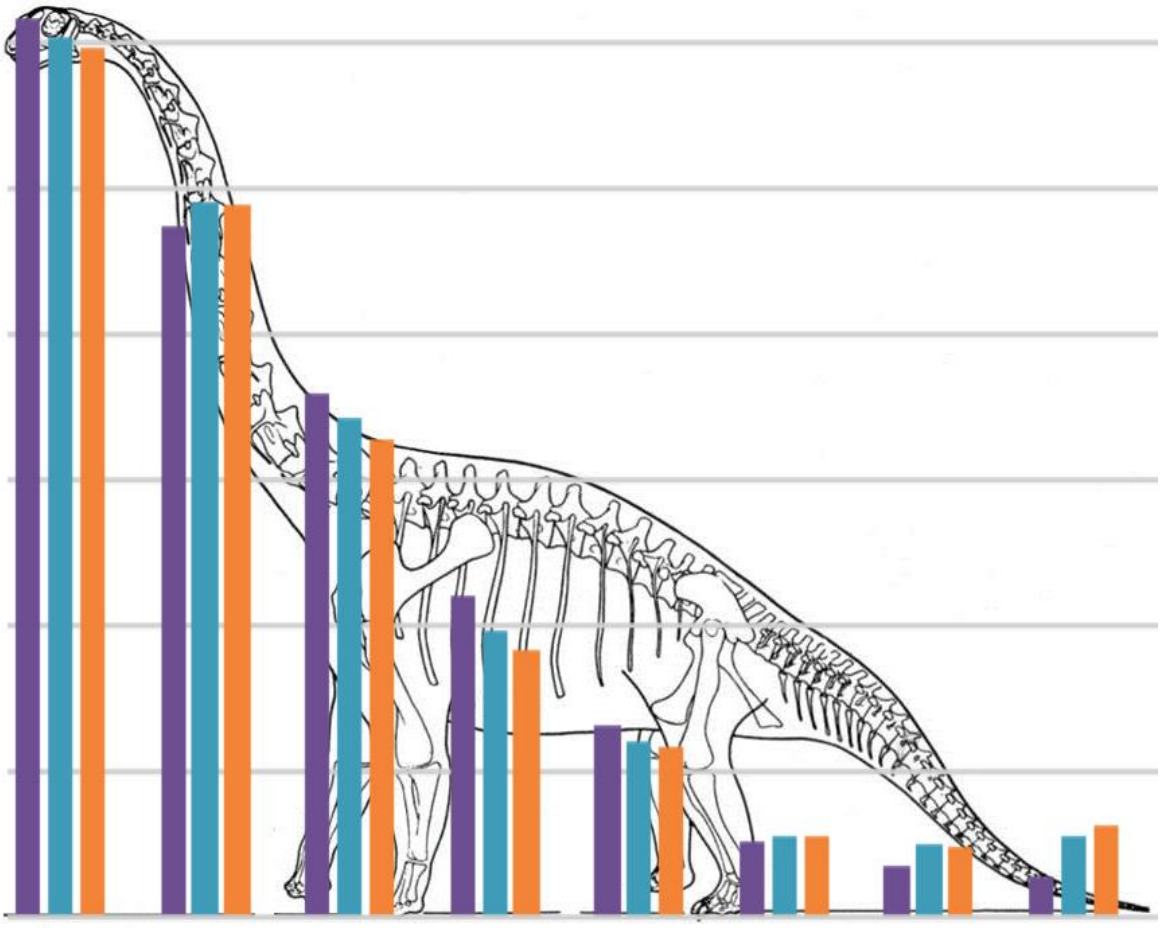
PLAN-E Platform of National eScience Centers in Europe

The long tail of Science and Data

A PLAN-E workshop in the context of the EOSC

WORKSHOP report

PLAN-E Plenary Paris, 19-20 April 2018



1 Contents

Contents

2 About the organisers

2.1 About PLAN-E

PLAN-E (plan-europe.eu) is the Platform of National eScience Centers in Europe. It consists of representatives of the major centers in Europe concerned with eScience, including Data Science and Computational Sciences, and which have -by rule or *de facto*- a national or regional role in their home country.

