HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

# Open Science – A Necessity and it's Challenges

**Hans Pfeiffenberger**

Alfred-Wegener-Institute for Polar and Marine Research,
Helmholtz Association - Germany

NRC Lithuania 2015-07-08, Vilnius


AWI

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

## Agenda

- A Bit of History in 350 A.T.
- Reproducibility and Trust in Research
- Re-Use and Progress of Research
- Current Best (?) Practise
- Summary

## Royal Society: Science as an Open Enterprise (2012) [1]

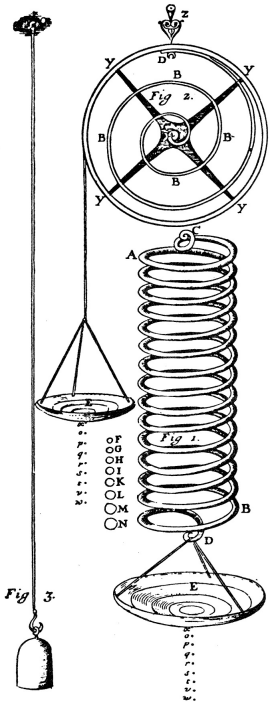
- **Open enquiry has been at the heart of science** since the first scientific journals were printed in the **seventeenth century**. …
- Science's capacity for **self-correction** comes from this openness to scrutiny and challenge.
- **RS take on data:**

