REPORT

PLAN-E Workshop
Data in Depth
Lugano (CH), 8 September 2016

ed. Patrick J.C. Aerts

On September 8th 2016, as part of the 5th plenary meeting of PLAN-E, a workshop was dedicated to the triangle: data-infrastructures, data-services and data (user-) requirements. The three angles to this discussion were introduced by:

- Damien Lecarpentier (CSC) - On EUDAT
- Peter Doorn (DANS) - On Data services
- Jason Maassen (NLeSC) - On user requirements
- (CSCS) - On the CSCS vision on Data services and infrastructures

1 Introduction

For the efficacy of Data Infrastructures it is essential that infrastructures, pan-European services, national and local services are well integrated and fit current user requirements as well as requirements for long term preservation. This is represented in this basic triangle:

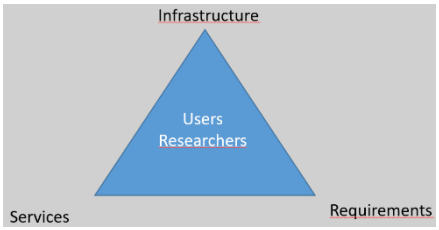

Figure 1 Data interaction model

EUDAT represents in this ecosystem “the infrastructure”. EUDAT is a Pan-European, FP7/H2020 supported co-operation between thirty-five institutes across Europe. But it is relevant noting that not all European countries are involved and surely some countries are more involved than others.

Figuur 2 EUDAT participation across Europe

EUDAT relies on active delivery of services and facilities by its members. It focusses on delivery to major European endeavors, such as the ESFRI-projects[1] and operates in practice at the back ground. This means that it does not target individual researchers, but rather acts through its member organisations, to whom it provides an extension to their operating domain so as to enlarge it, make it more robust and helps integrating their services. More on EUDAT will be discussed later in this report.

CERN	CH/EU
DKRZ	D
Umweltbundesamt	D
Uni. Tübingen	D
Forschungszentrum Jülich	D
MPI für Meteorologie	D
KIT	D
GFZ	D
MPCDF	D
EMBL	D/EU
BSC	Es
CLARIN	ESFRI
CERFACS	F
CINES	F
LIBER	F/EU
eScience Data Factory	F
CSC	Fi
University of Helsinki	Fi
GRnet	Gr
CINECA	It
INGV	It
DANS	NL
KNMI	NL
SURFsara	NL
Uninett/Sigma2	No
PSNC	Pl
BioSense	RS
SNIC	Se
KTH	Se
LUND University	Se
STFC	UK
UCL	UK
Trust-IT	UK

Figuur 1 EUDAT Participants per country

DATA services are at the core of what makes data valuable assets. Without being able to find data through meta tags, have them securely stored and managed and have them retrievable, data would be as good as lost. Value adding organisations, such as DANS (Data Archiving and Networked Services) provide such services at local and national level and may, for example through EUDAT, deliver these services European-wide.

Parties like DANS do provide services directed to institutions (universities, government, agencies and publishers) and the individual researcher/user, from low level services to store data files with their meta data tags to high level archiving and long term storage. Such organisations are professional data service providers and may or may not have their own physical storage infrastructure. DANS is an example of an organization that provides basically all data-related services but has all of its physical data stored at a professional data center, in this case VANCIS (until recently a daughter enterprise of SURFsara), through SLA’s.

A representative of the user community, in this case NLeSC, shows aspects of user requirements. As NLeSC basically covers all science domains, NLeSC may be an a-typical user, but in contrast will encounter all sorts of data types researchers will be faced with. Typical Data Shapes are:

- Sensor streams
  - Astronomy, high energy physics, ecology
- Structured Data (grids, point clouds, ...)
  - Climate, meteorology, hydrology, archeology
- Databases
  - Life sciences, chemistry, ecology, materials science, archeology
- Linked Data
  - Life sciences, chemistry, humanities
- Unstructured Data (text, images)
  - Humanities, forensics

And of course one needs to access, analyze, store and access, archive and share and reference data.

[1] In particular: CLARIN, ENES, ELIXIR, EPOS, ICOS and ELTER

Moreover: Data need to be or become "FAIR": Findable, Accessible, Interoperable and Reusable. And this is neither trivial nor obvious, usually.

