Sustainable Business Model for Data Repositories

Sun Kun OH (Konkuk University)

2 December 2016

SUT, Thailand

In yesterday's talk..

A "data cloud" for HEP

- Few – O(5-10) - large centres
  - Multi-Tb private (SDN) network between them
  - Treat as a single "virtual data centre"
    - Policy replicates data inside for security and performance
    - Think of RAID across data centres
  - Store all of the "AOD" data here; Do not replicate data to global physics institutes (major cost)
- Pluggable compute capacity:
  - HEP resources at these centres & other large centres
  - Commercial compute
- Model allows commercial data centres
  - For storage – enough redundancy that a commercial centre could unplug
  - For compute
  - Relies on networking and GEANT/Esnet etc. connections to commercial entities, policy
- Users access data in this cloud remotely
  - Eventually download "ntuples" – or equivalent
  - All organised processing is done in this model
- Enables new analysis models: all data can be seen as colocated
  - Get away from the "event-loop" → queries, machine-learning, etc.

This idea has been discussed in the WLCG community (e.g. see I. Fisk CHEP plenary)

- Hybrid model:
  - HEP-resources at a level we guarantee to fill → cost-effective
  - Commercial resources for "elasticity"
- Needs new funding models

In yesterday's talk..

Beyond HEP

- This virtual data cloud model may be very interesting for other sciences
  - E.g. SKA Regional Centres
  - Works also for DUNE, Future facility development, others
  - Can provide resiliency and long term preservation capabilities
- Integrating commercial resources
  - Requires (potentially) significant changes in funding models
  - Can we actually procure commercial resources at large-enough scale to get economy?
    - HNSciCloud as a proof-of-principle of joint procurement
    - Can we purchase from the largest cloud vendors? Politics?
    - Real cost-efficiency and elasticity requires a "spot-market" price
  - How do we arrange performant and secure network connections to commercial resources?

A workshop was held on

SUSTAINABLE BUSINESS MODELS FOR DATA REPOSITORIES

Overview Presentation

Paul F Uhlir, Consultant
3 November 2016
OECD HQ, Paris

OCDE
DES POLITIQUES MEILLEURES POUR UNE VIE MEILLEURE

Meeting minutes

- Date: 3-4 November 2016
- Venue: Headquarter OECD, Paris
- Attendances: 15 experts

Main issues were about ..

SUSTAINABLE BUSINESS MODELS FOR DATA REPOSITORIES

The primary objectives are to cast light on the following issues:

1. How data repositories are currently funded;
2. What additional, innovative, income streams are available to data repositories;
3. How various income streams may fit together into a business model; and
4. How various business models match budgetary structure and the willingness and ability to pay of the various stakeholders.

Attendants include

SUSTAINABLE BUSINESS MODELS FOR DATA REPOSITORIES

EXPERT COMMITTEE

Co-Chairs: Allen Dearry, NIH (US); Ingrid Dillo, DANS (NL); Simon Hodson, CODATA (FR)

Members: Francine Berman, RPI (US); Phillippe Cudre-Mauroux, University of Fribourg (CH); Klaas Deneudt, VLIZ (BE); Martie van Deventer, CSIR (ZA); Michael Diepenbroek, PANGAEA (DE); Uri Gabai, Ministry of Economy (IL); André Golliez, Impact Hub (CH); Peter Grolimund, Teradata (CH); Natalie Harrower, Digital Library Ireland (IE); Kazuhiro Hayashi, NISTEP (JP); Irina Kupiainen, CSC-IT, (FI); Brian Lavoie, OCLC (US); Wainer Lusoli, EC (BE); Devika Madalli, ISI (IN); Hiroshi Manago, Government of Japan (JP); Cameron Neylon, Curtin University (AU); Seo-Young Noh, KISTI (KR); Sun-Kun Oh, Konkuk University (KR); Happy Sithole, CSIR (ZA); Roar **Skålin, Research Council of Norway, (NO); Kihoko Suda, Government of Japan (JP); Andrew Treloar, ANDS (AU)**

Observers: Hyung-Jin Lee, KISTI (KR); Mustapha Mokrane, WDS (JP)

Consultants: John Houghton (AU); Paul Uhlir (US)

OECD, Global Science Forum: Carthage Smith (FR), Taro Matsubara (JP)