  **Intelligent Openness**





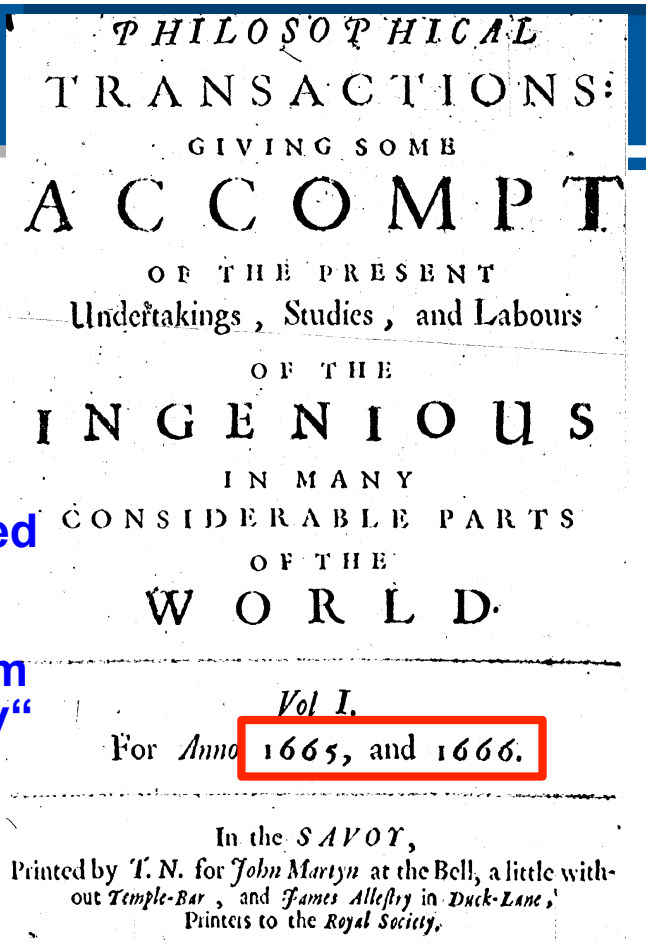

## Openess in the 17th Century




**Hooke, published his law**

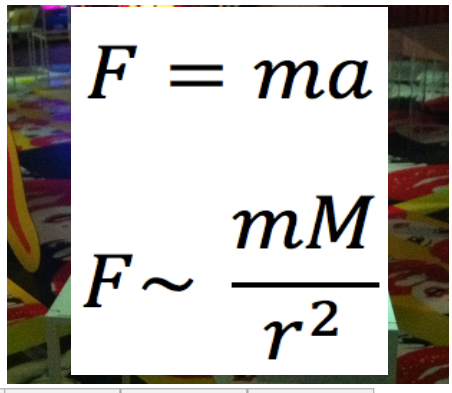
**1676** by anagram „ceiiinosssttuv“

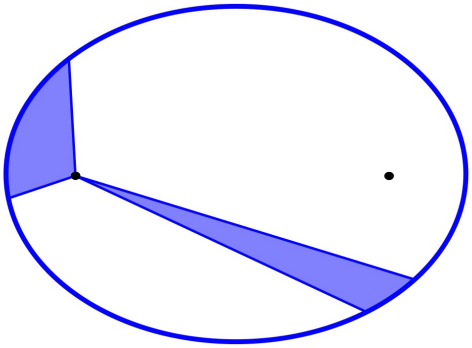
**1678 in booklet**


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

## Modern Science is based on data – since Renaissance!

- 1606 - 1618: Kepler‘s Laws
  - reduced Tycho Brahe‘s quality data
- 1684 – 1687 Newton De Motu – Principia
  - explained (!) Kepler‘s laws



| Planet | $T$ | $d$ | $T^2$ | $d^3$ | $T^2/d^3$ |
|---|---|---|---|---|---|
| Merkur | 0,241 | 0,387 | 0,058081 | 0,057960603 | 1,002077221 |
| Venus | 0,615 | 0,723 | 0,378225 | 0,377933067 | 1,000772446 |
| Erde | 1 | 1 | 1 | 1 | 1 |
| Mars | 1,881 | 1,524 | 3,538161 | 3,539605824 | 0,999591812 |
| Jupiter | 11,863 | 5,203 | 140,730769 | 140,8515004 | 0,999142846 |
| Saturn | 29,458 | 9,555 | 867,773764 | 872,3526289 | 0,994751131 |

$T$ = siderische Umlaufzeit in trop. Jahren $d$ = große Halbachse in astronomischen Einheiten (Abstand Erde–Sonne)





## Agenda

- A Bit of History
- Reproducibility and Trust in Research
- Re-Use and Progress of Research
- Current Best (?) Practise
- Summary

## Reproducibility

- "Reducing waste from incomplete or unusable reports of biomedical research" The Lancet (2014)
- "… studies of published trial reports showed that … 40–89% were non-replicable"
- Required solution is publishing and linking all text, data, software …

- Making data available on request is out (PLoS); It has been shown over and over that requests are not honoured.



HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

## Open Science Challenge #1

The Lancet article (2014) offered a long laundry list of "Components of study documentation" to be published:

1 The protocol and related documents, such as details submitted for study registration

3 Supplementary materials, such as education materials for patients, clinician training resources, and videos

7 The primary data, data manuals, and statistical code for analyses

9 Reliable and stable bidirectional linkages between all these elements

# PLoS Data Policy (2014)

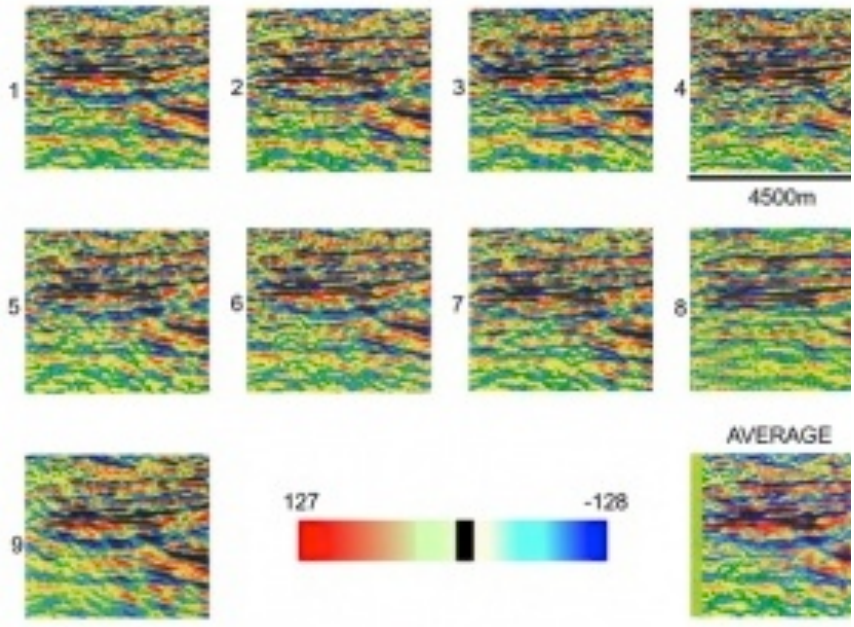
- **Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection.**
- **PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data … .**
- **If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.**





