D3.3: EUDAT Conference Outputs and Recommendations

Author(s)	Sara Garavelli (TRUST-IT), Vasso Kalaitzi (LIBER)
Status	Final
Version	v1.0
Date	26/03/2018

Abstract: This document summarises the outputs of the EUDAT Conference, held from 22nd to 25th of January 2018, in Porto, Portugal. It also highlights recommendations for future EUDAT developments and emphases future trends that emerged in the course of the event.

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECT e-Infrastructures. Contract No. 654065

Document identifier: EUDAT2020-DEL-WP3-D3.3	
Deliverable lead	TRUST-IT
Related work package	WP3
Author(s)	Sara Garavelli (TRUST-IT), Vasso Kalaitzi (LIBER)
Contributor(s)	Heli Autere (CSC), Abdulrahman Azab (UiO), Michaela Barth (KTH), Rob Baxter (EPCC), Daan Broeder (Mertens Institute), Shaun de Witt (CCFE), Leon du Toit (UiO), Giuseppe Fiameni (CINECA), Licia Florio (GÉANT), Marjan Grootveld (DANS), Margareta Hellström (ICOS), Maria Francesca Iozzi (UNINETT SIGMA 2), Christos Kanellopoulos (GÉANT), Wolfgang Kuchinke (HHU), Yann Le Franc (e-Science Data Factory), Damien Lecarpentier (CSC), Ellen Leenarts (DANS), Barbara Magagna (Umweltbundesamt), Dick Schaap (MARIS), Gergely Sipos (EGI), Deboara Testi (CINECA), Luca Trani (KNMI), Mark van de Sanden (SURFsara), Matthew Viljoen (EGI); Angus White (DCC)
Due date	28/02/2018
Actual submission date	26/03/2018
Reviewed by	Anni Jakobsson (CSC)
Approved by	PMO
Dissemination level	PUBLIC
Website	www.eudat.eu
Call	H2020-EINFRA-2014-2
Project Number	654065
Start date of Project	01/03/2015
Duration	36 months
License	Creative Commons CC-BY 4.0
Keywords	Conference, European Open Science Cloud, European Data Initiative, EUDAT CDI, EUDAT Services, GDPR, Cross-collaboration

TABLE OF CONTENT

LIST OF FIGURES

EXECUTIVE SUMMARY

What does putting the EOSC vision into practice imply? Which factors are fundamental to its success? What role will research communities play? What are the common goals of EDI & EOSC, and how can they work best together?

These are just a few of the questions that over 230 participants attending the EUDAT Conference "Putting the EOSC vision into practice" set out to answer at a 3-day meeting held in Porto, Portugal. Attendees included policy makers, service providers and research communities from 25 countries, all involved in various data challenges and disciplines.

The conference was opened by Augusto Burgueño Arjona, Head of the "eInfrastructure" Unit, DG CONNECT, who presented EOSC as one of the key instruments available to support collaboration between e-Infrastructures & research infrastructures, and promote open science: "EOSC has to be an inclusive ecosystem where horizontal and thematic service providers work together to meet the user needs", he stated.

The discussion on how to make the EOSC a reality continued with panel discussion involving Mr. Burgueño Arjona and moderated by Annabel Grant, Senior Stakeholder Engagement Manager, GÉANT.

This panel discussion included Françoise Genova, Researcher at Centre de Donnees astronomiques de Strasbourg (CDS), Per Öster, Director, CSC & EOSC-hub Project Director, Grazia Pavoncello, representative of the Italian Ministry of Education, University and Research (MIUR) and Alex Vermeulen, Carbon Portal Director of ICOS ERIC, participated.

Panel recommendations:

- EOSC should focus on interoperability and integration issues − Previous investments (existing services) and experiences (competences and use cases) must be leveraged;
- EOSC should not be seen from a purely technical standpoint − EOSC is not intended to become a new platform, but a federation of existing services provided by different organisations with different governance and funding mechanisms. EOSC needs to define clear terms of use considering access rights, conditions to access services, and business models;
- EOSC has to be an "affordable", "inclusive" ecosystem − EOSC aims to take different countries, disciplines and institutions onboard and offer equal opportunities to all of them to create data-driven science. The access conditions need to be affordable from a cost perspective to encourage participation;
- Users must be in the driving seat − EOSC should permit a smooth onboarding process for mature communities and should act as an enabler for new communities;
- Research Infrastructures (RIs) play a dual role in EOSC: as users, and as thematic service providers. It is clear that EOSC presents the opportunity for RIs to expand their user base. Multi-disciplinary use cases would provide a good instrument to assess the value and the benefits that EOSC can bring to RIs, as well as to attract them;
- Measuring usage of the EOSC services is fundamental to assess its impact − A Virtual Access mechanism is already in place. Measuring usage is not only in terms of impact but also to compensate the service providers. Services that are used need to be funded;
- Data quality, trust, and user feedback are still open challenges.