EMBL-EBI: Providing Bioinformatics Research Infrastructure for the Life Sciences

Rolf Apweiler
Director, EMBL-EBI
www.ebi.ac.uk

EMBL-EBI

EGI-Engage

EGI Input for OECD Meeting

Matthew Viljoen
Senior Operations Officer
EGI Foundation

matthew.viljoen@egi.eu

OECD Global Science Forum Workshop on
Revenue Sources and Cost Optimisation
OECD, Nov 3-4 2016

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under grant number 654142

Figshare + Sustainability

figshare
credit for **all** your research

Dan Valen – Product Specialist, fig**share** 11/2016

ELIXIR Europe

A distributed infrastructure for life science data

OECD, Paris, 3-4 November 2016
Andrew Smith

elixir

www.elixir-europe.org

International Collaboration for **Data Preservation** and **Long Term Analysis** in High Energy Physics

CERN / WLCG Data Repositories

OECD Global Science Forum Workshop on Sustainable Business Models for Data Repositories

https://indico.cern.ch/event/577772/

Jamie.Shiers@cern.ch

Data may be

- Articles, reports, documents;
- Pictures, figures, plots;
- Musics, vocal archives;
- Visual archives (AP, NS);
- Experimental data before and after analysis (CERN, LIGO).

Some terminologies

- Data repository = data storage or data warehousing, a particular kind of setup which an enterprise or an organization has chosen to keep certain kinds of data.
- Data stewardship = data governance, focusing coordination, implementation, and maintenance of data repositories in an organization.
- Data stewards enable an organization to take control and govern all the types and forms of data and their associated libraries or repositories.

Nature of open data

- Who (claim to) produce the data?
  - Individuals, e.g. researchers
  - Institutions = employers of individuals, e.g. research laboratories, universities, governments
  - Collaborations = organizations of individuals, e.g. ATLAS

Nature of open data

- Cost of producing the data
  - Personal money
  - Research fund from private or public sector
  - Institutional budget which may be the public spending
  - Voluntary or obligatory contributions requested from collaborations

Nature of open data

- Ownership of the data who (decide to) open
  - Individuals
  - Institutions
  - Organizations
  - Private company
  - Government representing public

Nature of open data

- Data management and storage
  - In owners' PC
  - Special public institutions or organizations for logistics, e.g. WLCG
  - Private companies, e.g. Google
  - Funding agencies representing public sector or government

Nature of open data

- Open to whom?
  - General public
  - Members or registered customers but anyone may apply for the membership or registration
  - Restricted application and/or limited membership
  - Exclusive communities

Nature of open data

- By what extent and how?
  - Duplicable
  - Downloadable in restricted formats
  - For eyes only
  - Free access (any time anywhere)
  - Online request and authorization required
  - From an established access points

Nature of open data

- Cost of access payable to the owners or their representatives
  - Free
  - Membership fee
  - Microcharge for each access
  - Annual subscription
  - Institutional MOU (free to the employees of the institutions)

Some science data repositories in Korea

- Korea Biodiversity Information Facility (KBIF), National Science Museum
  - http://nabipos.kbif.re.kr/index.jsp
- In KISTI, National Science & Technology Information Service (NTIS)
  - http://www.ntis.go.kr/ThMain.do
- In KISTI, NDSL (National Digital Science Library)
  - http://www.ndsl.kr/index.do
- And Tier-1 Center of WLCG at GSDC, KISTI

HEP data and WLCG

Open (Access to) Data

- **It took a long time to get where we are with Open Access to publications**
- HEP has made the "data behind publications" available for decades (HEPdata)
- "The" data is much more complex: may well require significant amounts of **documentation**, **software**, **storage** and **computational / network resources** + **SUPPORT!**

➢ **Perhaps this deserves its own workshop series?**

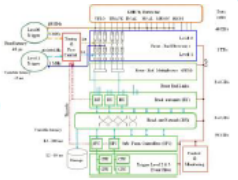
HEP data and WLCG

How Much Data?

- **100TB** per LEP experiment: **3 copies** at CERN (1 on disk, 2 on tape) (+ copies outside)
- **1-10PB** for experiments at the HERA collider at DESY, the TEVATRON at Fermilab or the BaBar experiment at SLAC.
- The LHC experiments is already in the multi-hundred PB range (**x00PB**)
- **10EB** or more including the High Luminosity upgrade of the LHC (HL-LHC)

HEP data and WLCG

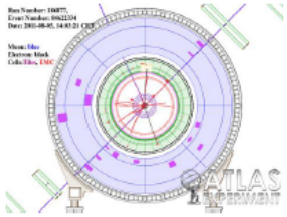
What Makes HEP Different?