# The Dangers of Working in Closed Silos – „Does computation threaten the scientific method?“

- „using the same processed data from eight other companies, the **same algorithms** in the **same** programming **language**, using the **same input data**, just **coded independently**
- L.Hatton, A. Giordani ISGTW



# Agenda

- A Bit of History
- Reproducibility and Trust in Research
- Re-Use and Progress of Research
- Current Best (?) Practise
- Summary





TECH 10/18/2013 @ 11:11AM | 6.640 views

# IBM's Watson Now Tackles Clinical Trials At MD Anderson Cancer Center

+ Comment Now + Follow Comments

IBM continues to expand the use of its Watson supercomputer from winning Jeopardy to handling incoming call-center questions to guiding cancer doctors at Memorial Sloan Kettering to better diagnoses. Today it announced a new pilot program for Watson at Houston's renowned MD Anderson Cancer Center. The institution has been trying out Watson for a little under a year in its leukemia practice as an expert advisor to the doctors running clinical trials for new drugs.

„guiding cancer doctors ... to better diagnoses“

„an expert advisor“

# ODE „Tales“

Karin Lochte **(2011)**
(Alfred Wegener Institute for Polar and Marine Research)

“[Researchers would prefer] just one point of access to all data, which would be simple to use and ‘fool proof’.”

But she suspects it is **wishful thinking** to ask for Google-like simplicity when one looks for “chlorophyll data in the Atlantic at 200 meters depth”





# The „economic“ case: Primary data made available doubles the amount of knowledge gained

- **Hubble Space Telescope data**
- **ENCODE (“Human Genome 2.0”)**
  - “clumsy etiquette-based restrictions” … “starting to show their age and a lack of clarity”
    Birney, The making of ENCODE, Nature 2012, doi:10.1038/489049a

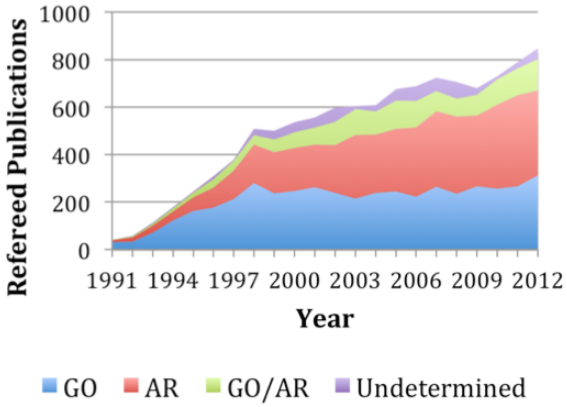



*Figure 1. Number of refereed publications based on Hubble Space Telescope data in the Multi-mission Archive at the Space Telescope Science Institute. GO = guest observer programs (papers published by the principal investigator and immediate collaborators), AR = archival research (papers published by researchers not affiliated with the principal investigator), GO+AR = papers that include both GO and AR data, and Undetermined = papers for which the origin of the data is unclear.*

# Funders‘ Policies (1)

- **NSF Post Award Requirements**
  - **Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, …**
  - **in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division/ Office for a particular field or discipline …**

**(http://www.nsf.gov/bfa/dias/policy/dmp.jsp)**





# Funders‘ Policies (2)

- **NSF Proposal Preparation Instructions (Jan 2013) Proposals / PIs’ CVs must contain:**
  - **“A list of: (i) up to five products most closely related to the proposed project; … Acceptable products must be citable and accessible including but not limited to publications, data sets, software, patents, and copyrights.”**
  - **“Plans for data management and sharing of the products of research. … no more than two pages”.**
- **see San Francisco declaration … DFG: “Quality not Quantity”**