PLAN-E is based on voluntary co-operation and knowledge sharing, in particular regarding the further development of eScience, the status of people working in the eScience domain, the proliferation of knowledge about the impact of (big) data, the importance of proper data management and the relation to proper academic conduct and so on. But also the more traditional topics, like Computational Science, involving modelling, are covered by most of the PLAN-E member organisations.

PLAN-E Plenary meetings are hosted twice annually in turn by different countries. The meetings involve mostly two or three different topics addressed in workshops.

2.2 About the Workshop

The Long tail of Science and Data was selected as a topic for a workshop by PLAN-E, because of a growing concern that the attention for the Big Sciences and Big Data in European and national policies leads to an unbalance in available resources and support for the vast majority of science domains and the larger part of different data sets.

Based on the insights of the PLAN-E members and factual information about the numbers behind the long tails, PLAN-E arrives at some conclusions it seeks to share with the policy makers and service providers in order to create greater awareness of this matter, now that through the EOSC basically all scientific disciplines are encouraged to open up and share data.

2.3 About the report

This report is the direct reflection of the discussions during the PLAN-E Plenary meeting in Paris, April 19-20 2018 and the introductory presentations on the topic of the Long Tails of Science and Data. The examples derive from daily practice of persons and organisations directly involved in providing public services in the fields of data, compute and ICT-infrastructures. The observations and conclusions are direct translations of the discussions in different workshop sessions of the participating members and invited speakers.

3 Executive summary

1) PLAN-E put the topics of the long tail of data and the long tail of science on the discussions agenda, because these topics are felt important while discussing and establishing the EOSC. And because these topics are considered underrepresented during the ongoing implementation discussions on the EOSC.
2) The devil is not only in the *de*tails but also in the *long* tails: "Big" as in many or large is the easy part of data, "Big" as in complex, varied or dynamic is the devilish part of data. And reaching out to those disciplines that are visible as big sciences or big science projects is simple compared to reaching out the complete pallet of sciences that need to get involved.
3) It must be recognised that only a small fraction of all scientists in Europe use national of European ICT-infrastructures for data or computing. France reports less than 10%, a broad Netherlands survey (2016) showed similar figures when asked about acquaintance with European ICT infrastructures. It must be avoided that European ICT-infrastructures are only for the happy few by lack of awareness with the rest.
4) Intensifying and targeted user support does show direct effect in the usage of external ICT-resources (experience from France). Essential therein are:
    a. Ease of use of the e-services
    b. Training
    c. User support.
5) One step away from the domains where Terabytes are the norm, the most common data sets are of the nature small (5-20 Mbytes) to very small (<2 Mbytes). Data of the "common researcher[1]" larger than 1 or 2 Gbytes are rare (like less than 3%).
6) The ideal of reproducibility in science, for domains where it applies (where experimental and measured data, survey data and data from modeling play a role) is hardly achieved in practice and warrants an open data policy.
7) Inclusion of *all* major disciplines in the set up and operational scope of the EOSC is essential.
8) eScience introduces a way of working in research that can practically address the issues of inclusion (of all sciences) awareness creation and disciplinary cross over.

4 Recommendations

1) Extend the operational scope of the EOSC to explicitly address the long tail of science and the long tail of data.
2) The EOSC should not take the route of least resistance, by starting with already established infrastructures and research communities that already operate at a European level.
3) Strengthen the activities to proliferate knowledge and create awareness about European ICT-infrastructures, including the EOSC.
4) Create and support practical efforts to dedicate attention to modern academic good practices, involving data handling, proper coding, multidisciplinary research, privacy and security in early scholars education.
5) Design within the EOSC national support structures that can advise on the use of ICT tooling, resources, infrastructures, data handling, FAIR principles, etc. Without such support the broader usage and broader data sharing will not come about (accept for the "happy few").

