The opinions expressed in this presentation are my personal opinions and do not necessarily reflect the draft report of the High Level Expert Group for the European Open Science Cloud.
Open Science as a Social Machine

Where (the…….) are the Data?
Lamenting
Neelie Kroes (then Vice-President of the European Commission, responsible for the “Digital Agenda” for the European union) When she announced the EU’s Open Data Strategy she opened with “Data is the New Gold”. We wish it were that simple.

The value of data

Nature Genetics, 43, 281–283 (2011)

Barend Mons¹–⁴, Herman van Haagen¹, Christine Chichester²,⁴, Peter-Bram ‘t Hoen¹,⁴, Johan T den Dunnen¹, Gertjan van Ommen¹,⁴, Erik van Mulligen³,⁴, Bharat Singh²,³, Rob Hooft²,⁴, Marco Roos¹,²,⁴, Joel Hammond⁵, Bruce Kiesel⁵, Belinda Giardine⁶, Jan Velterop⁴,⁷, Paul Groth⁴,⁸ & Erik Schultes¹,⁴
DATA is the new oil
5 min.
of
Contemplation
From Individual Brilliant Minds to Social Machines
Open Science as a Social Machine

WE ARE HERE!...where are you !? 

Data Repositories (OA/linked if lucky)

Christmas trees of Hyperlinks

HTML/XML

PDF

Print
**human knowledge bubble:**
a wave of current trends in the field that factors in only small amount of the knowledge space. Most scientists are not aware of historical and lateral information.

**knowledge space:**
boundaries of biomedical knowledge (exponential growth)

**computational knowledge bubble:**
computer can expand information awareness
10 min.
of
Reality
Data loss is real and significant, while data growth is staggering.

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady.
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade.

‘Oops, that link was the laptop of my PhD student’
Current scholarly publishing and the Elusive Explicitome Phenomenon

example from: & Verspoor 2013

The Open Access (article) fight !!!!!!

The Elusive Explicitome: what escapes us (95%)
The EXPLICITOME

everything we have ‘claimed’ in science

Estimate today (LS) : \(10^{14}\) associations…….
The EXPLICITOME

Open Science ≠ OA articles
5 min.
of
Open Science
Simplified e-Science

Ridiculogram
AERIAL SURVEY
pattern recognition in
Ridiculograms

FAIR for computers

X

FAIR for people

HUMAN EXCAVATION
rationalisation and
‘confirmational reading’

‘Why would I believe this association’???
The Explicitome is spread over **Thousands** of databases

The Explicitome is estimated at $10^{14}$ assertions

The ‘Cardinal Explicitome’ is estimated to be ‘only’ $10^{11}$ assertions
We publish about less than a million LSConcepts!

> 100 K predictions
[disease] – [gene]
- Link over 70 databases
- Discovery of indirect relationships
- Rationalisation (workflows)
- Dig into evidence via provenance
The ...... Knowlet is mostly connected to the IL-6/IL-1/TNF/AR via RNA binding, drug interactions, cell proliferation, cellular targeting, proteolysis and kinase activity.
new variant

variant in patient (LOVD)

TSS (FANTOM5)

GWAS (everywhere)

Isoforms (NextProt)

Gene (ENTREZ)

Proteins (UniProt)

Tissue location (HPA)

Tissue-Disease (PubMed)

Pathways (wikipathways)

SEED articles (EURETOS)

new variant
secure (polymorphic encrypted) Query

User Decides

encrypted Query to API

Secure (polymorphic encrypted) Query

User Decides

User has subscription to (all) publishers

New S-(p?)-O's

Scholar

annotation

EURETOS

EURETOS

I open access

EURETOS

EURETOS

I open access

EURETOS

I open access

EURETOS

annotation

nanopublication

assertion

provenance

publication info

all references in PDF > Lazarus EURETOS-Graph…

User Decides

User Decides

seed article
5 min. of Vision
ELIXIR: An international distributed infrastructure for biological data

Technical platforms

- Data
- Standards
- Tools
- Compute
- Training

User communities

- Marine metagenomics
- Crop and forest plants
- Human data
- Rare diseases
European Open Science Cloud
European
Open Science
Cloud
The EOSC

Open Science

OA (articles)
complexity (and 4V) of data

disciplines involved

Data Stewardship Challenge
EOSC: Framing

- Trusted access to services & systems
- Re-use of shared data
- Across disciplinary, social and geographical borders
- Federated environment, across Member States
EOSC: ‘Internet approach’

- **Minimal** international guidance and governance
- **Maximum** freedom to implement.
- **Globally** interoperable and accessible
- **Globally** embedded in a ‘Commons’
EOSC: Scope

- Human expertise
- Core resources
- Standards, best practices
- Underpinning technical infrastructures
- A web of Services
Open Science

Open Innovation

Systematic and professional data management

Long term data stewardship
The majority of the challenges are **social** rather than **technical**.