The conference continued with an inspiring keynote by Michael Wise, Head of Astronomy, ASTRON − the Netherlands Institute for Radio Astronomy, who presented the data challenges posed by the Square Kilometer Array (SKA) Project. "Based on current projections, the SKA Observatory, once operational, is expected to produce an archive of standard data products with a growth rate in the order of 300 petabytes per year. Any further processing and subsequent science extraction by users will require significant, additional computing and storage resources".

This set the scene for a second panel, chaired by Rob Baxter, EPCC, University of Edinburgh on "The European Data Infrastructure (EDI) and the Data Challenge", focused on understanding the role of HPC in the EOSC and EDI landscape. Serge Bogaerts, PRACE, Giuseppe Fiameni, CINECA, Kimmo Koski, CSC, Sinead Ryan, Trinity College Dublin, and Michael Wise, ASTRON all contributed to the discussion.

Panel recommendations:

- EDI should be the traditional vehicle to provide HPC resources to EOSC;
- Albeit the funding streams for EuroHPC, EDI and EOSC come from different funding sources, user communities and ICT providers share mutual goals and need to work together to address societal challenges (climate changes, natural disasters, etc.);
- EDI and EOSC should be backed up by strong scientific cases, fundamental to demonstrate their true impact. Ideally, these cases should be multidisciplinary;
- Scientists managing exascale data need to focus on the science problem – The data and HPC resources are merely the instruments. ICT providers need to put scientists in the position to perform science.

The panel recommendations were further discussed during a set of breakout sessions, which explored the overarching themes of crucial importance in the creation of a thriving data economy, such as legal issues, interoperability of services, the role of research infrastructures as thematic service providers, business models and the sustainability of data infrastructures.

Nine co-located events, organised by ENVRI, EOSCpilot, the EUDAT Working Groups on Sensitive Data, Semantic and Array Databases, GÉANT, LIBER and SeaDataCloud complemented the EUDAT conference, allowing participants to deepen specific topics and to establish new collaborations.

Finally, the conference was the opportunity to engage stakeholders in discussion of the future of EUDAT and the EUDAT Collaborative Data Infrastructure (CDI). Already counting more than 20 members including leading European research organisations, data and computing centres, the EUDAT CDI will continue to develop and run an interoperable layer of common data services to support research in Europe. The participation of the EUDAT CDI in the EOSC-hub project (www.eosc-hub.eu) will also allow EUDAT to play a concrete role in the EOSC ecosystem and will guarantee a continuous interaction with the user communities that have been at the heart of the EUDAT strategy since it started in 2013.

1. PLENARY SESSIONS

1.1. Plenary 1: The European Open Science Cloud – Putting the Vision into Practice

The EUDAT conference was opened by **Kimmo Koski, CSC Managing Director and the EUDAT Coordinator** who described EUDAT's major achievements since its beginnings:

Figure 1: Kimmo Koski, CSC Managing Director & EUDAT Coordinator

- the design, implementation and delivery of a complete set of services to address the full lifecycle of research data, the so called EUDAT Service Suite or B2Services[1];
- the wide adoption of EUDAT services thanks to a service-centric strategy based on service co-design, involving users in all phases of the service definition, implementation, configuration and operation lifecycle (current services have been developed in close collaboration with over 50 research communities);
- the establishment of the EUDAT CDI[2]: With a network of over 20 European research organisations, and data and computing centres in 14 countries, the EUDAT Collaborative Data Infrastructure (CDI) is one of the largest infrastructures of integrated data services and resources supporting research in Europe. It is designed to address the full lifecycle of research data, representing a strategic solution to the challenge of data proliferation in Europe's scientific and research communities. This is achieved through ongoing collaboration between Service Providers and Research Communities, working within a mutually beneficial framework for the development and operation of an interoperable layer of common data services.

All these activities were performed constantly keeping the European Open Science Cloud (EOSC) scenario to which EUDAT also contributes in mind, and providing relevant recommendations for its shaping.