[1] The term "Common researcher" is admittedly very vague, but refers to researchers not specifically working in the domains of particle physics and astronomy, (gen-)omics or climate research.

5 Background to the workshop

The subject The Long tail of Science and Data was suggested by the local host, CNRS, after discussions with Federico Ruggieri. It concerns a serious matter, because the attention at the European level for Big Sciences has been quite dominant. For understandable reasons: some sciences cannot be conducted innovatively at any national level, but only by combining all European forces and funds. However, now that the EOSC is being established, *all* sciences in *all* disciplines are to take part in this endeavour and open up and share their data cross disciplinary. It is in this context that we encounter a new reality, where the issues with the long tails may be dominant and disproportionately complex. A good reason for PLAN-E to address this topic and try to distil some advice from the facts and experiences and discussions.

6 Plenary contributions

Two plenary contributions introduced this topic: "Going the extra mile", by Vincent Breton and "The long tail of research data" by Peter Doorn. Both touch upon the subject by pointing out that most science is done outside the scope of the "big sciences" and most complexity in data management does not derive from their sheer volume. Reaching out to science at large is labour intensive and mismatches between what funding agencies want to achieve with requiring research data management plans and so on and the basic understanding of researchers who should deliver are more common rather than incidental. Qualitative research in many domains hardly encounters "data" as in the context we have "data" in mind if we discuss data management planning.

The long tail of data consists of very many small data sets indeed, of which the administrative efforts to keep them findable may easily be heavier than keeping their volume stored. Anyway, the issues at hand differ dramatically from the problems with keeping Big Data or the sets of large homogeneous data forms from Big Science projects.

These matters were discussed at the PLAN-E Plenary in Paris, introduced by the following plenary introductory contributions.

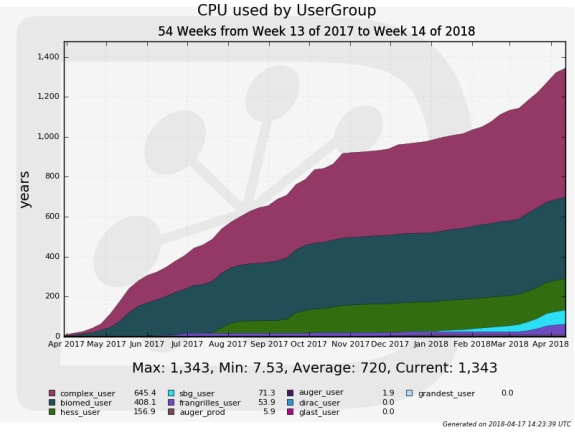
6.1 The Long Tail of Science (Vincent Breton)

One indication of what the long tail of science entails is to refer to "scientists (in Europe) that are not using data and computing infrastructures". For France, with 156000 researchers (2017) including 77000 PhD students, an estimated less than 10% is using these infrastructures. But the long tail of science also concerns scientists outside Europe in developing countries.

Year	CPU hours consumed by french certificate owners	CPU hours contributed by France and used by EGI	Share of France resources consumed by CERN
1-3/2018	185 M	813 M	84%
2017	676 M	3,2 G	83%
2016	760 M	2,6 G	77%
2015	850 M	2,5 G	76%
2014	700 M	1,5 G	72%

This table shows that about 80% of all French grid resources are being used by a single party "CERN", so all other users together make up for only 20% of the resources used.

Another figure shows significant growth over time in used CPU capacity, due to targeted user support.



With a distributed approach (through "salesman"), first at country level, then at regional level. Users could be reached and encouraged in their using the e-infrastructure. Three elements turned out to be key prerequisites:

- Ease of use of the e-services
- Training
- User support.

So, targeted support seems an effective way to encourage scientists beyond the traditional circle of users to make use of e-infrastructures that otherwise would have been outside their scope of methods.