Not just the size of data, but in particular **complex data** and **analytics across domains**.

Shortage of **data experts** globally and in the European Union.

**Archaic system of rewards** and **funding** of science and innovation.

‘Valley of death’ between **(e-)infrastructure providers** and **domain specialists**.

**Short funding cycles** of core research infrastructures are **not fit for purpose**.

**Fragmentation** between domains causes **repetitive** and **isolated** solutions.

Distributed data sets increasingly **do not move** (**size & privacy** reasons).

Centralised HPC is **insufficient** to support **distributed meta-analysis and learning**.

However, the **major components** for a **first generation EOSC** are largely ‘there’

But ‘**lost in fragmentation**’ and spread over 28 Member States.
EOSC: **Key requirements**

- **New modes** of scholarly communication
- **Modern reward** and recognition practices need to support data sharing and re-use
- **Innovative**, fit for purpose **funding schemes** for sustainable underpinning infrastructures
- Core **data experts** need to be trained and their career perspective significantly improved
- Cross-disciplinary **collaboration-specific measures** for review, funding and infrastructure
- Support for the transition from **scientific insights** towards **societal innovation**
- The EOSC needs to be developed as an **eco-system of infrastructures**
- Key Performance Indicators should be developed for the EOSC
- The EOSC should **enable automation of data processing** and thus **machine actionability** is key.
- **FAIR principles**
EOSC: Policy Recommendations

- P1: Take immediate, affirmative action in close concert with Member States
- P2: Close discussions about the ‘perceived need’
- P3: Build on existing capacity and expertise where possible
- P4: Frame the EOSC as supporting Internet based protocols & applications
EOSC: **Governance Recommendations**

- G1: Aim at the lightest possible, internationally effective governance
- G2: Guidance only where guidance is due
- G3: Define Rules of Engagement for formal participation in the EOSC
- G4: Federate the Gems across Member States
EOSC: Implementation Recommendations

- I1: Turn this report into an EC approved White Paper to guide EOSC initiative
- I2: Develop, Endorse and implement a Rules of Engagement scheme
- I3: Fund a concentrated effort to locate and develop Data Expertise in Europe
- I4: Install a highly innovative guided funding scheme for the preparatory phase
- I5: Make adequate data stewardship mandatory for all research proposals
- I6: Install an executive team to deal with international coherence of the EOSC
- I7: Install an executive team to deal with the preparatory phase of the EOSC
FAIR (meta)data (RDF, XML etc.)

processed data (any storage format)

Raw data (many formats)

Hard infrastructure and repositories

Analysis transformation (mostly interlinking and optimisation)

EOSC API's

FAIR transformation

FAIR download (in local format)

provenance

provenance

confirmational reading

Actionable Knowledge
5 min.
of
Conclusions
The Data Stewardship Cycle

5%
Malpractices......

‘supplementary data’

Journal Impact Factor

No data stewardship plan

Ignore Altmetrics

Obstruct Tenure

Data Experts

Knowledge Sharing Impaired
YOU ARE HERE

CWA ..........................................................
Open PHACTS .............................................
ODEX4all .............................. FDG
FAIRdom, CEDAR, EUCAT, PHT, FDG, etc. etc
FORCE11, JDDCP, FAIR
EXCOR ..........................................................
EOSC ..........................................................
ELIXIR ..........................................................


95% MS

Data Management Plans
Mandatory for Research Projects H2020 & Member States

Long-Term Data Stewardship
How to finance ESFRI’s and EBI SIB type + infra
Mainly private for reliability

Interoperability Backbones, Standards, Procedures
Mainly H2020 + ESFRI-type domain expertise

5% EC
2.51 min.
ofPHT
The Personal Health Train