EUDAT recommendations for a successful EOSC:

- Bring e-Infrastructures together through cooperation, rather than competition, and clearly define the EOSC & EDI roles from the outset;
- Build trust between RIs & E-INFRAs, agreeing roles & responsibilities, co-designing services, and avoiding a "one-stop shop" approach;
- Delineate European & national roadmaps: sustainability of the EOSC has to be based around national funding. This is why EOSC needs to solve real needs at a national level to be supported by member states and the EOSC roadmap needs to be consistent with national roadmaps & investments;
- Never forget the Scientific Case: Jobs, growth, and economies of scale do matter… but science & research should be the main drivers!

Kimmo Koski's introduction was followed by the keynote talk delivered by Augusto Burgueño Arjona, Head of the "eInfrastructure" Unit, DG CONNECT, who presented EOSC as an significant instrument to support collaboration between e-Infrastructures & research infrastructures, and promote open science: "***EOSC has to be an inclusive ecosystem where horizontal and thematic service providers work together to meet the user needs".*** In order to maximize previous investment the EOSC ecosystem should rely on existing services and infrastructures, as well as commercial services that can complement public resources to meet user demand.

Figure 2: Augusto Burgueño Arjona, Head of Unit "eInfrastructure", DG CONNECT

[1] https://eudat.eu/catalogue
[2] www.eudat.eu

Figure 3: Per Öster, CSC Director & EOSC-hub Project Director

Per Öster, CSC Director & EOSC-hub Project Director, picked up on several of the points raised by Mr. Koski and Burgueño Arjona in their presentations. "All actors involved in EOSC have clear in mind that EOSC is not about creating a new single platform that will solve all the issues of the users. **EOSC is about integration and interoperability** which are also the objectives of the recently funded EOSC-hub project[3]. **EOSC-hub will create and provide the Hub as a federated integration and management system for the future EOSC.** It will implement the Hub as **an open, community-led framework** and will **provide a wide range of services from numerous major digital infrastructures and research communities** starting with those of EUDAT [4], EGI [5] and IndigoDataCloud[6]".

Figure 4: The EUDAT conference participants

A panel moderated by Annabel Grant, Senior Stakeholder Engagement Manager at GÉANT followed. This included Françoise Genova, Researcher at Centre de Donnees astronomiques de Strasbourg (CDS), Per Öster, Director, CSC, and the EOSC-hub Project Director, Grazia Pavoncello, ministerial representative at the Italian Ministry of Education, University and Research (MIUR) and Alex Vermeulen, Carbon Portal Director of ICOS ERIC, together with Mr. Burgueño Arjona.

Panel recommendations:

- **EOSC should focus on interoperability and integration** – Previous investments (existing services) and experiences (competences and use cases) must be leveraged;
- **EOSC should not be viewed from a merely technical standpoint** – EOSC is not expected to be a new platform rather a federation of existing services provided by different organisations with different governance and funding mechanisms. EOSC needs to define clear terms of use considering access rights, conditions to access services and business models;
- **EOSC has to be an "affordable", "inclusive" ecosystem** – EOSC aims to take on board different countries, disciplines and institutions and offer all of them equal opportunities to carry out data driven science. The conditions for access need to be inexpensive to encourage participation**;**

[3] www.eosc-hub.eu
[4] www.eudat.eu
[5] www.egi.eu
[6] www.indigo-datacloud.eu/

- **Users must be in the driving seat** – EOSC needs to allow a smooth transition of mature communities and be an essential enabler for new communities;
- **Research Infrastructures (RIs) have a dual role in EOSC: both as users and thematic service providers** – It is evident that EOSC provides the opportunity for RIs to expand their user base. Multi-disciplinary use cases could provide the perfect tool to assess the value and benefits that EOSC can bring to RIs and to attract them;
- **EOSC service usage has to be measured to assess impact** – A Virtual Access mechanism for this is already in place. Gauging usage is not only fundamental in terms of impact, but also to compensate the service providers. It is the services that are used that need to be funded;
- Data quality, trust, and user feedback are still open challenges.

"***From an artistic point of view EOSC is a movie, which is art, but behind it there is an industry that has to organise the infrastructure, the governance and the sustainability model."*** *Augusto Burgueño Arjona, European Commission*

Figure 5: Panel Plenary Session 1: The European Open Science Cloud – Putting the Vision into Practice

1.2. Plenary 2: The European Data Infrastructure (EDI) and the Data Challenge

The conference continued with an inspiring keynote by Michael Wise, Head of Astronomy, ASTRON – the Netherlands Institute for Radio Astronomy, who presented the data challenges behind the Square Kilometer Array (SKA) Project.

