‘Making Data Work for Cross Domain Grand Challenges’:

Working Groups and Dagstuhl Workshop 2020

Version 7, 24 March 2020

Purpose of this Document

In 2021, CODATA—on behalf of the International Science Council and in collaboration with a large number of partner organisations—will be launching a Decadal Programme ‘Making Data Work for Cross-Domain Grand Challenges’, aimed at addressing the critical issues of data access, interoperability and reuse across domain, disciplinary, and institutional boundaries.

This document provides information about a set of initial Working Groups, and a Dagstuhl Workshop, that will play an important role in the preparatory phase of the Decadal Programme and, in due course, in the substantive work of that initiative.

Scope and Background

To face many of today’s global grand challenges, data is needed from different domains and disciplines, and from different institutional levels, and it must be interoperable to be useful. Research projects in such fields, whether for policy or scientific purposes, often involve the use of data from a wide variety of sources, ranging from specific, local data sets to those supplied by higher-level national and international organizations. A huge proportion of research effort is expended to integrate and harmonize this data so that a meaningful analysis can be conducted.

New approaches to discovery and analysis of data allow for truly large-scale projects which rely heavily on automation, cutting across traditional boundaries between domains and disciplines. Such approaches require that the data resources themselves be prone to automated discovery, access, and use. Some of the basic concepts behind such approaches can be found in the FAIR Data Principles.

At heart, however, these ideas require that the multiplicity of models used to structure and describe data be themselves prone to interchange at a computational level. This demands standardization and mapping across various data and metadata models, efforts which have been to some degree ongoing for many years. By themselves, however, these efforts are

insufficient. What is needed are models/frameworks at a higher level of abstraction, which can support such harmonization at a computational level. Efforts to develop such a model are emerging but are still nascent. Further, a community of widespread practice is needed so that such models will be adopted and used.

This problem is a critical one, for both obvious and non-obvious reasons. In order to use computational techniques which can process huge amounts of data, access to that data in usable form is required. Once that problem is solved, however, another emerges: if data are accessed and harmonized computationally, visibility into the provenance of the resulting data - used as the basis of analysis – must exist. Without knowledge of the computational processes which provided the data for analysis, the findings of such analysis cannot be fully trusted, whether for scientific or policy purposes.

Machine learning approaches can be applied to this problem, but have not gained currency in data management practice, nor been expressed as standard models/specifications to address it. Further examination is needed to understand how emerging machine-learning models can be used as a basis for widespread practice and benefit from richer provenance (and other) metadata.

Global grand challenges require data coming from a wide range of domains and institutional levels, presenting us with diverse issues:

- Semantics, classifications, and terminology must be clear not only across domains and national boundaries, but also vertically within chains of data reporting and use
- Metadata specifications for different purposes must be comprehensible at a computational as well as human-readable level, requiring both harmonization/alignment and better machine-actionable models and techniques
- The provenance and processing of data must be made explicit in a fashion which supports further computation, enabling machine reproducibility of findings
- The connection between scientific micro-data and official statistics at the national and international level must be strengthened, to improve both usability and quality for policy and scientific researchers alike

This work brings together experts from both the world of official statistics and global policy monitoring data, technologists, and researchers with a scientific and academic focus. Technologies which address the creation, management and exchange of metadata will be central to this work, to support discovery, analysis, automated processing, and enhanced reusability of data. Further, the intersection of these technologies with machine learning approaches will be considered. A broad range of standard models and specifications in these areas will serve as a focus of the effort, looking not only at how such models can be aligned, but also how best to perform computation across them.

Progress and Collaboration

A number of activities have been undertaken by CODATA and partners over the last two years to explore the issues described above and to develop approaches to help address them. This has included: case studies on infectious disease monitoring, disaster risk reduction (including climate change and adaptation), and the development of resilient cities; engaging with a number of partner organisations; and—with the DDI Alliance—two Dagstuhl Workshops in 2018 and 2019 on the alignment of standards and technologies for cross-domain data combination. The first two workshops in this series have produced draft guidelines and use case documentation to provide insight into the challenges which form the focus of the Decadal Programme. They also helped identify and frame the Working Groups below.

The third workshop—again organized with the DDI Alliance—will take place at Schloss Dagstuhl in Wadern, Germany on 11–16 October 2020. The event will act as the first face-to-face meeting and as a sprint for the Working Groups described below, following and augmenting their online collaboration.

The Working Groups

The four initial Working Groups will each build on work coming out of earlier efforts. It should be noted that they will not be limited by this earlier work, but will be building on it to establish a better-defined set of outputs in light of the overall goals emerging from group discussions.

It is recognized that meaningful solutions to the problems identified will not come from any single group as described here. It is the intention of this effort that cross-fertilization of ideas and interactions between groups are central to how this work will be conducted. Consequently, members of the different groups can expect that they may be drawn into the work of other groups at appropriate points. (In our experience, Dagstuhl provides an excellent environment for this type of collaborative effort.)

Each group will contain experts in the use case and its data, as well as experts in relevant standards, technologies, and semantic/conceptual models.

Group 1: Semantic Interoperability and Conceptual Framework

This group will focus on conceptual and semantic issues impacting the problem space. There are draft guidelines for vocabularies (a.k.a. terminologies, classifications, ontologies, etc.) and how they are managed and published. These will form one basis of the work. Additionally, a conceptual framework for data-sharing activities has been discussed and drafted but is in early stages of development. This framework is critical in communicating

and organizing the outputs from all of the other working groups. Other activities may be identified as the work proceeds.

A key task of this group will be to help other groups coordinate in those areas where semantic and conceptual alignment is needed.

