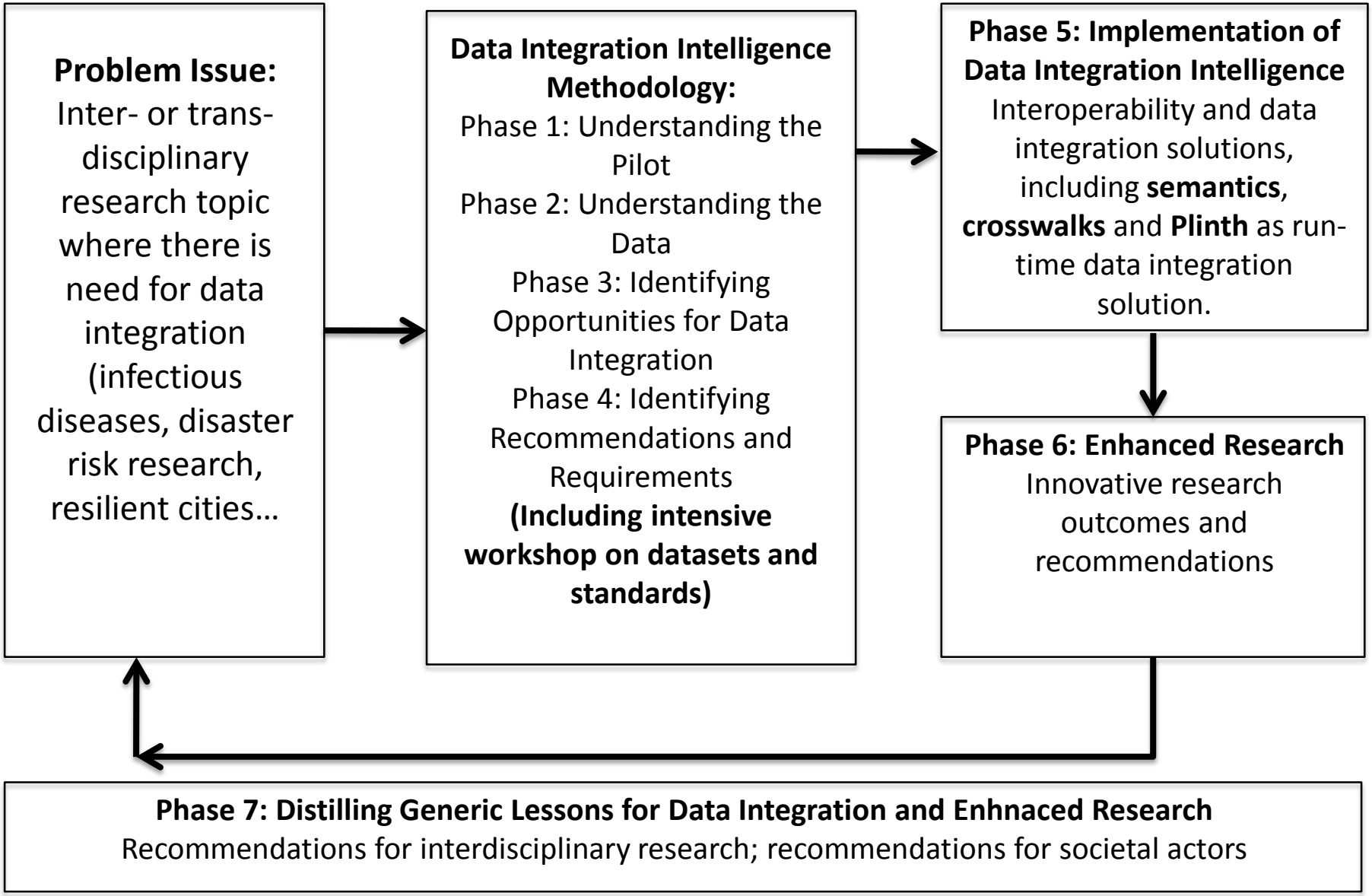
Pilots / Case Studies for Interoperability and Integration