- We **throw away** most of our data before it is even recorded – "triggers"
- Our detectors are **relatively stable** over long periods of time (years) – not "doubling every 6 or 18 months"
- We make "**measurements**" – not "**observations**"
- Our projects typically last for **decades** – we **need** to keep data usable during at least this length of time
- We have **shared** "data behind publications" for more than 30 years… (HEPData)

HEP data and WLCG

What is the problem?



- The data from the world's particle accelerators and colliders (HEP data) is both **costly** and **time consuming** to produce
  - That from the LHC is a particularly striking example and ranges in volume from several hundred PB today to tens of EB by 2035 or so.
- HEP data contains a wealth of **scientific potential**, plus high value for **educational outreach**.

- Given that much of the data **is unique**, it is essential to preserve not only the data but also the full capability to reproduce past analyses and perform new ones.
  - **This means preserving data, documentation, software and "knowledge".**
- There are numerous cases where data from a past experiment has been re-analyzed: we **must retain** the ability in the future

HEP data and WLCG

A bit about CERN / WLCG

- Experience from LEP / WLCG gives concrete numbers of costs for 10TB / 100TB / 100PB / (10EB) data stores
- WLCG Tier model allows us to compare costs at different scales
- CERN “managed storage” costs (budget) roughly flat from LEP (500TB in 2000) to HL-LHC (10EB)
- An order of magnitude reduction in scale (e.g. a Tier1) “saves” only a factor in manpower costs
- A multitude of “data repositories” is about the most inefficient way to do it – “small” = expensive
- Most likely WLCG will reduce # storage sites – even if “politically” difficult

HEP data and WLCG

If you want to save money...

- **Don't build multiple "small" bit repositories**
  - These will cost more (much more when integrated) and be less reliable
    - WLCG Tier0 (aka CERN) will manage from ~1PB to ~1-10EB with constant manpower and ***falling*** h/w costs over a period of 3 to 5 decades!
  - Don't imagine that a bit repository - even a "certified one" - can do everything!
- Regulations may favour multiple repositories - changing this may be a key message regarding sustainability...
- **WLCG has "shown" that locality of data (given sufficient network bandwidth) is no longer critical**

- **"Common solutions", e.g. CernVMFS, Zenodo (INSPIRE) are another, all too obvious(?) way of saving money**

Data price change

- Production cost per unit amount of HEP data will become cheaper and cheaper.
- The cost of logistics (storing, distributing, etc.) will confirm “the economy of scale” (smaller tiers are less economical than larger ones) for busniness.
- The consumers (end-users or researchers) may pay for the data, or paid by funding agency

The challenge

The Challenge: Sustainable Business Models for Data Repositories

- Research funder policies – quite rightly – mandate data stewardship.
  - OECD Principles and Guidelines, 2007
  - G8 Science Ministers Statement, 2013
  - Major funders in US, UK, EC Horizon 2020 data policy etc.
- Increasing need for data repositories and data stewardship.
  - Increasing volume presents a challenge.
  - Requirements for stewardship present a greater challenge.
- **Sustaining digital data infrastructure is a major issue for science policy!**
- Genuine concern that current funding models will prove inelastic and not meet the growing requirements – concern on the part of repositories and funders.
- Witnessing Innovation
  - Changes in funding / business models (ADS, TAIR; DANS, ICPSR)
  - Innovative business models (Dryad, FigShare)

