## Advancing Interdisciplinary Global Challenges Research Through Data Integration

Context and case studies for the DDI-CODATA Dagstuhl Workshop ‘Interoperability of Metadata Standards in Cross-Domain Science, Health, and Social Science Applications’ are provided by the ISC CODATA Initiative ‘Advancing Interdisciplinary Global Challenges Research Through Data Integration’. That initiative is described below.

**Context and Importance of the Initiative**

The pressing and major global scientific and human issues of the 21st century (including climate change, sustainable development, disaster risk reduction) can only be addressed through research that works across disciplines to understanding complex systems, and which uses a transdisciplinary approach to turn data into knowledge and then into action. The digital revolution offers major opportunities and challenges: some disciplines have made dramatic progress in their ability to generate and analyse data, other less so. Yet, the ability to integrate diverse data from different disciplines so as to model and identify patterns in systems relating to the major challenge areas remains relatively rudimentary. In the compelling phrase of the UN Expert Report on the Data Revolution, if we are to have ‘a world that counts’, a world in which data is used to understand complex systems and inform action, then it is essential that attention is given to improving the collection, interoperability, integration and use of data.

**ISC and CAST Funding for the Pilot Activities**

In the context of the creation of the new International Science Council – which provides a unified voice for the human and social sciences *and* the natural and technical sciences – the Council’s Committee on Data (ISC-CODATA) is proposing a timely decadal initiative to improve and scale capacity for data integration in order that interdisciplinary research areas may take full advantage of the opportunities of the data revolution.

Funding has been secured from the China Association for Science and Technology to support a pilot of for this programme which will refine the methodology and approach; to prepare a fuller proposal to ISC; and to explore funding opportunities.

**Project Activities**

The project activity focuses on three pilots which will test and refine the methodology in three important interdisciplinary research areas that depend on a wide range of heterogenous data and encounter fundamental challenges of interoperability and data integration.

The pilot areas are:

1. **Infectious Disease Outbreaks**, led by the Infectious Disease Data Observatory, University of Oxford: pilot lead, Dr. Laura Merson.
2. **Resilient Cities**, led by Resilience Brokers, in partnership with the ISC International Programme on Urban Health and Wellbeing; in the first instance Resilience Brokers are working with Ruta Medellín but more cities may be added: pilot lead, Stephen Passmore.
3. **Disaster Risk Reduction**, led by Public Health England, in partnership with the ISC International Programme on Integrated Research for Disaster Risk (IRDR, Data Working Group) and the CODATA Task Group on Linked Open Data for Global Disaster Risk Research: pilot lead, Prof. Virginia Murray.

The pilots are supported by a ‘data science’ activity which is supporting their collaboration with international scientific unions, with organisations responsible for data standards and ontologies, with interoperability experts, and with technological solutions to improve data discovery and integration. The principle experts involved in this activity are: Phil Archer, Standards and Interoperability Consultant; John Broome, CODATA; Simon Cox, CSIRO; Geoffrey Boulton, CODATA; Bob Hanisch, NIST; Simon Hodson, CODATA; Bill Michener, DataONE; Joachim Wackerow, GESIS and Data Description Initiative; Lesley Wyborn, Australian National University.

Work is proceeding by means of regular calls and some face-to-face meetings between the pilots and data science experts: particularly Archer and Hodson.

**Project Workshop Beijing**

A workshop was held in Beijing in late July 2018 with 25 participants from the project, CODATA, ISC Programmes and Chinese research activities. The workshop had a three-fold objective:

1. To acquaint Chinese colleagues, including representatives of CAST, with the programme, and seek their engagement and involvement in it.
2. To explore and deepen collaboration with major Chinese and international research programmes, such as the Digital Belt and Road, Integrated Research on Disaster Risk and Urban Health and Well-being.
3. To plan development of these strands of work as a basis for major funding and so that they may be presented to the new International Science Council (ISC) as a fundamental part of its future work programme.

Important outcomes of the workshop were greater collaboration between the pilots and the two ISC programmes and with the Chinese Center for Disease Control and Prevention and with the Digital Belt and Road initiative. The project also refined its methodology and this is being put into practice by the three pilots. The documents which lay out this approach have also been shared and are discussed further below.

The project is now working closely with the pilots and preparing documentation for a major, week-long workshop, for a total of 25 participants, at the Leibniz Foundation’s Schloss Dagstuhl. The attendees include representatives of the pilots and an impressive range of data and metadata experts from a variety of domains. This is organized in collaboration with the Data Description Initiative (DDI) and is a remarkable opportunity for intensive work on data interoperability in relation to the pilots.

**Progress of the Pilots**

The pilot case studies are preparing for the Dagstuhl Workshop by providing detailed information about the data sets required to answer particular research or developmental questions. The ‘methodology’ that came out of the Beijing workshop describes a series of phases and steps and is detailed below as an Appendix.