**www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#IIC2fic**

# Agenda

- A Bit of History
- Reproducibility and Trust in Research
- Re-Use and Progress of Research
- Current Best (?) Practise
- Summary





# Status of compliance with Berlin Declaration

- We have (2015) ca.
  - 20-30% OA to articles
  - 1% to data (with disciplinary exceptions!)
- Why is appealing to researchers, citing the public good, not sufficient?
- As long as there is (perceived) risk and/or cost, but no rewards for compliance …
- Now, funders are getting out the sledgehammer
  - Netherlands: 60% by 2016 or else …

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

## 2013: CO above Troll Station, Original Data

```
BAS microwave radiometerCO profiles acquired at Troll station, Antarctica between Feb 2008 and Jan 2010
Contact: Patrick Espy, tel:  +47 73 55 10 95, email: patrick.espy@ntnu.no
-------------------------------------------------------------------------------
date [UT]:                          2009-10-19 10:44:06
apriori contribution: The profile is most reliable where the contribution from the a priori profile is less than approx.
                      Negative values are a scaling artifact and should be regarded as close to 0.

The 2-sigma systematic errors provided have been determined using perturbation calculations:

temperature error:  error induced by the temperature profile (estimated error = 5K) needed as
                    additional information for the retrieval, mainly random
calibration error:  error induced by the calibration of the measured spectrum (estimated error = 10 percent), can be sys
spectroscopy error: we used lineintensity from HITRAN 2004 with an estimated error of 2 percent, systematic
channel shape error: uncertainty due to the use of a modified channel response function in the retrieval in order to cor
                    for an instability in one of the radiometers local oscillators after 2008-08-09, systematic
Error from measurement noise [K]:  0.1510, random
Smoothing error:    This error only needs to be considered if the profiles of the BAS radiometer are compared to
                    profiles with a significantly larger vertical resolution. For such a comparison the
                    better way would be to convolve the high-resolution profile with the AVK of the retrievals.

Sum of errors:      To build the sum of certain errors they are added up as follows sqrt(error1^2 + error2^2)
-------------------------------------------------------------------------------
```

| pressure [hPa] | altitude [km] | vmr [ppmv] | apriori contribution [percent] | temperature error [ppmv] | | calibration error [ppmv] | | spectroscopy error [ppmv] |
|---|---|---|---|---|---|---|---|---|
| 0.749894 | 50.679 | 0.060 | -5.939 | 0.003 | 0.048 | 0.010 | 0.234 | 0.011 |
| 0.562341 | 53.021 | 0.065 | -20.151 | 0.002 | 0.056 | 0.011 | 0.319 | 0.012 |
| 0.421697 | 55.337 | 0.072 | -27.600 | 0.002 | 0.061 | 0.012 | 0.349 | 0.013 |
| 0.316228 | 57.609 | 0.080 | -29.442 | 0.004 | 0.067 | 0.013 | 0.298 | 0.015 |

Sun-earth Interactions

measurements carried out in order to study the dynamical context.
The data set covers the period from February 2008 to January 2010, however, due to very low CO concentrations

- Storage
- Contraints

General Information
Submission
Review

**Abstract.** This paper presents mesospheric carbon monoxide (CO) data acquired by the ground-based microwave radiometer of the British Antarctic Survey (BAS radiometer) stationed at Troll station in Antarctica (72° S, 2.5° E, 1270 a.m.s.l.). The data set covers the period from February 2008 to January 2010, however, due to very low CO



HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

## Fluxes of sedimenting material from sediment traps in the Atlantic Ocean

**S. Torres-Valdés[1], S. C. Painter[1], A. P. Martin[1], R. Sanders[1], and J. Felden[2]**

[1]Ocean Biogeochemistry and Ecosystems Research Group, Southampton, SO14 3ZH, UK

[2]Center for Marine Environmental Sciences, Universität Bre Bremen, Germany

**Review Status**

This discussion paper is under review for the journal Earth System Science Data (ESSD).