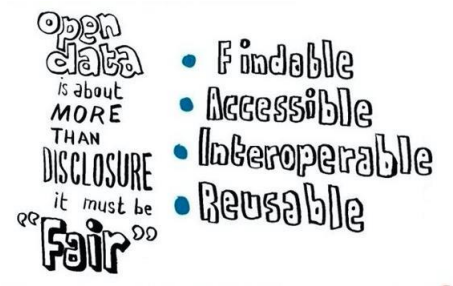

Figure 2 The FAIR Guiding Principles for scientific data management and stewardship. Mark D. Wilkinson et al. http://dx.doi.org/10.1038/sdata.2016.18

Finally, in short, there are typical discipline-specific issues, such as:

- Formats and Accessibility (life sciences humanities)
- Volume (astronomy, high-energy physics, climate)
- Privacy (medical sciences, archeology(!))
- and more.

2 Discussion

The workshop discussion topics covered the following items:

- Short, medium and long term storage/archiving
- Integration issues/overlap
- Gaps/deficiencies
- Future of EUDAT.

The presented EUDAT portfolio of activities is listed in figure 3.

Service	Function	Status	Individual Researcher	RI/ Community Manager	Service Provider
Data Discovery					
B2FIND	Multi-disciplinary joint MD catalogue	Active	✓	✓	
Metadata Catalogue	MD extraction, MD store, index	Under develop.		✓	✓
Data Hosting, Registration & Management & Sharing					
B2DROP	Cloud storage, sync & exchange	Active	✓	✓	✓
B2SAFE	Policy-driven data management	Active		✓	✓
B2SHARE	Repository for sharable digital objects	Active	✓	✓	✓
B2HANDLE	Policy-based prefix & PID management	Active		✓	✓
Data Type Registry		Under develop.		✓	
Data Access, Interface & Movement					
B2ACCESS	Federated multi-protocol IAM	Active	✓	✓	✓
Generic API	Common data interface service	Under develop.	✓	✓	✓
B2STAGE	Data staging service CDI → ext.	Active	✓	✓	✓
Subscription	Data transfer subscription	Under develop.		✓	✓
Consultancy					
Training	on services & data management	Active	✓	✓	✓
Consultancy	on licensing, certification, data privacy, data system design	Active	✓	✓	✓
Helpdesk	Support and enabling	Active	✓	✓	✓
Operations					
Service Hosting	PaaS, IaaS, SaaS	Under develop.		✓	✓
Monitoring	Availability & reliability monitoring	Active		✓	✓
Accounting	Storage & Data Usage Reporting	Under develop.		✓	✓
SLC Management	Service Portfolio & Catalogue	Active	✓	✓	✓
Coordination	Project Implementation, Service & Resource Provisioning	Active	(✓)	✓	✓
Site Registry	Site, Service & Service Groups	Active	(✓)	✓	✓

Figure 3 EUDAT portfolio of activities and services

2.1 Short, medium and long term storage/archiving

A strong definition of short, medium or long term for storage and retrieval does not formally exist. There may, however, be good reasons to start agreeing on these terms for data preservation.

As reference it may help that the term for storing data that belong to published scientific papers is set to 10 years after publication (Several Codes of Conducts for research and funders set this requirement). This may define the limit of medium storage. So, long term then means longer than 10 years (and in principle indefinite). Short term then refers to day-to day accessibility for five years (yet an arguable time frame), but at least during the active research phase.

The distinction between these time frames is important, because they require different levels of curation, infrastructure (facilities), services and funding models.

In the short term domain quick access, delivery and sharing are the main elements for current data.

In the medium range some elements of curation already come into play, because the technological innovation cycle may be only four years. After four years hardware formats (tape, disc) are becoming obsolete already and refreshing storage to new media becomes necessary. Also original creators may have changed jobs or become otherwise less traceable.

For the long term, for those data that pass the relevancy criteria (according to selection criteria, which have to be specified too), real trustworthy curation and preservation are required. The goal is to have these data readable, findable and accessible no less than for books are since written history

started (although for some 20 years defines long term). This in turn requires sustainable file and data formats, often sustaining the software indispensable to read and/or handle the data, stronger disaster management, dependable funding and proper business models, etc.

For all commercial and semi-commercial (not for profit, but yet private) organisations, including such services as Figshare, there is no guarantee that data delivered there will be available in due course. But even for the public organisation EUDAT, there is no exit strategy defined (yet). By this is meant: what to do if EUDAT is threatened in its very existence.

Also only rather few partners in EUDAT are willing to engage into long term data preservation and only few can offer quotes for real long term storage, which just demonstrates how difficult it actually is to rely on the long term of data storage.