Challenges faced were and still are:

- Cost: who covers what, how to get there, difficult to assess and compare;
- Limited interest from the e-infrastructure stakeholders and funders. They leave too much to the user or other local help desks as if it were not their concern.

It is a matter of experience that EGI competence centers are ESFRI-centric (this is a personal observation). There are also only limited places to share experience with European e-infrastructures. A representative user forum, including domain prospects is missing and the information that there is is centered around the different facilities and therefore also rather technical in nature.

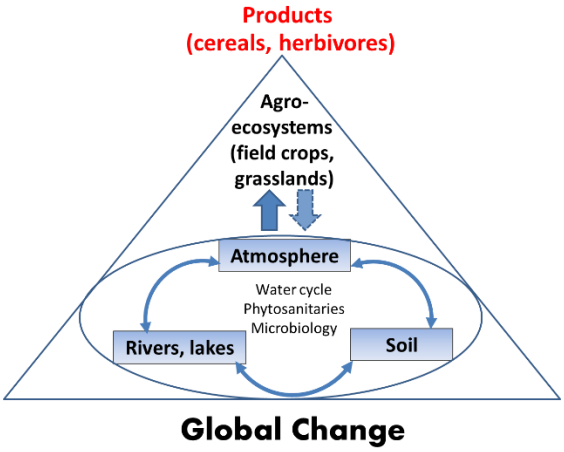
EGI owes much of its success to the LHC, not to ease of use or their addressing large user communities. HPC should be able to have more scientists enjoy its capabilities, but it doesn't yet. So the expectation for the EOSC is, that it again will serve only the inside communities (those that are involved anyway already) unless it responds to specific needs or addresses scientific challenges.

6.1.1 An example

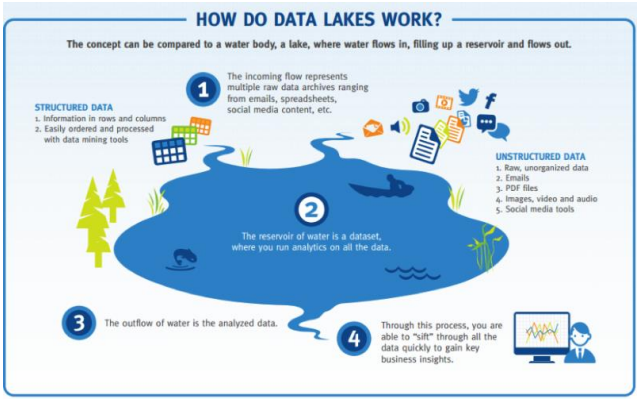
An example from practice comes an agro ecosystem, where data do play an important role. So, experience was obtained while trying to build an environmental cloud for agriculture at a regional level. The circle of life for that domain is depicted in the figure below.

From building this infrastructure, a few conclusions can be drawn: there is

- Need for flexibility in data sharing policy
- Need for flexibility in data structures
- Need for friendly interfaces for data exploitation.



In as far as data are involved, the data infrastructure can be seen as a "data lake".

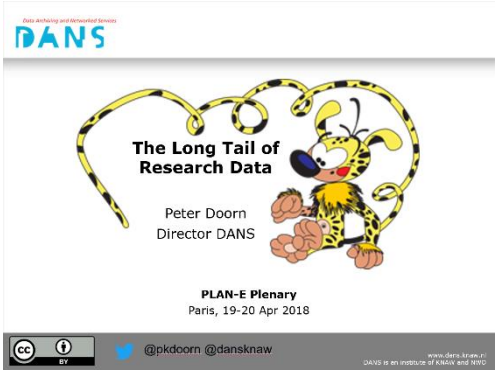


6.1.2 Conclusions

The experiences gained in France, regarding the proliferation of the use of national and international (mostly European e-infrastructures),with additional hands on experiences from a developing country (Vietnam; not re-reported in this extract from the original presentation) are the following:

- Reaching out to the long tail of science is possible, provided it is made to a priority and implemented at a proper granularity level;
- The EOSC needs emblematic science use cases
- The EOSC needs a clear role for users (from all domains) in the EOSC governance (and surely beyond the ESFRI project reps)
- The EOSC should be open to the long tail of science outside of Europe, supporting developing countries.