The key tasks of the pilots or ‘case studies’ in preparation for the Dagstuhl workshop is to identify the research questions or issues, to list the data sets that are needed for such research and then to characterize the data, paying particular attention to format, metadata, defined concepts and attributes, use (or not) of controlled vocabularies and ontologies etc. This information will provide a basis for discussion in at Dagstuhl.

The current focus of each pilot is as follows:

**Infectious Disease Outbreaks:** the team at IDDO has been surveying activities and possible questions and data sets. They have now identified the following three activities, to be pursued in order, subject to capacity:

1. Sub-Optimal Medicines: IDDO is preparing a database that takes data from a variety of sources about sub-optimal medicines. The data sources include in situ testing, news stories, official reports (Promed, WHO etc.).
2. Identifying Populations at Risk from Sub-Optimal Treatment and Infectious Disease Outbreaks: building on existing work in relation to malaria treatment. This work looks at clinical outcome data both from actual treatment (?) and clinical trials in different populations. The analysis also needs data on age, co-morbidities (including PEPFAR data on AIDS treatment); census data; population distribution; global burden of disease data; medicine quality data (as above).
3. Understanding the Ebola outbreak: aggregating data of significance to the Ebola outbreak: clinical data and outcomes, where it happened and when, what could have impacted the spread. International response (WHO, MSF); rainfall; population density; transport routes; Ebola treatment units opening; census demographics for at risk populations; rumour datasets (twitter); commencement and acceptance (and location) of cremation etc.

**Resilient Cities:** the team at Resilience Brokers is working with Ruta Medellín and other stakeholders in the city. The team has identified the correlation of air pollution with health issues on the one hand and economic impacts on the other. The sort of questions identified are as follows:

1. What is the correlation between air pollution and respiratory diseases (according to time of exposure to contaminants, location, travel undertaken etc)?
2. How does the quality of the air impact the economic development of Medellín?
3. What are the costs associated with the care of respiratory patients vs the costs of reducing pollutants?

The pilot is currently working to identify types of data sets and the specific data sets needed and available for these analyses and it will then conduct the exercise of describing these data sets. The data sets in scope are likely to include:

* Pollution data (nitrogen monoxide and nitrogen dioxide, ozone (O3), PM 1, PM 2.5, PM10, sulfur dioxide (SO2));
* Travel information (destination-origin survey), travel information from Waze (Medellín, Bello and Envigado, including duration of waze trips; zones with a higher concentration of pollutants;
* Lived environment and weather data (cloudiness, air quality, temperature, wind etc);
* Economic indicators (GDP, taxes, budget use);
* Clinical data on respiratory disease, cost of medical interventions, cost of drugs etc.

**Disaster Risk Research:** the team at Public Health England and colleagues have identified disaster mortality data as the area of focus. Mortality is the key indicator for disasters and an important component to the Sendai Framework reporting process. Nevertheless, many countries have significant issues in reporting mortality due to disasters. Challenges lie in the following areas:

1. Definitions (in terms of the disaster and the relationship of the death to the disaster)
2. Data Sources
3. Techniques for estimating mortality
4. Interpretation of the data
5. Creation of the indicator

The team has produced discussion papers on the challenges involved with mortality data. An important next step will likely be to identify a case study or two and examine in more detail the specific data sources, their characteristics and challenges.

**Appendix 1: The Methodology, Phases and Steps in the Data Integration Pilot**

|  |  |
| --- | --- |
| Phase 1: Understanding the Pilot  | Step 1: What is the research topic? |
|  | Step 2: What are the core research questions? |
|  | Step 3: What data types are needed? |
|  | Step 4: What specific datasets are needed? What are readily available? |
| Phase 2: Understanding the Data | Step 5: What are the access and usage characteristics of the datasets? |
|  | Step 6: What is the format, structure, definitions and descriptions of the data? |
| Phase 3: Identifying Opportunities for Data Integration | Step 7: What are the opportunities and challenges in relation to interoperability and data integration? |
|  | Step 8: What are the data preparation / transformation functions required to make the data analysis ready? |
| Phase 4: Identifying Recommendations and Requirements | Step 9: Run intensive data interoperability workshop. |
|  | Step 10: Present and iterate outputs with the initiative and the discipline community. |
| Phase 5: Implementation of Data Integration Intelligence | Step 11: Planning step / meeting |
|  | Step 12: Improve stewardship and FAIRness of data |
|  | Step 13: Refine Semantics |
|  | Step 14: Implement 'plinth' for machine assisted discovery and interoperability of data. |
| Phase 6: Enhanced Research | Step 15: Conduct and publish analysis using the integrated datasets. |
| Phase 7: Distilling Generic Lessons for Data Integration and Enhanced Research | Step 16: Evaluate the research and action benefits of data integration using this model  |
|  | Step 17: Draw general lessons for data integration programme. |

**Appendix 2: Enhancing Data Integration Intelligence**

Process for data activities in the proposed larger programme with a process and feedback to improve the science.