- ☞ Start defining the terms *short*, *medium* and *long term* in relation to data storage and software sustainability;
- ☞ Have EUDAT define an exit strategy for the sake of medium and long term data preservation;
- ☞ Define and share cost models for long term data archiving.

2.2 Integration issues/overlap

2.2.1 General

While EUDAT is extending its portfolio of activities and services, the question arises of overlaps in services with national, regional and local services. But is this really the case? And what if it does?

EUDAT is a meta-infrastructure, based on software and protocols on top of the existing (hardware)infrastructure provided and hosted by the EUDAT service providers. As such EUDAT does not compete with any of the service providers. On the other hand, while extending the service portfolio some services may be overlapping and of course by opening up national services to a European audience, services provided in one country may overlap with existing services in other countries.

- ☞ EUDAT is a meta-infrastructure on top of the existing national infrastructures. As such EUDAT does not compete with any of the service providers.

2.2.2 Workshop observations

EUDAT is considered not overlapping. But it is a strong requirement for EUDAT to understand synergies and to work together with successful established national infrastructure / service providers to prevent that it will. So it will need the participation of the infrastructures. By engaging with existing successful national and commercial service providers it can and should identify how to best integrate its services to form an easy-to-use ecosystem.

- ☞ EUDAT overlapping slightly with existing services is considered not harmful, but EUDAT should engage closely with existing national infrastructures and services in order to best add value.

2.2.3 Quality and user requirements

In order to remain attractive as added value service, EUDAT needs to offer services that are at least as effective or better/more efficient than what we are currently using. If it does, the other (national and European) infrastructures will start using EUDAT in addition to their own infrastructures or instead of and start or keep on building further on that.

Generally speaking, EUDAT will not be able to compete on quality with (very) specialized services tailored to a specific group, but it does not need to either. However, CESSDA presently unites social

science data archives from over 20 European countries, each with its own systems. The question arises: should CESSDA-partners, instead of using and maintaining their own systems, start offering the EUDAT services? Digital preservation software is expensive. It is more cost effective if one uses a shared service. CESSDA in this case is a (practical) example and presently is recognized as an ESFRI Landmark in the ESFRI 2016 Roadmap.

On the question how EUDAT can make services with quality for the customers, it is first of all necessary to be involved with the communities in order to learn and know about their current and future needs. This workshop can help establishing contacts with knowledgeable user communities and get insights from there.

- ☞ Quality of services is the single best asset EUDAT will have -and largely under its control- to sustain and (geographically) extend its activities. EUDAT should connect closely to user communities to make sure it focusses on the right services.

2.2.4 Visibility and engagement

The PLAN-E community generally operates closely to the national and European e-infrastructures. Nonetheless the scope and primary targeted audience for EUDAT are not really well known to even the PLAN-E audience. This may raise some concern for future sustainability. So who are the EUDAT customers?

It was thereafter convincingly argued that EUDAT is supposed to operate in the background, roughly comparable to GéANT. EUDAT is a background infrastructure in support of existing national data-infrastructures. In addition EUDAT does engage directly with the major European infrastructure projects ("ESFRI") and infrastructures ("PRACE, EGI"). This does not mean all concerns are addressed: most communities have no explicit ESFRI support, and the support from only the national infrastructure and services providers may not be enough to get enduring financial (funded) support, if the larger community is not aware of the impact of its services. The ESFRI-projects are relevant to collect user requirements and provide real use cases where EUDAT services can be tested and promoted. But EUDAT's efforts should not be limited to that.

EU's e-infrastructures are generally unknown to the general (scientific) public, as was revealed by recent surveys. In that sense, EUDAT is no better or worse than the other European e-infrastructures. So, how to broaden the name recognition of European infrastructures, such as EUDAT.

- ☞ Some suggestions:
  - Dataverse is organising "community meetings" annually, in which users, developers and service providers meet. These are very successful and EUDAT might do something similar.
  - EUDAT should have a clear idea of who they're targeting, and ensure they are disseminating this clearly, also to funders and "rectors of universities".
  - Complement EUDAT services to national infrastructures and ESFRI-projects with actions directed to end-users.
  - Broaden the scope to all domains in science.

In all cases -again- the quality and cost-efficiency of the services will be the best selling point and in all cases EUDAT must devote time and efforts to engaging communities beyond ESFRI.

