Action Theme 3
Scoping Workshop

Belmont Forum
Data Management and e-Infrastructure CRA
Paris, November 28-29, 2016

Contacts:
Jean-Pierre Vilotte vilotte@ipgp.fr
Mark Asch mark.asch@u-picardie.fr
Tsair-Fuh Lin (MOST) tflin@mail.ncku.edu.tw
Michael Vogelsanger (JST) belmont@jst.go.jp
# Agenda

### DAY 1

**Monday 28th November 2016**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Topic</th>
<th>Presenter(s)</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>9h00-9h30</td>
<td>Session I</td>
<td>Opening: Welcome, Objectives, Agenda</td>
<td>R. Samors, M. Asch, J-P. Vilotte</td>
<td>CRA, Belmont Forum</td>
</tr>
<tr>
<td>9h30-10h15</td>
<td></td>
<td>ESGF: a federation for data analysis</td>
<td>S. Denvil</td>
<td>ESGF</td>
</tr>
<tr>
<td>10h15-10h45</td>
<td></td>
<td>Coffee Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10h45-11h30</td>
<td></td>
<td>DIAS: Interoperable and interdisciplinary data</td>
<td>A. Kawasaki, E. Ikoma</td>
<td>U. Tokyo, Japan</td>
</tr>
<tr>
<td>11h30-12h15</td>
<td></td>
<td>CMIP: data and model inter comparison</td>
<td>S. Joussaume</td>
<td>IPSL, France</td>
</tr>
<tr>
<td>12:15-12:45</td>
<td></td>
<td>Open Discussion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12h45-14:00</td>
<td></td>
<td>Lunch Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00-18:00</td>
<td>Session II</td>
<td>Project presentations identifying gaps and barriers (see next page)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00-15h30</td>
<td></td>
<td>Project presentations identifying gaps and barriers (see next page)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15h30-16h00</td>
<td></td>
<td>Coffee Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16h00-18h00</td>
<td></td>
<td>Project presentations identifying gaps and barriers (see next page)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18h10-20h00</td>
<td></td>
<td>Cocktail Dinner</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Context

• Data Explosion - in volume and complexity (large instruments, monitoring networks, large simulations)

• Open data / Open science context (toward reproducible science).

• Inter- and Trans-disciplinary data use for environmental change

• Critical need to bring together application domain scientists and “digital” scientists (computer scientists, data analysts, statisticians, mathematicians).

• Need for reliable decision-making tools and decision-supporting predictions (see DMIPs), particularly for risk/hazard policies for prevention and mitigation.
Context: “big-data paradigm”

The Fourth Paradigm
Data-Intensive Scientific Discovery

4th silo...

Experimental

Thousand years ago
Description of natural phenomena

Theoretical

Last few hundred years
Newton’s laws, Maxwell’s equations…

Computational

Last few decades
Simulation of complex phenomena

The Fourth Paradigm

Today and the Future
Unify theory, experiment and simulation with large multidisciplinary Data
Using data exploration and data mining (from instruments, sensors, humans…)

Crédits: Dennis Gannon
Context: a new paradigm

The scientific discovery process = the inference cycle
## Context: convergence

<table>
<thead>
<tr>
<th><strong>BDEC</strong></th>
<th>see <a href="http://exascale.org">exascale.org</a></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>EOSC</strong></th>
<th>see <a href="http://ec.europa.eu/research/openscience/">http://ec.europa.eu/research/openscience/</a></th>
</tr>
</thead>
</table>
Workshop outcomes

• Input for a draft call text, based on our discussions, covering:

  • Common data and e-infrastructure gaps and barriers where collaboration between existing projects, sharing common research practice, would be beneficial and extensible to other projects.

  • Data and Model Intercomparison Projects (DMIPs).

  • An initial evaluation framework.
AT3 Timeline

• Early November 2016: Feedback from Belmont Forum in Doha.
• Late November 2016: Scoping workshop in Paris.
• March 2017: Publication of call.
• June 2017: Proposal submission deadline.
• Late 2017/Early 2018: Launch of the funded projects.
Not a single dimensional challenge

Data Generation

Data-intensive Challenges

Data analytics, Mining, unsupervised learning

Data processing, transformation

Discovery, Insights, Prediction

Data management

Data reduction query

Data visualisation

Data and method sharing

Adapted from Choudhary

Data -> Extraction/cleaning -> Integration/aggregation -> Learning models -> trigger / question -> predict
A research-driven variety of infrastructures

Ashby’s Law of Requisite Variety

Only variety absorbs variety
Federating autonomous infrastructures

A research-driven strategy .....
E-infrastructure enabling interdisciplinary studies

- Interdisciplinary and trans-disciplinary research conducted around problems rather than in silos.
- Drivers for the co-evolution of data-intensive e-infrastructure system,
- Agreement around interchange and inter-operability of data and metadata.