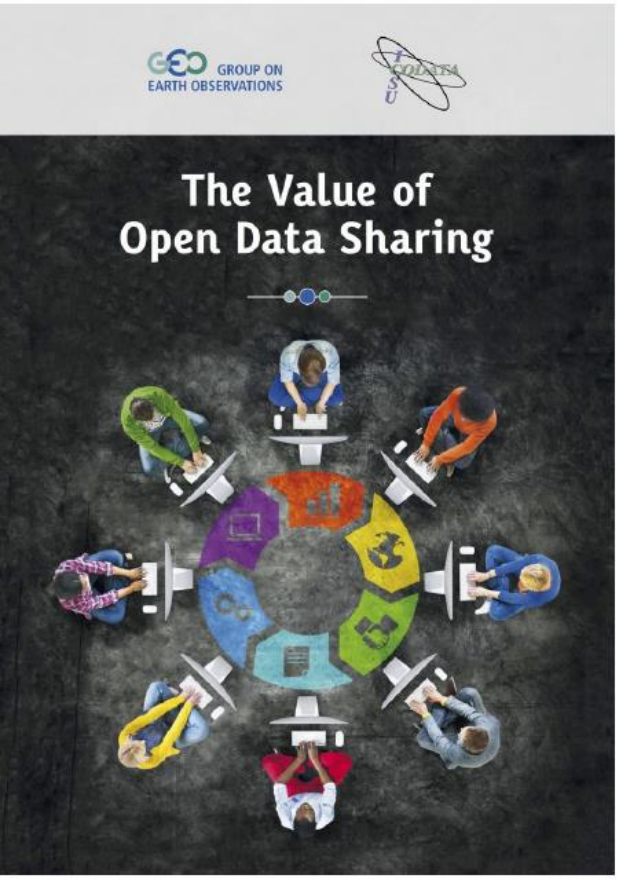
Data flow

Where should research data go?

Homogenous data collections essential for research	• Earth observation data; • Genetic data; • Social science survey data…	National and international data archives
Significant data outputs of publicly funded research	• Significant data outputs from funded projects; • Raw and analysed experimental data…	National or institutional data archives; data papers
Data underpinning research publications	• Raw and analysed data for reproducibility (evidence); • Data behind the graph…	Dedicated data archives (e.g. Dryad)

Data sharing

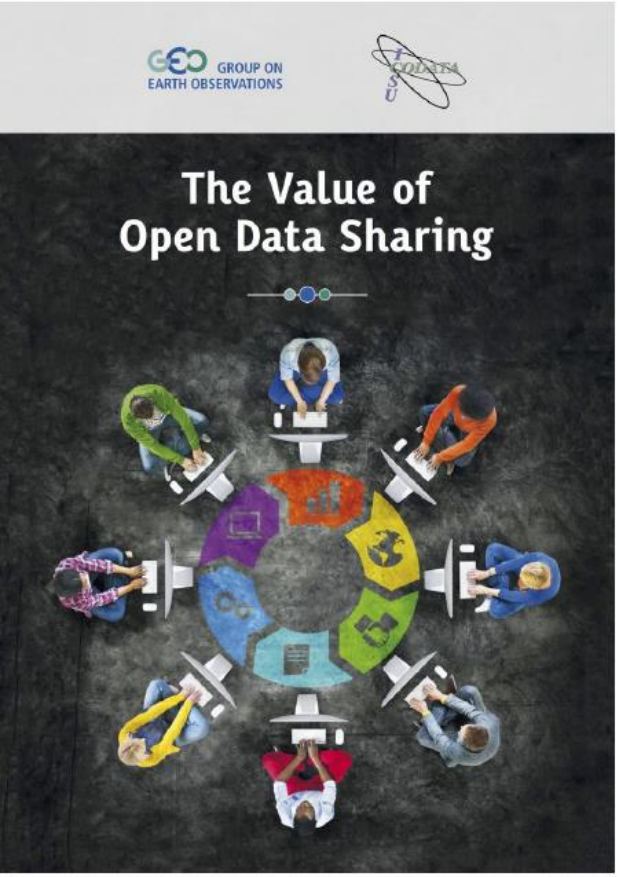
The Value of Open Data Sharing



- Report by CODATA for GEO, the Group on Earth Observation.
- Provides a concise, accessible, high level synthesis of key arguments and evidence of the benefits and value of open data sharing.
- Particular, but not exclusive, reference to Earth Observation data.
- Benefits in the areas of:
  - Economic Benefits
  - Social Welfare Benefits
  - Research and Innovation Opportunities
  - Education
  - Governance
- Available at http://dx.doi.org/10.5281/zenodo.33830
- **GEO DSWG is building on this work with further examples: would be valuable to work with this community.**

Data sharing

Economic Benefits of Data Sharing



- 'Many studies and reports have documented the positive value of openness for EO data, specifically, and for various other types of data and information, more generally.'
- Weiss 2002: quantified considerable economic benefits of making meteorological data open ($400-700M in gross receipts; businesses and employment).
- Houghton 2011: apart from economic benefits, gross saving for Australian Bureau of Statistics of AU$3.5M by eliminating charging and management structure.
- Houghton 2014: Estimate unrealised benefits of research data of AU$1.4-4.9BN set against estimated AU$130-200M cost of data infrastructure.

- **Interested to know what studies of the benefits of data availability have been conducted in this area of research?**

Data flow

- Like commodity, at each transferring, the value of data is increased, i.e., the value-added price is applicable.


감사합니다
Natick
Grazie
Danke
Ευχαριστίες
Dalu
Obrigado
Thank You
Köszönöm
Tack
Спасибо
Dank
Gracias
Seé
谢谢
Merci
ありがとう