Science of the (near) Future: Its power and requirements

Robert Gurney University of Reading
“Big Data” problems in environmental science

• Volume

• Variety
  – Different data sources
  – Different formats
  – Different semantics
  – Complex grids

• Veracity
  – i.e. Data quality

• Velocity
  – EO datasets terabytes per day
  – Numerical simulations petabytes per day
Growth in “Big Data” in climate

J T Overpeck et al. Science 2011;331:700-702
Open Data

• Lots of public sector data have been made open and freely available through various government and other portals globally, but in unconnected ways.
Data Silos

- All these data still exist in different data silos (e.g., different EO archives).
What are the opportunities offered by new cloud technologies?

- The key to the cloud approach is the representation of everything as a service, that is our data, models, visualisation tools and expert knowledge.
- A space for:
  - Exploring data
  - Linking models
  - Accessing knowledge
  - Visualisation tools
What is “Cloud Computing”?

- Scientists
- Policy Makers
- Local Communities
- General Public

Web Interface

Web Services

Virtual Resources

Hardware Resources

Models

Data
Why the Cloud?

• Easy access to resources
   Infrastructure as a Service - IaaS
   – IaaS: hardware resources as a utility.
   – Allows the infrastructure to scale to meet user demand and maintain quality of service.
   – Ease of mind: issues of reliability, performance and security at the hardware level are outsourced.
   – Allows us to focus on solving domain-specific problems.
   – No usage quotas (unless you want to).
   – Very few technical hoops to jump through.
MyOcean: Copernicus Marine Core Service
(http://www.myocean.eu)
Visualisation of in situ observations (floats, moorings etc)
Comparing models and observations

Intercomparison Test
> Sea Water Salinity Intercomparison

Units: psu
Time: 2011-02-09 00:00:00.000Z +/- 1 day
Depth (m): 130

Open in Google Earth  Permalink  Email Link  Export to PNG
Anatomy of a typical MELODIES service

(there is a lot of variation on this theme!)

Earth observation imagery (raster)

EO Data processing

Discrete features (RDF)

Linked Data sources

Linked Data integration (Strabon)

User interface (WebGIS) driven by SPARQL queries

Contextual information

e.g. Maps, Administrative boundaries

e.g. Population data, CORINE land cover
Realising the potential of environmental data, models and tools.
Realising the potential of environmental data, models and tools
Realising the potential of environmental data, models and tools
Realising the potential of environmental data, models and tools.
Realising the potential of environmental data, models and tools
Realising the potential of environmental data, models and tools.
Common Themes

Processing of large data stores
  – Especially Earth Observation and models (tera- to peta-scale)
  – **Big Data “volume” problem!**

Combination of data from diverse sources
  – Earth Observation (e.g. Copernicus)
  – GIS / mapping (e.g. INSPIRE)
  – Socioeconomic and population (e.g. data.gov)
  – **Big Data “variety” problem!**

Which data are right for each user? How much can we trust the data?
  – **Big Data “veracity” problem!**
Established to foster global environmental change research; Initiated June 2009 by NSF and NERC, building on IGFA
E-infrastructures and Data Management

CRA

• “....the need to address global environmental challenges requires a more coordinated approach to the planning, implementation, and management of data, analytics and E-Infrastructures” through international collaboration.

• Belmont Forum, New Delhi, February 2013

NB: all funding agencies, like all publishers, have problems of reproducibility
The Belmont Forum adopts this data policy and the following principles to widen access to data and promote its long-term preservation in global change research; help improve data management and exploitation; coordinate and integrate disparate organizational and technical elements; fill critical global e-infrastructure gaps; share best practices; and foster new data literacy.

The Belmont Forum recognizes that significant advances in open access to data have been achieved and implementation of this policy and these principles requires support by a highly skilled workforce. The Belmont Forum recommends a broad-based training and education curriculum as an integral part of research programs and encourages researchers to be aware of, and plan for, the costs of data intensive research. The Belmont Forum’s ambition is that this policy and these principles will take positive steps toward establishing a global, interoperable e-infrastructure based on cost-effective solutions that can help enable actionable and societally beneficial science.

Data should be:

- Discoverable through catalogues and search engines
- Accessible as open data by default, and made available with minimum time delay
- Understandable in a way that allows researchers—including those outside the discipline of origin—to use them
- Manageable and protected from loss for future use in sustainable, trustworthy repositories

The Belmont Forum and its members will support and promote this data policy and principles with the intent of making these data principles enforceable over time.
Foster Communication, Collaboration and Coordination through Establishing a Data and e-Infrastructure Coordination Office

• The Belmont Forum is establishing a data and e-Infrastructure Coordination Office to foster communication, collaboration and coordination across Belmont Forum-funded projects and agencies and engage with the wider global e-infrastructure community.

• Activities include:
  • 1) implement, monitor, support, and evaluate interoperability capabilities among Belmont Forum-funded projects and participants;
  • 2) reduce or avoid duplication;
  • 3) leverage resources (including those in the other action themes);
  • 4) map the data and e-Infrastructure landscape;
  • 5) facilitate global interoperability convergence and communication.
The Belmont Forum will promote good data stewardship in all Belmont Forum funded research.

This action mandates funders to require funded researchers to provide and comply with an Enhanced Data Plan (EDP). This Plan should cover:

- the management of data
- connection of the data to associated context (research project, resulting publications, researcher details, data provenance, etc.)
- facilitating the discoverability of the data
- ensuring that the data is capable of reuse/exploitation.
Demonstrate international best and common practice, and inform Belmont Forum research e-infrastructure policy, through identification and analysis of cross-disciplinary research examples

- The Belmont Forum should implement scoping workshops and international calls for case studies, and then share infrastructure to define and develop best practice. Reports from the case studies will additionally determine the policy for data e-infrastructure deployment and sharing in future Belmont Forum-funded projects, ensuring best and common international practice for optimized data exploitation, led by science need and not technological availability.
Support the development of a cross-disciplinary training curriculum to expand human capacity

- There is considerable evidence of the shortage of skilled people worldwide with both a well-developed understanding of environmental science and good data and informatics skills and knowledge.
- The Belmont Forum will
- 1) support the development of a holistic international training and education curriculum in data-intensive science, data-aware Information Technology and Informatics for Environmental Science, for delivery to environmental, social, and computer scientists;
- 2) launch the curriculum through the creation of a number of international short courses and immersive winter/summer schools.
Timetable

• Policy adopted October 2015
• Detailed Implementation Plan published March 2016
• Champions appointed Spring 2016
• Workshops through 2016
• Calls and funding start to flow late 2016 onwards
The need to address global environmental challenges requires a more coordinated approach to the planning, implementation, and management of data, analytics and e-Infrastructures through international collaboration. — Belmont Forum, New Delhi, February 2013

The Belmont Forum e-Infrastructures and Data Management Collaborative Research Action brings together domain scientists, computer and information scientists, legal scholars, social scientists, and other experts from more than 14 countries to establish recommendations on how the Belmont Forum can implement a more coordinated, holistic, and sustainable approach to the funding and support of global environmental change research.