6.2 The long tail of data (Peter Doorn)



Peter Doorn, director of DANS (Data Archiving and Networked Services, an organization shared by both the Netherlands Organisation for Scientific Research and the Royal Netherlands Academy of Sciences), addressed four topics in his contribution:

- Data big & small: Big Data/Long tail definitions: the 4 V's and methodological challenges
- How realistic or false are the promises of data intensive research? Replication crisis?
- The "4th paradigm" of data intensive science includes danger to mix up statistically significant with meaningful results
- Volume and Variety of data production in the humanities and social sciences.

The four V's in data science (some have 6 or 7 V's), most relevant in Humanities and the Social Sciences are:

1. Volume, the scale of data, individual files to collections
2. Velocity, analysis of the data flow, rate of change, etc.
3. Variety, the different forms "data" can take (photo, sound, text, numbers, mixed)
4. Veracity, the trustworthiness of the contents of data or the trustworthiness of the process leading to the data or their subsequent handling.

Typically for the Social Sciences and Humanities are the following characteristics:

- Currently >40,000 "data sets" in DANS archives
- Data set: collection belonging to a research project
- Every data set consists of 1 or more data files, up to 25,000+
- Most data sets are small (96% < 1 Gb)
  - For example, the entire population census of 1960 (>11 million records) is about 500 Mb
- Total number of data files about 4.5 million
- Challenge: data management operations on the whole archive -- slow and problematic
  - Mass conversions (e.g. thumbnails of images)
  - Data integrity control (checksums)
  - Compressing the data
- Trend: "data publication package" belonging to a publication as an extract of the raw & processed data

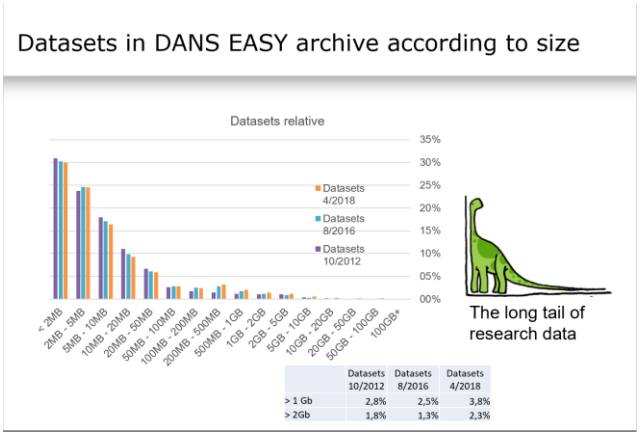


Figure 1 Breakdown of data sets in EASY according to size

EASY is one of the data depositing systems that DANS serves. It requires from users to add meta data to their data and delivers a DOI in return for later reference.

Promises: see the projections made for Big Data

Domain	Datasets	% Datasets
Behavioural and educational sciences	1.351	2,3%
Economics and Business Administration	235	0,4%
Humanities	38.363	66,3%
Interdisciplinary sciences	5.609	9,7%
Law and public administration	817	1,4%
Life sciences, medicine and health care	6.593	11,4%
Science and technology	162	0,3%
Social sciences	4.708	8,1%
Total	57.838	100,0%

Note: including ca. 7000 data sets in more than one domain

Figure 2 Breakdown of data sets in EASY according to discipline, per April 2018

Of course "data" did not (and will not) replace conceiving new theories or conducting experiments. Rather it adds to the repertoire of tools one can use to establish or falsify new theories. Also Artificial Intelligence will help in this respect, through training on big data sets. But also as a tool, not a replacement of theory.