The SKA is an ambitious project to construct one the world's more powerful radio telescopes and enable transformational science across a wide range of research areas. Based on current projections, once operational, the SKA Observatory is expected to produce an archive of standard data products with a growth rate in the order of 300 petabytes per year. Although the challenges posed to populate and maintain the SKA science archive are already significant, these data products currently represent only the first part of the full science extraction chain. Any further processing and subsequent science extraction by users will require significant, additional computing and storage resources.

"*The potential is there and we have trouble to extract the science because of the storage and the computing resources and the complexity of the data. Given the data volumes and processing scales expected from the SKA, the large-scale infrastructures must provide a processing environment that allows astronomers to manipulate and analyze large data collections in flexible and familiar ways, but at SKA data scales. A European Open Science Cloud for research represents an important resource to provide the astronomy community with the scale of computational infrastructure necessary to maximally exploit the scientific potential of the SKA.* ***By providing access to large-scale storage and computational resources as well as expertise in high-performance and high-throughput computing and software middleware, the EOSC could represent a core foundation upon which the SKA is built,***" commented Mr. Wise.

Figure 6: Michael Wise, Head of Astronomy, ASTRON – the Netherlands Institute for Radio Astronomy

Serge Bogaerts, Managing Director at PRACE[7], the Partnership for Advanced Computing in Europe, introduced the audience to the High-Performance Computing (HPC) world, remarking that there is a changing trend in the domain: historically, the main objective of HPC was to provide powerful supercomputers to solve complex computational problems. **Today in the HPC world, the data intensive aspect has to be considered and this will be the focus of the European Data Infrastructure (EDI) initiative**. Over the past 30 years, a number of domains like climate research, weather forecast, astrophysics, and chemical engineering have been benefitting from HPC and today new scientific domains are realising that HPC is key, such as biology life science, material science, as well as domains such as social science and humanities. Industry has also relied heavily over the years on HPC especially in sectors such as aeronautics, oil and gas, automotive and finance, and today HPC is a key enabler for medicine, developing nanotechnologies, renewable energies, etc.

Figure 7: Serge Bogaerts, Managing Director at PRACE

"*HPC is more and more becoming an instrument supporting public decision-making processes (for example it has been very successful in predicting trajectories of hurricanes), that's why* ***in PRACE we have decided to move towards a more data centric approach as data is the common denominator of all scientists****. We need to learn how to handle a large volume of data that are generated on these machines, or by experimental facilities like SKA and offer proper computing facilities for that.* ***If we manage to align EOSC and EDI properly we can provide better support for science***".

These presentations provided the framework for a second panel chaired by Rob Baxter, EPCC, University of Edinburgh on "The European Data Infrastructure (EDI) and the Data Challenge" which focused on understanding the role of HPC in the EOSC and EDI landscape. Serge Bogaerts, PRACE, Giuseppe Fiameni, CINECA, Kimmo Koski, CSC, Sinead Ryan, Trinity College Dublin, and Michael Wise, ASTRON contributed to this panel discussion.

Panel recommendations:

- **EDI should be the traditional vehicle to provide HPC resources to EOSC**
- **User communities and ICT providers should work together to address societal challenges** such as climate change, natural disasters, etc. Even if there are different funding streams behind EuroHPC, EDI, and EOSC, the objectives are common
- **EDI and EOSC should be supported by strong scientific cases, which are fundamental to demonstrate their real impact. Ideally, these cases should be multidisciplinary ones.**
- **Scientists managing exascale data should focus on the science problem** – The data and HPC resources are just the instruments. ICT providers need to put scientists in a position to perform science.

[7] http://www.prace-ri.eu

Figure 8: Panel Plenary Session 2: The European Data Infrastructure (EDI) and the Data Challenge

2. BREAKOUT SESSIONS

The panel recommendations were further discussed in a set of breakout sessions which explored the overarching themes of crucial importance to creation of a thriving data economy, such as legal issues, interoperability of services, the role of research infrastructures as thematic service providers, as well as business models and the sustainability of data infrastructures.

2.1. "EUDAT what's next?: CDI & EOSC" and "EOSC: what's in it for researchers & service providers?"

This session presented the recent developments of EUDAT in its effort to move from an EC project to a sustainable collaboration between partnering organisations, through the EUDAT CDI Collaboration Agreement. It also provided some information to organisations interested in engaging with CDI on the various opportunities to do so, including how to formally join the CDI.