**A huge work to find, assess, collate (quality) data;**

**24 out of 43 text pages are source data references!**

**Abstract.** We provide a data set assemblage of directly observed and derived fluxes of sedimenting material (total mass, POC , PON , $BSiO_2$, $CaCO_3$, PIC and lithogenic/terrigenous fluxes) obtained using sediment traps. This data assemblage contains over 5900 data points distributed across the Atlantic, from the Arctic Ocean to the Southern Ocean. Data from the Mediterranean Sea are also included. Data were compiled from a variety of sources: data repositories (e.g., BCO-DMO, PANGAEA), time series sites (e.g., BATS, CARIACO), published scientific papers and data provided by originating PI's. All sources are specified within the combined data set. Data from the World Ocean Atlas 2009 were extracted to coincide with flux

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

# Does citation already work as an incentive?

Home
Online Library ESSD
Online Library ESSDD
- Papers in Open Discussion
- Volumes and Issues
- Special Issues
- Most Commented Papers
- Full Text Search
- Title and Author Search

Alerts & RSS Feeds
General Information
Submission

Earth Syst. Sci. Data Discuss., 5, 491-520, 2012
www.earth-syst-sci-data-discuss.net/5/491/2012/
doi:10.5194/essdd-5-491-2012
© Author(s) 2012. This work is distributed
under the Creative Commons Attribution 3.0 License.

Article | Discussion | Related Articles

**Global marine plankton functional type biomass distributions: coccolithophores**

C. J. O'Brien, J. A. Peloquin, M. Vogt, M. Heinle, N. Gruber, P. Ajani, H. Andruleit, J. Arístegui, L. Beaufort, M. Estrada, D. Karentz, E. Kopczyńska, R. Lee, T. Pritchard, and C. Widdicombe

**Interactive Discussion** — Status: Open (indefinitely extended)

AC: Author Comment | RC: Referee Comment | SC: Short Comment | EC: Editor Comment

**Reviewer: „no effort appears to have been made to engage the specialist scientists who have spent months or years at sea collecting such data. “ - not knowing that:**

**Authors asked 164 potential contributors – got answer from 13!**



HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

## 2012: Nature Climate Change, ESSD and CDIAC - interlinked

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | **Terrestrial $CO_2$ sink (positive values represent a flux from the atmosphere to the land)** | | | | | |
| 2 | | All values in petagrams of carbon per year (PgC/yr), for the globe. For values in carbon dioxide ($CO_2$), multi | | | | | |
| 3 | | 1PgC = 1 petagram of carbon = 1 billion tonnes C = 1 gigatonne C = 3.67 billion tonnes of $CO_2$ | | | | | |
| 4 | | **Cite as:** | | | | | |
| 5 | | CLM4CN | Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, | | | | |
| 6 | | HYLAND | Levy, P. E., M. G. R. Cannell, et al. (2004). "Modelling the impact of future changes in clim | | | | |
| 7 | | LPJ-GUESS | Smith, B., I. C. Prentice, et al. (2001). "Representation of vegetation dynamics in the mod | | | | |
| 8 | | LPJ | Sitch, S., B. Smith, et al. (2003). "Evaluation of ecosystem dynamics, plant geography and | | | | |
| 9 | | O-CN | Zaehle, S., P. Ciais, et al. (2011). "Carbon benefits of anthropogenic reactive nitrogen offs | | | | |
| 10 | | ORCHIDEE | Krinner, G., N. Viovy, et al. (2005). "A dynamic global vegetation model for studies of the | | | | |
| 11 | | SDGVM | Woodward, F. I. and M. R. Lomas (2004). "Vegetation dynamics - simulating responses to | | | | |
| 12 | | JULES | Clark, D. B., L. M. Mercado, et al. (2011). "The Joint UK Land Environment Simulator (JULE | | | | |
| 13 | | VEGAS | Zeng, N., A. Mariotti, et al. (2005). "Terrestrial mechanisms of interannual CO2 variability | | | | |
| 14 | | | | | | | |
| 15 | | Terrestrial CO2 sink as a residual | | **Models** | | | |
| 16 | Year | of the global carbon budget | | CLM4CN | HYLAND | LPJ-GUESS | LPJ |
| 17 | 1959 | 0,42 | | 0,79 | 2,02 | 0,42 | -0,83 |
| 18 | 1960 | 1,14 | | 0,75 | 1,53 | 1,16 | 0,81 |
| 19 | 1961 | 1,20 | | 0,30 | 1,71 | -0,07 | -0,55 |
| 20 | 1962 | 1,76 | | 0,79 | 2,37 | 1,25 | 0,57 |
| 21 | 1963 | 1,72 | | -1,20 | 1,81 | 0,26 | -0,37 |