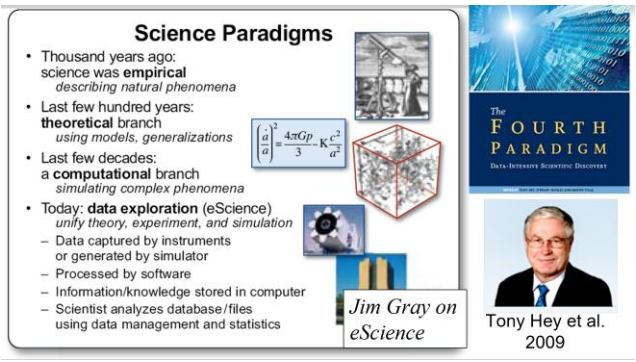


Figure 3 The "fourth paradigm" concept

The claim for "data" as 4th paradigm may still hold. Again as an extension of the scientific methodology. Collecting data is in itself nothing new, as it is just what researchers starting to do a few thousand years ago. It is the combination of volume and AI that adds to the scope of scientific methodologies. But it remains prerequisite to properly pre-define one's research methodology in order to prevent unsustainable claims on the basis of sloppy data-based research. This became all the more apparent in Nature's re-research effort:

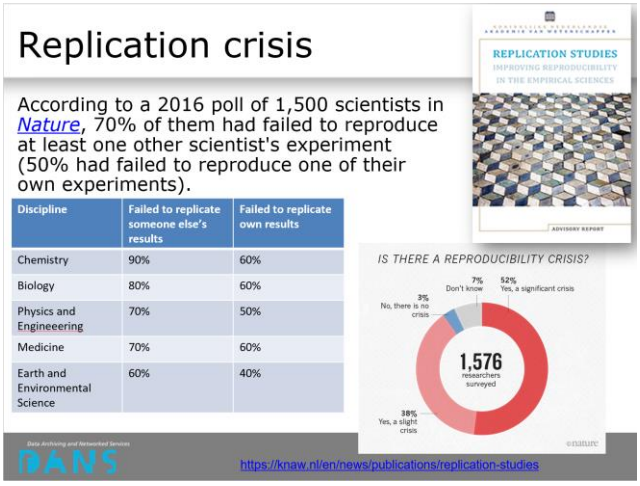


Figure 4 The replication crisis, as found by *Nature*

Data production in the Social Sciences and Humanities comes most from the following sub-domains:

Humanities:

- archaeology: excavations and surface surveys
- history and cultural studies: digitized/transcribed
- archival sources
- library holdings (books and other texts with images)
- museum holdings (artworks, images and descriptions)
- linguistics: text, human speech (audio/video)

Social and behavioral sciences:

- social sciences:
- social surveys
- qualitative interviews (audio/video+transcriptions)
- censuses and registration data
- psychology: data from experiments

Big data -to the measures of the domain- come from:

Born digital

- administrative processes: government administrations
    - taxation, population registers, school data, traffic flows
- commercial processes: business and financial transactions
    - banking, sales (goods, real estate), stock exchange
- socially produced: social networks

    - Twitter, Wikipedia, Facebook, YouTube, Flickr
- personal devices: GSM, GPS
- simulation data

Mass digitization

- images
- OCR of images: text & numbers
- audio-visual

Data generated by individual people tend to be small and by collaborative groups of modest size. Data generated by social processes, transactions, administrations and personal devices tend to be BIG. Data preserved from the past tend to rather big and fuzzy and complex. There is a small but growing number of "big data" projects in SSH. The uptake of HPC will remain modest

- Millions of digitized books ("Culturomics")
- Analysis of twitter feeds and social media:
- Sentiment analysis to predict markets and economic trends
- Linguistic analysis
- Traffic flows using GPS.