Figure 9: Debora Testi, CINECA

It continued with a discussion about the role of the research communities within the CDI. How can EUDAT best support the involvement of research communities in the CDI? What are the appropriate mechanisms to be able to influence the EUDAT agenda?

The session concluded with a slot dedicated to the presentation of the workplans and joint activities of the EOSC-hub and OpenAIRE-Advance projects, recently funded as part of the EC call e-Infra-12 to lay the technical foundations for the European Open Science Cloud with a discussion on the principles of engagement focused on what it is becoming a service provider in EOSC context. Several discussions took place on how to organise the onboarding of new service providers and how to get agreement (The detailed notes of the session are reported in annex A.1).

Conclusions reached

- The added value of the EUDAT CDI is now becoming clearer to the audience and the fact that the CDI is an open endeavour welcoming new organisations to join was seen as a very positive aspect of the collaboration and whole enterprise.
- The active involvement of research communities in the CDI is critical for the success of EUDAT. It is its trademark and should always be pursued.
- A board of users seems a good way to keep users involved in the future of EUDAT services.
- Definition of the exact role and responsibilities of the users' board should be achieved before defining the composition and membership of the board precisely.
- Involving the communities in the discussion related to the principles of engagement is critical to obtain their ultimate buy in.

2.2. EUDAT Collaborative Data Infrastructure Services

The first EUDAT CDI service session, attended by approximately fifty participants, was dedicated to presenting the status of the EUDAT services, their background, how they have evolved, the current status and the plan for the future. It is clear that there is still the ambition to improve the services, extend them and make them more usable to the users. The second session was dedicated to showcasing demos of each of the services and some of the use cases developed really addressing the community needs. The demo sessions were all particularly well attended and proved that there is still a high demand for the type of services being offered. One particularly interesting recurring question was: "How to obtain a 'good' long-term agreement with an EUDAT service provider". (The detailed notes of the session are provided in annex A.2).

Conclusions reached

- There is still a great deal of interest in the EUDAT services and, coincidentally, the B2 branding is very well recognized.
- The demo session was organised with a marketplace set up (seven tables equipped with a screen in one big room). Each table was assigned different EUDAT service for which a demo was provided) and this format was hugely successful, with many very technical discussions. This seemed to be a good format to advertise services.
- Plans for further service developments are much appreciated, even if it's not sure how all of the plans will be funded.

Figure 10: Two views of the EUDAT marketplace session

2.3. Harvesting results from the EUDAT Community: Demonstrators & Pilots

In this session, communities and pilots, especially the EUDAT extended data pilots, were invited to showcase real examples of EUDAT services enabling research via demos. The most surprising aspect of the session was to see how users have employed the different services in different ways, even for purposes that the providers hadn't thought of. After the demo session, flash presentations were made by the communities: they were asked to make a statement on how EUDAT impacted the landscape and to report the lessons learnt. The main concerns raised were related to the future: stability of services and the stability of the offer of storage. On the other hand, all communities recognized that the two EUDAT projects have been quite special, in the way that data centers were able to discuss and relate to communities and they hope that this will continue in the future with the EUDAT CDI. (The detailed notes of the session are provided in annex A.3A.2).

Conclusions reached:

- Those communities that used APIs (in B2SHARE and B2DROP especially) have been successful in implementing solutions that meet their needs;
- The interdisciplinary nature of EUDAT tools is demonstrably useful;
- Implementation was difficult mostly because APIs were evolving throughout the project – bleeding edge;
- B2ACCESS is useful for enabling access and also for reporting on how data is actually used;
- Using EUDAT services provides incentives to standardise tools, data and metadata which save time when experiments are repeated – sometimes there is resistance in the short term, but this of benefit in the long term for the community;
- Integration with EUDAT services can foster the renewal of existing community tools;
- Communities are concerned about maintaining transparency during EOSC;
- EUDAT has become an influential player; it is important that EOSC fosters sustainability of services and APIs;

- Service stability is essential before production release;
- Communities need to provide clear answers about what will happen to the underlying resources after EUDAT2020 (e.g. what will happen to the storage?).