Goals of the group:

- Develop and finalize a draft of the Conceptual Framework
- Further develop the Guidelines for Vocabularies (initiated in the 2019 workshop)
    - Finding a terminology/vocabulary
    - Adopting a terminology/vocabulary
    - Adapting/designing a terminology/vocabulary
    - Making a terminology FAIR
- Collaborate with and support work in other groups
    - Positioning of group guidelines/outputs within the Conceptual Framework
    - Indicators Graph in the Policy Monitoring Indicators group
    - Terminological/semantic harmonization within the Infectious Disease and Resilient Cities group
    - Others as needed

Group 2: Policy Monitoring Indicators

This group will focus on the SDGs, DRR, and other policy monitoring indicators, and the relationship of these data to scientific work in the disaster risk reduction, infectious disease, resilient cities groups. One area of focus would be technology alignment in support of researcher use of international aggregates (like the SDGs), covering DataCube, SDMX, and various related aspects.

Another major thread would be the intersection of semantic technologies with policy monitoring indicators and their use by researchers. One part of this is the graph of relevant indicators and their components/dimensions (variables), against which high-level data sets at the international level can be mapped. The central graph, in turn, would be mapped to vocabularies of greater interest to researchers (OBO Foundry, etc.)

The use of scientific data to validate/enhance international aggregates could also be addressed.

Goals of this group:

- Develop and finalize a draft of the technology alignment approach for national and international organizations working with external users
- Develop and finalize a draft of the Policy Monitoring Indicators graph, including input variables/components
- Develop process for simply mapping aggregate indicators against this graph

- Develop mapping to selected OBO Foundry (or other) vocabularies of interest to researchers
- Establish approaches for enhancement/validation of SDG data using integrated scientific microdata (LSHTM HIV example).

Group 3: Infectious Disease

The focus of this group is on the end-to-end flow of data in infectious disease research. This will initially cover the HDSS infrastructure work from last year and builds on the LSHTM case integrating this data with other data, for end users' analysis and for better using SDG data. It will in turn be expanded to include an examination of pandemic epidemiology data, particularly with reference to COVID-19. All of the draft guidelines in these areas would be further explored. Relationships to DRR and the Indicator Graph would also be explored. This group covers many technical areas, including reusable approaches to data integration and intersection with Schema.org.

The work on the data reuse example from the 2019 workshop can serve as a basis for further work, illustrating the practical technical requirements for data sharing and reuse.

Goals of this group:

- Further develop draft guidelines at all levels from those coming out of 2019 work/DIDA
- Expand and refine these guidelines to meet the data needs of pandemic epidemiology concerning COVID-19 monitoring and research
- Expanded approach to supporting users and user integration of data, including focus on documentation approaches (proposed to WHO?)
- Build on the 2019 data reuse example to illustrate end-to-end cross-domain reuse.

Group 4: Resilient Cities

It is proposed that this group will focus on the topics emerging from recent work in Medellín, Columbia, and on other examples, including some from India and elsewhere. Specifically, data related to transportation, health, planning, and measuring economic impacts will be examined, as a means of approaching challenges for data integration. The use of international data for references purposes will be considered, as will needed harmonization around time, geography, and other issues identified in earlier workshops.

The relationship of data to emerging maturity models for stewardship and use of data in resilient will also be a possible topic: does "(meta)data preparedness" fit into this way of looking at the problems of resilient cities?

Goals of this group:

- Develop guidelines for cities using data from across multiple sources
- Identify shareable techniques/tools for integrating data
- Extend/apply the data reuse example from the 2019 workshop, in coordination with the Infectious Disease group
- Evaluate the relationship between maturity models and data preparedness

Working Groups and the 2020 Dagstuhl Workshop

Working Groups participants will be asked to participate in online meetings and collaboration prior to the event which frame up the work to be conducted in the 'sprint' at Dagstuhl, the exact outputs of the Working Groups, and establish the mechanism for interactions between the groups.

It is envisaged that Working Groups should meet roughly fortnightly online. Initially a convenor will be proposed but a more formal structure, with co-chairs, may evolve if so wished. CODATA will provide secretariat support and will liaise with partners and members to ensure that online collaboration is facilitated effectively.

The 2020 Dagstuhl Workshop is intended to provide an opportunity for face-to-face work and a 'sprint' for the Working Groups. Building on previous work and the ongoing collaboration, it is intended to refine and expand guidance documents, and to produce clarity in how affected organizations can improve practice moving forward. The outputs would be concrete proposals and guidance based on the cases considered, at several levels: for institutional decision-makers and funders, for business users and data managers, and for technology implementers.

Participants and Areas of Expertise

This work brings together experts from both the world of official statistics and global policy monitoring data, technologists, and researchers with a scientific and academic focus. Technologies which address the creation, management and exchange of metadata will be central to this work, to support discovery, analysis, automated processing, and enhanced reusability of data. Further, the intersection of these technologies with machine learning approaches will be considered. A broad range of standard models and specifications in these areas will serve as a focus of the effort, looking not only at how such models can be aligned, but also how best to perform computation across them.

Among the technical standards and specifications being considered in this work are the Data Documentation Initiative (DDI – especially the DDI – Cross Domain Integration specification); Schema.org; the Business Process Modelling and Notation Standard (BPMN); Health Level 7's Fast Healthcare Interoperability Resources (FHIR); many W3C specifications including the

Data Catalog Vocabulary (DCAT), The PROV Ontology (PROV-O), the SOSA/SSN Ontology for sensor networks and spatial data, the DataCube Vocabulary, and others; and models and standards which have emerged from the official statistics domain, including the Statistical Data and Metadata Exchange (SDMX), the Generic Statistical Business Process Model (GSBPM), and the Generic Statistical Information Model (GSIM). (This list is not exhaustive but indicates the type of technical specifications which will be considered.) Additionally, many domain-specific ontologies, vocabularies, and classifications will be included in the workshop.