**Action articulated around a step-by-step and multi-level strategy:**

1. **Survey, analyse and promote collaboration between cross-disciplinary case studies:**

   Accelerate *inter-disciplinary and trans-disciplinary* – natural, social and economic - global change research and improve the quality of decision by **enabling effective use and valorisation of data** (international monitoring and observation systems, large-scale earth systems simulation);

   Emphasise “**going the last mile**”, i.e., transforming scientific knowledge into actionable information for society and achieving influence;

   Increase **quality of science and decision making** through relevant and standardised framework for collaborative interdisciplinary data-and-models inter-comparison.

2. **Distill and collate findings to inform the BF strategy**

   toward a **holistic cross-disciplinary data infrastructure, training and “intellectual ramps” in harmony with interdisciplinary and trans-disciplinary research practice.**

**A multi-level approach:** research domain specialists, data scientists, IT researchers, data-aware engineers, and critical stakeholders, i.e., including infrastructure providers.

**Cycle-up series:** **scoping workshops, competitive call for interdisciplinary case-study collaboration**

**Start with existing supported projects by Belmont Forum and other international initiatives.**
Action Theme 3: Objectives

1. **Identify** a first set of active *interdisciplinary data-use projects and use-cases* federating data- and e-infrastructure for environmental and global change problems and **foster coordination/collaboration between some of them** around common research practices with the aim to develop *mutual understanding and address collaboratively well-identified gaps and barriers*.

2. **Identify** large-scale *Data and Model Inter-comparison Projects* that are relevant for global change and natural risk research, and **foster coordination and collaboration between some of them** with the aim to develop *mutual understanding and address collaboratively well-identified theoretical and practical issues*.

3. Through the above two, **inform the data- and e-infrastructure policy** with bacon of best practices responding to concrete issues, and the **human capacity action theme** with well identified needs in *training and “intellectual ramps” to be developed collaboratively*. 
Continuous update of indicators and evaluation matrix for Belmont projects

Analysis of existing Data and Model Inter-comparison

Steering, monitoring & analysis of the case studies call through a series of workshops

Inter-workshop coordination

Elaboration and publication of the call

Selection process

Competitive call

2 year call: Inter- and trans-disciplinary case studies

Periodic evaluations of coherence and of impacts

Phase 1: 2016

Phase 2: 2017

Phase 3: 2018-2019

ANR (France)
JST (Japan)
MOST (Taiwan)
NSF (USA)
NRF (South Africa)
NERC (UK)
CSIRO (Australia)
EC
Targeted Projects of the Belmont call

Any project with well-identified e-infrastructure, data analysis workflows and data management related problems.

**Interdisciplinary** and trans-disciplinary projects where:

- **Big Data and 4Vs** issues with *multi-type and multi-disciplinary data* are present;
- **Findable, Accessible, Interoperable, Reusable Data** concepts are present;
- **Data management and Data stewardship** are present;
- **Environmental, Social and Economical challenges** are present;
- Needs to *federate data and compute infrastructures* to address the above are well identified and timely to address collaboratively.
The Open Data Iceberg
partly FAIR, partly Cloudy

Technology

The Technical Challenge

Processes & Organisation

The Ecosystem Challenge

The Funding Challenge

The Support Challenge

People

The Skills Challenge

The Incentives Challenge

The Mindset Challenge

motivation and ethos.

Technical challenges

Challenges for federated data-analysis platform

Foster international collaboration and community building toward know-how exchange for

- Storage and computing architecture in support of massive and complex interdisciplinary data
- Streaming data analysis workflow orchestrating analysis of distributed data sources with pervasive provenance systems
- Network-based and provenance-based data movement between different and distributed data and computing sources honouring data and AAI policies
- Concurrent data access and data representation for data-intensive analysis
- Adding access, data analysis and visualisation services on top of the data
- Energy and Green technology challenges
- Collaboration with private providers: public cloud and others

Bridging the gap between multi-type and multi-disciplinary data

- Data stewardship, data and metadata formats, data exchange protocols
- Credential and interoperability at the data level
- Implementing FAIR data principles
- Structured/unstructured data
- Dealing with and assimilating different data spatial and temporal scales
- Strengthen the use of data by and from other communities especially socio-economical communities

Data Model Inter-comparison - validation - prediction

- Identify trans-national expertise and beacons of good practices
- From model to coupled models framework
- Bridging scientific-driven and policy-driven concerns into a framework
- Extension to other socio-economic and health issues
- Foster standardisation of protocols and methods across disciplinary
The ecosystem challenges

Federating autonomous data and compute infrastructures ecosystems

- Research-driven strategy
- Hourglass architecture individualising and isolating layers supporting different concerns
- Data policies: across different disciplines, data providers and countries
- Involving data and compute resources providers, national science agencies

Federating data policies across domains and national boundaries

- AAI and data licensing
The funding and support challenges
The incentive challenges

Data publication and citation
Data plan for cross national boundary projects
Intellectual ramps
Skills and mindset challenges

Data literacy
Data analytics literacy including statistics and machine learning