2.3 FAIR principles and data services

FAIR principles require implementation in different practical environments. One kind of implementation is through the Data Seal of Approval (DSA), an international crediting service for qualified archives. Comparing the DSA and FAIR principles yields the following resemblance:

DSA Principles (for data repositories)	FAIR Principles (for data sets)
data can be **found** on the internet	**f**indable
data are **accessible**	**a**ccessible
data are in a **usable format**	**i**nteroperable
data are **reliable**	**r**eusable
data can be **referred** to	(citable)

2.3.1 To the F

It is generally agreed, at least during the workshop, that it all starts with *findability*. Is there a "Google/Bing" to find what you are looking for? It is the opinion of the audience, that one cannot beat parties like Google in its own operating domain. So rather than compete it is better to use Google, to become findable in the first place. This is why *metadata* are so important. EUDAT (or any organization providing data services) should provide for metadata and metadata services, both at the structural and at the descriptive level. Through the metadata, resources can be found by general purpose search engines. Metadata should be provided as close to the place where the data are created or published as possible. So basically end users (the data creators) should do this or be supported in doing this. If there are overlaps in search facilities: all the better. In addition it is always good to have a search engine that searches within certain domains, provided these engines can be found first by general purpose search facilities ("Google", "Bing").

Does EUDAT compete with private companies offering data services, such as Figshare or Mendeley? Basically not. EUDAT can, however, help serving such services, through the national centers, if these have a particular involvement in these or comparable private services.

2.3.2 To the A

Through the EUDAT "B2"-services, data findable at EUDAT are accessible as well. In short the B2-services encompass:

B2DROP - sync research data (similar to Dropbox)
B2SHARE - store and share research data (similar to Zenodo, Figshare, Dryad)
B2SAFE - replicate research data (similar to [C]LOCKSS consortium)
B2STAGE - stage data to computation (similar to GridFTP, GlobusOnline, HEP staging tool)
B2FIND - find research data (similar to national catalogues,)

The national services linked into EUDAT are in the business of data services and provide professional services for accessibility as well. Also the national professional data services organisations (DANS) provide for accessibility of data sets in the form of direct access or obtaining copies.

2.3.3 To the I

Interoperability comes in different flavors, depending on the discipline or type of data. This is why the implementation of the I is rather difficult. What does it mean if one hosts and services files, documents and archives of many different natures and disciplines. The FAIR principles themselves do not really give enough details. However, for data to be interoperable the principles do describe:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data.

(see https://www.force11.org/group/fairgroup/fairprinciples)

It is generally considered very important to have standards, use them and promote their use. EUDAT may consider setting standards (in cooperation with RDA) and help defining them, where necessary. For the rest it may seem more something to be left, in an organized way, to the different user communities, but the care for interoperability/exchangeability may be a matter that goes beyond the individual user communities.

2.3.4 To the R

Reusability is equally difficult to interpret in a practical context as Interoperability. It is not considered a principle concern for EUDAT as such. It is implicit to the background of the FAIR principles that reusability is a proper goal from a general point of view, but where it comes to services, such as to be provided by EUDAT, this is more a matter of the communities that share their data, combined perhaps with sufficiently long preservation times, so that other communities, becoming aware of the existence of certain digital information, will still have access to these data. So the R should be more the concern of national, local and disciplinary service providers.

- ☞ EUDAT should learn from professional (public) service providers how to best apply FAIR principles to different kinds of data.

2.4 Gaps/deficiencies

There are two types of gaps that come to mind in this workshop: gaps in national/regional coverage and gaps in services.

2.4.1 Geographically

Although EUDAT is a Pan-European service and infrastructure, not all countries are actually involved in it and so are lacking the links to this infrastructure. Among those countries are notably: Ireland, Denmark, Portugal, Czech Republic, Slovakia, Hungary, Bulgaria, (Turkey) and more. It seems in the general interest, if only for enlarging the scope of knowledge and data exchange beyond institutional and national boundaries, that EUDAT in due course will be present in all European countries. It is argued that the quality of the services beyond what can be offered on a national scale, would be the best selling point for getting these countries involved after all.

- ☞ EUDAT should strongly extend its geographical operations domain so as to cover the whole European Union and so add value to the already existing services.

2.4.2 Service portfolio

Basically, the discussion on gaps in the range of services yields the advice that EUDAT should do a continuous gap analysis of the services required by the ESFRI-projects, by other communities, research group and by individual researchers. Beyond that,

- ☞ Several suggestions come forward during the discussions for the portfolio of activities:
- Provide overviews of user community usage:
    - What combinations of services do different user communities use?
- Make it as easy as possible to use combinations of services within the Open Science Cloud environment (so including also the other European (e-)infrastructures).
- Provide privacy-aware and safe environments for the larger eHealth domain.
- Provide and advocate secure storage (meaning Raid-5 type services on a distributed organisational level).
- Clarify EUDAT's policies on data integrity and security.