Points for discussion are:

1. The need to acknowledge that in all domains most researchers still work with modest volumes of data
    1. Investments need to reflect this
    2. Plan-E also seems to have favored Big Data above Small Data
    3. Uptake of HPC, Grid, etc. will will remain low
2. Do "data publication packages" represent the original (raw and processed data) in an acceptable (FAIR) way?
    1. Pro: Publication packages contain valuable additional information, including syntax/code
    2. Con: This is an escape not to make available the actual data
    3. 4 V's require very different methodological and technical solutions; focus of e-science and data science has been on volume & velocity; little attention has been paid to variety & veracity challenges
    4. Data-centric research contributed to the replication crisis
    5. Long-tail data can be curated, managed, archived and made accessible by repositories for small to modest size data; facilities for big volumes need to incorporate trust functions and get certified separately

Long tail data remains typical for the Humanities and Social Sciences. But fruitful it is, combining data from different sources. We few examples are given. Collaborative work: bringing together data from many scholars

1. Historical shipping
2. Digitized censuses
3. Global inequality
4. Holocaust studies
5. Dendrochronology

Ad 1. Bringing together shipping records from projects over the decades: South Chinese Sea Trade (1681-1792); Dutch-Asiatic Shipping (1602-1795); Climate of the World Oceans (weather observations from ships' logs, 1750-1854), Atlantic Connections, Trans-Atlantic Slave Trade, etc.

Ad 2. Census digitization projects since 1996. Collaboration with Statistics Netherlands, 40,000+ pages of tables turned into numbers, Images of the original source books, Up to 60,000 users per year.

Ad 3. Data collection from thousands of sources from all over the world by hundreds of specialists, solving massive problems of data interpretation, cleaning, linking, harmonization, comparison…

Ad 4. Holocaust Researchers Catalog 42,500 Nazi Ghettos, Camps; Numbers Are 'Unbelievable'.

Ad 5. Dendrology.

Private sector in The Netherlands (6000 BC-present):

- 2000 research projects
- 20.000 measurement series of 13.000 trees (60% dated)

Private sector and universities in Germany:

- Archaeology: e.g. Dorestad
- Cultural heritage: many objects from The Netherlands and Flanders
- Architectural history: North and East NL, Amsterdam

With this exposé and with these examples, it is to be clear, that the long tail of data holds a serious complexity but also an abundance of very worthy, useful data and a broad spectrum of information. It is to be treated no less careful than the Bigger data that tend to attract most of the attention. All the more in the context of the EOSC, this long tail requires and deserves its place and care.

7 Elements from the discussions

7.1 A matter of serious concern

Broadly shared in the PLAN-E community is the observation of a clear bias in the EOSC discussions so far towards Big sciences and Big data. ESFRI-type of projects and examples seem leading in the visions on the EOSC. This bias is, according to the participant's opinions, not related to the value, importance or projected impact of those Big sciences and data, but merely to the fact that some research just happens to require big equipment or produce mass amounts of data that can only be realized or handled by sharing the limited resources at European, rather than national or institutional level. Whereas PLAN-E members fully support such major infrastructures being funded and organized, and most are member of one or more of those projects, it is felt that the *EOSC should be different* in that respect and not just be implemented to support the Big stuff.

The question why this would be a matter of concern is answered by the claim that in fact *the majority of scientific disciplines and research projects do not fit the profile of Big science and do not deploy or produce Big data*. But at the same time, the accessibility, re-usability, long term storage and findability matters apply or should apply to all scientific output from all disciplines and to and from all European member states if not internationally. And this requires the EOSC to be known to all researchers in all disciplines, be accessible and usable to all researchers.

7.2 Solution directions

The eScience community can be very instrumental to achieve this, because the many smaller research groups, communities and individual scientists do not generally have the means and now-how to find out how they can make all the available resources work for them. eScience communities can be formed by combining research groups qualifying for Big science activities with research groups that do not, for example by taking just "data" as a common denominator for cooperation, or "complexity". This as a means to get a cross-over of knowledge and experiences from domain to domain.