Figure 11: View of the session's participants

2.4. Cross e-infrastructure collaborations

This session was organised around four main themes:

- the EUDAT-PRACE collaboration, discussing the results obtained from combining HPC and data resources.
- the EUDAT-EGI collaboration, with success stories from the ICOS and the ENES use cases.
- the new projects due in the realm of extremely large datasets towards data exascale, understanding how these new solutions can be accommodated within the EUDAT suite of services and how EUDAT can extend its services to handle very large datasets.
- Assessment of the EUDAT services based on the results obtained by two small companies: SMEs are reluctant to release data, as there are no mechanisms to control who has access to it. It's important to extend the services provided by EUDAT to support very fine-grained authorization and access and provide information on the cost of this type of service.

The detailed notes of the session are reported in annex A.4.

Summary of Conclusions:

- The capability to couple data and compute resources together is key to fuel scientific innovation and advance research frontiers. This is why the EUDAT CDI will continue to collaborate with PRACE in the upcoming years;
- Involving user communities in the design process for the work conducted by EUDAT and EGI for the production of a cross-infrastructure offering users seamless access to data and high-throughput computing resources, helped both EGI and EUDAT to better tailor their services to match the real needs of specific user communities;
- Recommendations provided as to how to improve documentation and further testing prior to tools making a tool available to early adopter communities;
- Recommendations provided as to how to onboard further user communities in the best way;
- Overlaps and complementarities among initiatives focusing on computing e-infrastructure dealing with extremely large datasets, need to be further discussed. Channels need to be kept open for further collaboration. A session on this topic is planned as part of the first EOSC-hub week[8].

2.5. The impact of the policy framework on EOSC

The session focused on the difficulties that the General Data Protection Regulation (GDPR) and the various interpretations of this can bring; on what EOSC can do in terms of information governance and what we can do at an inter-governmental level. The first slot of the session comprised four talks, beginning with the policy framework of the new GDPR, then taking a look at the progress towards codes of conduct for research in the life sciences, followed by a report on a pilot study to apply the DataTags model to GDPR and concluding with an introduction to a 'data safe haven' service as a possible type of future EUDAT service. The second slot began by reporting the results of the EUDAT Sensitive Data Working Group, which was then discussed with the audience. Some highlights:

- Different communities, not surprisingly, have different new needs.
- There are a lot of challenges, both technical and ethical.
- The technical challenges are not the problem! (Echoing Rob's talk, or earlier.)
- There are specific/regional solutions available, and one key question is: should the data be moved, or the tool?
- How can we ensure reuse of data for research purposes?
- EU strategy and recommendation on principles, protocols and best practices are called for.
- We need more awareness-raising, and there is a clear need for cost-effective solutions: an EU-health information infrastructure?
- An inescapable conclusion was that it is the tools that need to be moved, not the data.

(Detailed notes of the session are reported in annex A.5).

Conclusions reached

- Data sharing for research just got a lot harder!
- Sending tools to data may be the best approach for now.
- Promoting standards and common approaches will help, even if sharing data is hard.
- The EOSC-hub collaboration of data hosts and data researchers might carry enough weight to lobby for relaxing restrictions related to research data sharing.
- Some see the GDPR as just another cost item to organisations, but other organisations see the benefits.
- Strategies are clearly emerging national to re(use) sensitive data, albeit with diverse opinions with respect to central vs. distributed sensitive data storage.

[8] http://eosc-hub.eu/events/eosc-hub-week-16-20-april-2018-malaga-spain

My personal data are mine; to abuse them is a crime; you cannot share; you must take care; or risk a hefty fine!

Figure 12: View of the session

2.6. User engagement & training

In this session, representatives of PRACE, EGI, OpenAIRE, EUDAT, ELIXIR, University of Goettingen and students who attended the EUDAT Summer School were asked to share "Hooray" and "Horror" stories from their training experiences. Many good training practices emerged. This is why in the second session the discussion turned as to how the sharing of good practices and experience can continue under a Community of Practice (CoP) for trainers. This could take the form of an informal network of trainers, independent of projects, allowing trainers to share expertise. The participants endorsed this idea, highlighting that the network should not involve all of the trainers, but the trainer co-ordinators. There will be some follow up activities on this. (The detailed notes of the session are provided in annex A.6).

Conclusions reached

- Coordinating or providing training, implies building trust at all levels. Targeting the right audience is essential – the human factor may be more important than content that perfectly matches the learning need.
- Organising and delivering training on the scale that the EOSC requires, would benefit from a Community of Practice for training coordinators/training managers and could provide a valuable instrument to share experiences and resources.
- A CoP looks to be most promising for training coordinators and training managers.

3. POSTER SESSION 1 MINUTE MADNESS & DEMO SESSIONS

At the end of the first day, 35 posters were presented to the audience[9]. Each poster presenter had to summarise the poster outcomes in a minute.