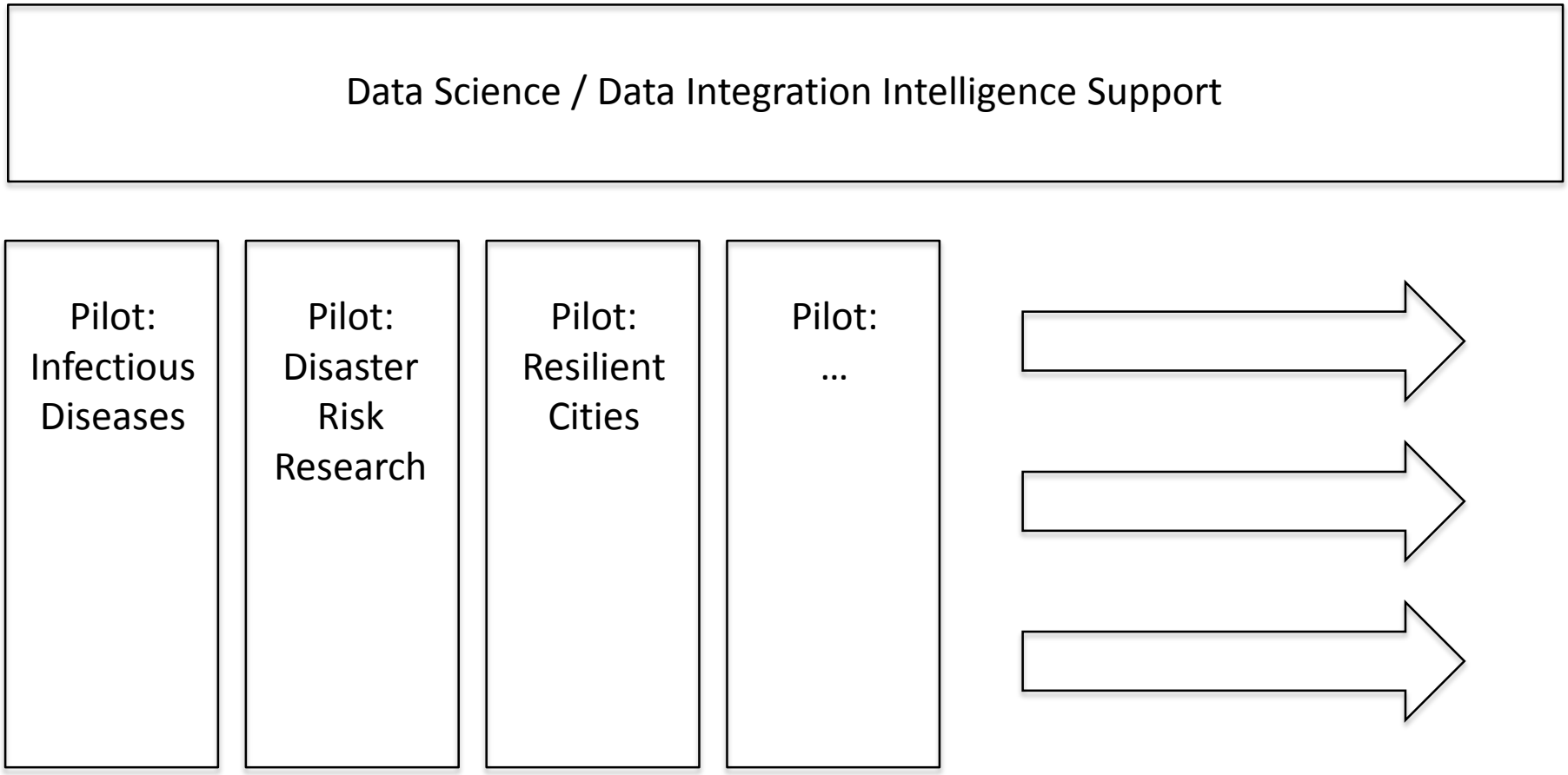
- **Infectious Diseases:** working with **IDDO, University of Oxford.**
  - Preparing a data platform for trial and clinical data relating to Ebola.
- **Resilient Cities:** working with **resilience.io** – international not-for-profit with data platform.
  - **Strong links with ISC Programme on Urban Health and Well-Being**
  - Pilot will focus on systems approach to air pollution, primarily using Medellín, Columbia as a case study. Also possibly engage with Ulaanbaatar, Mongolia; Accra, Ghana.
- **Disaster Risk Reduction:** with **members of CODATA TG on Linked Open Data for Global Disaster Risk Research, Public Health England**
  - **Strong links with ISC Programme on Integrated Research on Disaster Risk**
  - Pilot will focus on challenges of reporting mortality in the context of the **Sendai Framework for Disaster Risk Reduction**.
  - Complementary to the other pilots: significant policy dimensions, looking at different levels of data.



International Science Council

Enhancing Data Integration Intelligence





The Data Science / Data Integration Intelligence Support activity provides generic support for an expanding series of pilots.

Interoperability of Metadata Standards in Cross-Domain Science

- **Dagstuhl workshop in partnership with DDI, 1-5 October.**
- An opportunity to explore alignment of standards at collection and variable level, with reference to the use cases provided by the pilots.
- A pilot for a process of working between standards and pilots/case studies.
- Developed insights into how to address issues of interoperability and integration: what standards can and should be used; how implementing those standards may assist the pilots; what work is necessary on the standards to assist interoperability in these use cases.
- **Five research papers: one on the overall methodology; one each on the case studies; one on the alignment of DCAT and DDI.**
- **Endorsement in principle by ISC Governing Board. Programme Design and Fundraising.**

DATA DOCUMENTATION INITIATIVE

Workshop: Interoperability of Metadata Standards in Cross-Domain Science, Health, and Social Science Applications

Schloss Dagstuhl – Leibniz Center for Informatics, October 1-5, 2018 in Wadern, Germany

International Science Council

International Science Council

Thank you for your attention!

Simon Hodson
Executive Director CODATA
www.codata.org
http://lists.codata.org/mailman/listinfo/codata-international_lists.codata.org
Email: simon@codata.org
Twitter: @simonhodson99
Tel (Office): +33 1 45 25 04 96 | Tel (Cell): +33 6 86 30 42 59