- Play a role in tracking who is using what data - acting as a way of understanding usage and who gets the credits ("Google Scholar" for data sets).
- Openly compare and show differences and (dis-)advantages over commercial providers (e.g. Amazon).
- EUDAT could consider to provide data export on physical media services.
- EUDAT could consider to provide expertise: services to support data mining, data analytics, etc.
- EUDAT should consider establishing a user board, or an interaction platform with the science community (which could be (through) PLAN-E).

2.5 Research Data Management support

EUDAT has started to provide Research Data Management (planning) support. The question arose whether Research Data Management would be part of the infrastructure (EUDAT) or of the professional data service providers? Or both? And if so, where is the interaction between these parties on this topic?

Research Data Management Planning is an essential part of doing research properly in the present days. So every bit of help to get this off the ground is a good thing. However, it must be avoided for researchers to get flooded by well-meant advise, templates, models and more. An overview of what is going on:

- The RDA: The active data management group works on the development of DMPlans (https://rd-alliance.org/groups/active-data-management-plans.html)
- Openaire had a webinar on DMPlans.
- Science Europe has a working group on RDM. They work on protocols on a per-sub-discipline level. See: http://www.scienceeurope.org/policy/working-groups/Research-Data.
- The concept of (sub-)disciplinary research data management protocols or schemas can be found at https://dans.knaw.nl/nl/over/organisatie-beleid/informatiemateriaal/AConceptualApproachtoDataStewardshipandSoftwareSustainability_DEF.pdf.

☞ EUDAT is encouraged to follow the track of Science Europe, which means explicitly involving the researchers to compile (possibly pre-templated) models for Research Data Management, fitting their peer groups' needs. In due course a library of openly published RDMPlans should arise from this effort.

2.6 Future and funding

Various countries are evaluating or discussing their role towards EUDAT. For example Sweden is discussing the next generation of e-Infrastructure. Universities are being asked to contribute. Part of the discussion involves a decision between investment and utilisation of local, national or *European* services.

In Turkey, the decision making process is decentralised. What is offered should be based on users requirements.

Bulgaria and Romania are not included in EUDAT. Turkey is an observer. There are no known initiatives presently at the government level relating to research data. Decisions are in the hands of researchers, but EUDAT does not explicitly reach out to individual researchers.

It seems worthwhile for EUDAT to put efforts into better explaining the operating model of EUDAT. That it is not a European service that can replace national, regional or local services but that EUDAT extends such national, regional or local services to give them a European dimension, in return for which provides EUDAT the services mentioned above.

Obviously, participating in EUDAT requires national finding. This means that at national level decisions have to be made on spending funds on national data resources only or to both national and European data infrastructures. So in that sense national and European services are in competition.

- ☞ In order to achieve the best balance between what needs to be done nationally and European,
  - EUDAT is advised to engage with national funders and identify how they can raise the profile of EUDAT, and its services, and ensure investment at European level.
  - EUDAT could identify nations where there is not a well-used national infrastructure for certain parts of the data lifecycle, and make a benefits case for researchers, projects and institutions in that country buying their services.
  - EUDAT communicates how it intends to fund its future sustainability (can crowdfunding play a role?), if only to help guaranteeing the long term preservation of data.
  - EUDAT looks into the effects of European funding programs requiring data management plans that leverage European providers like EUDAT, EOSC?
  - EUDAT looks into what other user requirements are to be fulfilled by EUDAT once part of a broader Science Cloud environment
  - Have success stories available.

3 Conclusions

EUDAT's work and achievements are well appreciated by the workshop's participants. And all remarks are to be taken as positive encouragements to improve the efficacy and scope of EUDAT's endeavors.

Within the context of the European e-infrastructures growing towards an efficient and low-barriers European Open Science Cloud, EUDAT may and should play a binding role. If only for that, it is necessary that EUDAT in due course covers the whole of Europe, rather than the present set of participating countries.

Within the context of the triange: infrastructure-services-user requirements, the angles should become much closer to each other. Part of it can be achieved by more publicity about EUDAT's services and potential, but there remains a serious effort to be taken towards professional public data service providers and to user groups at large (beyond the "big" European funded facilities) to get better involvement and better integration.

PLAN-E is willing and suited to support EUDAT in these challenges and has demonstrated its serious interest in data matters through this workshop and these notes.