The best poster was awarded.

Figure 13: Poster session 1 minute madness

Four demos were also showcased during the coffee breaks:

- eInfraCentral – Your gateway for European e-infrastructures in action;
- Data collection with B2DROP and LabTablet;
- OpenAIRE Services: We need your input!
- Interactive showcase of the Dendro research data management platform.

[9] The full list of posters is available at https://eudat.eu/eudat-conference-porto-posters-demos

4. CO-LOCATED EVENTS

Nine co-located events organised by ENVRI, EOSCpilot, the EUDAT Working Groups on Sensitive Data, Semantic and Array Databases, GÉANT, LIBER and SeaDataCloud complemented the EUDAT conference, allowing participants to go deeper in specific topics and to establish new collaborations.

4.1. Sensitive data workshop

The main objectives of the workshop were to collect experience, evaluate technical solutions, and the challenges associated with the use of sensitive data in research. After a brief introduction given by the Sensitive Data Working group, a number of user experiences about carrying out research on personal data were presented, covering different research environments: linguistics, educational science, clinical practice, biodiversity.

The variety of domains shows how the issues related to carrying out research on sensitive personal data do not only affect research environments traditionally handling personal health data. Video data and audio data may also become sensitive, even though the research purpose under which they were collected was not to investigate sensitive topics. Furthermore, sensitive video and audio data are difficult to handle, as this is exposed as soon as they are played on a device, despite how secure the storage backend is. This makes the technical solution to analyse sensitive audio/video files very demanding and poses the strong need for metadata, that make the data discoverable without necessary transcribing them, and the need for widely recognised best practices. But also, anonymized data collected by sensors or social large data collections might become sensitive as they are aggregated.

Although the type of data/format/amount may be different, the challenges are quite similar and only regard the “degree of sensitivity” of the dataset, which in turn dictates the level of assurance and identity vetting needed to access the data and the security measures required to store the data. Several technical solutions were presented and provided locally/nationally, to support the use of personal data for research: from secure cloud solutions to store analyse and compute sensitive data to secure lab designed to optimise omic-analysis, from high-performance solutions to secure big data lake, from tagging strategies to archive solutions, all implemented in compliance with statutory act with regard to the treatments of digital personal and sensitive data.

But in the digital data era, huge research possibilities have been opened up by sharing sensitive data across different countries. In this context the Nordic countries, each one having a secure, national cloud solution and a homogenized set of statutory acts and best practices, represent a unique “sand box” for exploring solutions spanning national borders. This is the goal of the Tryggve project, a supported by all these countries as well as ELIXIR: to establish a network of interoperable secure solutions by facilitating the mechanism to enforce trust; to exchange data and to exchange computing tools.

The secondary use of health personal data, repository data and social data, albeit challenging from a legal point of view, provide research possibilities with the highest impact on society, and therefore is beneficial to the both the scientific community and society as a whole. The discussion on how to allow secondary use of health and social data is now taking place at different level. How Norway, Denmark, Finland and France are responding to this need was learned, but also how the problem has been tackled in a broader context, aimed at defining a European-wide infrastructure for health data. In all these processes of requirements, extraction and architecture-design definition, a common denominator once more appeared: it seems more feasible to move the processes rather than the data. That is, the aim should not be a global-wide data repository, rather a network of repositories accessible through proper analytical tools. This is also the vision provided by GoFAIR in the “FAIR train” model. Thus, this will open up the possibility to perform research jointly across disciplines with sensitive and open data sets.

The meeting concluded with the suggestion of preparing a memorandum as final product of the EUDAT sensitive data working group to provide a starting point for the next debate on the subject. Where and within which framework this discussion should take place is still under debate, but one possibility could be to

establish an interest group in the RDA (the detailed notes of the Semantic services in EOSC workshop are available in Annex B.1).