While not being blind for the fact that an eScience Community , as PLAN-E is, would be biased as well advising eScience as a solution to many such problems, it is yet claimed that stimulating interdisciplinarity in research and invoking innovative ways to deploy ICT in all its facets is the way to go. And it happens to be called eScience.

A serious and most robust contribution to address these matters in the future is through *education*. The way in which eScience thinks about doing research can be taught during the early stages of higher education. And it includes topics such as good academic practices regarding data (FAIR), compute (proper coding), proper referencing and crediting and of course about the very existence of infrastructures and resources beyond the own premises: regional, national, European, global.

7.3 Top down versus bottom up

For a practical realization of a major effort like the EOSC is, the PLAN-E community is in favor of a

blueprint-type of approach. Getting the community to support an approach where participation is voluntary and organically growing towards a final goal, but governed in a way, that milestones and deadlines can be set, with an underlying concept that is agreed upon etc. However, for the participation at large of all research communities and disciplines, the EOSC should also encompass grass roots developments and bottom-up activities. RDA is an example a bottom up activity, but there one sees that the step to participate is too high for certain categories of researchers. Grass roots activities focusing on young scientists may proof fruitful on the long term.

7.4 What PLAN-E can do

It all starts with awareness. PLAN-E members are aware of the long tails and can and will help promote awareness. PLAN-E, like eScience is neutral towards big, medium of small sciences, data or compute intensity. PLAN-E can help, at the local and national level, to promote interdisciplinary co-operations, so that the have-s and have-nots and the ICT or data knowledgeable can mingle with the ignorant.

Identifying the -sometimes few- experts in different (sub-)disciplines that do have the awareness and knowledge and support them in taking the lead for their discipline. For helping to translate FAIR principles into useful tools for their field, help harmonising Research Data Management Planning and getting access to EOSC-involved ICT-infrastructures, including data resources and storage facilities and tooling for meta data etc.

PLAN-E can and does actually promote FAIR principles across all disciplines and learn about the availability and use of ICT-infrastructures. Promoting FAIR principles, supported by increasingly more stringent requirements by funding agencies regarding data and software, seems an obvious route to get in touch will all disciplines and not just those associated with Big science.

PLAN-E members can help, locally and nationally, to translate technocratic policies into science applications and innovation and vice versa articulate the needs from the long tail of science into concrete infrastructural and policy requirements.

8 PLAN-E's Formal EOSC statement of support

This statement is the result of the discussions continued during the 7th Plenary meeting of PLAN-E in Oxford, 10-11 October 2017. It is repeated here because it fits the topic of the long tails suitably.

Formal PLAN-E statement of support for the European Open Science Declaration

The plenary of PLAN-E confirmed during its meeting in Oxford, October 10-11 2017, that PLAN-E supports the EOSC Declaration.

In detail PLAN-E can and will:

1. Help proliferate and implement FAIR principles for data in all disciplines and help refining its very definition in different application domains. Furthermore PLAN-E will help:
    a. extending the working domain of FAIR principles to software,
    b. promote FAIR principles within our communities, and
    c. implement FAIR data principles within our own institutes;
2. Upon invitation, act as an active stakeholder in the EOSC governance structure by translating scientific requirements into advice for practical services and physical components in the infrastructure;
3. Provide hubs, via its membership, for collaboration among the eScience community and provide knowledge on the deployment potential of ICT and available e-Infrastructure for domain researchers, largely at National level;
4. PLAN-E, based on understanding both vertical needs and horizontal demands for e-Infrastructure services, will be able to:
    i. identify gaps in service provision,
    ii. identify research profiles for domains,
    iii. oversee future needs;
5. Help harmonizing, within and among disciplines, Research Data Management Planning across institutions, from faculties/universities to national funding agencies and European funding organizations:
    i. in co-operation with national organizations for scientific research and especially eScience,
    ii. in communication with Science Europe,
    iii. by communicating best practice in data stewardship to implement DMPs;
6. Advocate the importance of Software Sustainability next to Research Data Management planning.