**GLOBAL CARBON ATLAS**

*The Global Carbon Atlas is a platform to explore and visualize the most up-to-date data on carbon fluxes resulting from human activities and natural processes. Human impacts on the carbon cycle are the most important cause of climate change.*

[7]


**Funded by BNP Paribas**

**Implemented by WeDoData**

**(„data journalism“)**





# Open Science Challenge #2

- **Trusted Environments for Protected Data**
  - **Science Europe Roadmap (2013):**
  - **“Identify where protected environments, or ‘safe havens’, for data are necessary, and promote the creation of policies, technical concepts and, ultimately, safe infrastructure for such cases.”**
  - **patient (health) data and proprietary data, …, are of crucial interest … balance of all ethical considerations … ensure trust amongst all stakeholders, including the public and researchers … privacy, confidentiality and consent are respected …**

RDA
RESEARCH DATA ALLIANCE

## Data Citation WG
## Making Dynamic Data Citable

https://www.rd-alliance.org/working-groups/data-citation-wg.html

Andreas Rauber, Ari Asmi, Dieter van Uytvanck

- **Goal**:
  - Ensure cite-ability of data at arbitrary levels of granularity, particularly when data is large-volume and dynamic
  - Machine-actionable, variety of data types





# Agenda

- A Bit of History
- Reproducibility and Trust in Research
- Re-Use and Progress of Research
- Current Best (?) Practise
- Summary

## The Imperative(s): Pro and Contra

- **Ethics: e.g., protecting privacy vs. saving lifes; protecting Ph.D. students vs. taxpayer Euros**
- **Good scientific practise: Openness and reproducibility at the heart of research (ethical limits apply…)**
- **Law: copyright vs. freedom of information acts vs. data protection vs. …**
- **Contracts and licenses: funders, project partners, publishers,…**

**=> Develop practises "easy" to comply with**

**=> Don't sign contracts without some serious thinking!**





## The Status Quo

- **Socio-cultural change is on the way**
  - **Need for change/quality is recognized (R.Soc./Lancet)**
  - **PLoS, Nature, … data policies**
  - **NSF/EC "5 products" rule offers "rewards" and the way out of the metrics dungeon**
- **"Technical" challenges remain, e.g.**
  - **Persistent repositories for computer code etc.**
  - **Quality assessment for data, software, "protocols" ...**
  - **Bidirectional linking of everthing open … (b-LEO)**
  - **Trusted environments for protected data …**

# ToDo‘s (1)

- Researchers need to develop (new) best practises
  - What to share when (Embargo timing)
  - No “legal tricks” (licenses) to enforce good scientific practise
  - Identify best repository and dissemination strategies (just as they do with journals and publishers)

=> have a plan!

  - Develop skills - and careers! - in data management and scientific programming (“Data Scientist”)

=> role of universities!!



HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

# ToDo‘s (2)

- Funders need to develop rules for funding and assessment
  - Require Open Products (articles, data, software)
  - Require data management plans
  - Abandon metrics, require 5 products (per person)
  -
  - Fund (new) information infrastructures:
    - In part (semi-)permanently (as with libraries)
    - In part through projects’ data management funding
    - In part through competitive R&D funding for innovation