Conclusions reached

- Existing tools and services are sufficient, the focus here must be on policies and processes.
- There is a wide range of cases where sensitive video data is used in research.
- The main issue is ethical requirements.
- Technical challenges exist but these can be solved in the long term.
- Applicability is possible in countries outside Europe, while in EU it is more challenging.
- Sensitive data can be data centric or person centric.
- Identity assurance is a good start but may not be sufficient.
- A EU health information infrastructure is still needed.
- Best Practices should be presented: specific examples with benefit/risk analysis
- Why are separate European Genome phenome Archives (EGAs) implemented? The local EGAs are for sensitive data (e.g. genomic data), B2SHARE is for open data. The aim is to extend this to the whole of Europe, enabling data linkage in a secure way. In contrast to the rest of Europe, the Nordic countries are homogenous, and regulations are similair. TSD/ePouta are prepared for EOSC. The aim is to generate trust.
- Processing large data is challenging both in terms of security and storage.
- Sharing and accessibility of metadata is more important than the data itself.
- The metadata architecture layer may not be sufficient, and needs to be extended with other vocabularies.
- It is an important principle that it's not the data that should be moved!
- Taking on the burden of composing EU strategies for sensitive data reuse is cumbersome.
- There are barriers for sharing data across countries, mainly due to the extremely diverse security rules and regulations.
- The trust mechanism between service providers and data owners should make provision for the data owner to maintain full control of the data management and access workflow.
- A Good solution is "Pooling data on the move", whereby data is stored in national repositories/clouds and pooled, upon approval, for specific analysis only.
- One solution is not going to resolve every issue. Different platforms to address different issues such as data security and integrity can co-exist. The matter is where Europe's metadata will be stored. A single repository is unrealistic. But the focus needs to be on free access to the metadata of sensitive data.
- Encryption of data is not sufficient. One possible method is fragmentation, and therefore only the part necessary for analysis.
- Finally, a sensitive data cloud is a Dream that should be possible to come true!

4.2. Semantic services in EOSC

Within the EUDAT CDI, a number of semantic services have been developed, and are either production-ready, such as B2NOTE, or as proof-of-concept for ontology aggregation and discovery, and Data Life Cycle modelling and provenance tracking. The objective of this meeting was to involve research communities in the discussion as to which kind of semantic services are needed to support crucial research challenges (the detailed notes of the Semantic services in EOSC workshop are available in Annex B.2).

Conclusions reached

- Semantic resources are first class citizens in the implementation of the FAIR data management system.
- There is a need for semantic services with multi-disciplinary coverage.

- There is a need among communities such as Biodiversity and Ecology to aggregate data from multiple disciplines.
- There is a need for Knowledge Engineers/Ontologists to access, aggregate and analyze semantic resources.
- There is a need for an Simple Knowledge Organization System (SKOS) vocabularies lookup service.
- There is a need for provenance of mappings and versioning of concepts.
- There is a need for a central hub for ontologists and knowledge engineers to access multi-disciplinary semantic resources and analytics.
- There is a need for a multi-disciplinary, curated semantic index that can be used by semantic tools (annotators, text miner, etc.).
- There is a need for a set of recommendations and standards to ease the interoperability of semantic resources.
- There is a need for potential business applications that could guarantee the sustainability of the resources.
- B2services are used by numerous infrastructures.
- B2NOTE and Semantic Look Up services are the EUDAT Semantic Services of interest to most RIs.
- It would be useful to have the possibility to connect to an external “synonym service” to enhance user experience while searching the ICOS data catalogue, such as the type proposed by AnaEE, for example.
- An annotation exchange service is a necessary extension of B2NOTE.
- An annotation of time series datasets to indicate regions of unusual activity, indicating an interesting event or a possible instrument failure is also needed, in general, annotated content of files would also be a necessary feature.
- Integrated automated text document annotation functionality in B2NOTE as part of an expert annotation workflow is also needed.
- A tool embedded in an ontology and thesaurus editor to comply with a metadata standard for semantic resources would be good to have.
- People should be able to vote or comment on an annotation stored in the B2NOTE database by adding some social media functionalities.
- Visualisation of the annotation needs to be improved and B2SHARE better integrated.
- A marketplace for ontologies is required.
- A semantic lookup service should be integrated with Semagrow.
- Metadata quality should be improved in repositories; towards a new standard for ontology metadata with the VSSIG.
- Ontology Repositories should be standardized.
- Collaboration between LifeWatch/LTER-Europe and AgroPortal to build a Biodiversity Community Portal should be encouraged.

4.3. Array databases for research communities

This workshop focused on the presentation of the results of the experiments selected in the EUDAT Array Database Working Group highlighting the benefits, limitations and challenges encountered.

Some positive aspects emerged:

- These help make life easier;
- They provide standardized access to data;
- They are less error prone to subsetting/assembling issues;
- They provide flexible access to relevant data partition;

And some negative ones:

- Input data has to be perfect, which is unfortunately not always the case in life;

- There is still need for caution in the database query when numerous processes are required (distributed installation may solve this).

The workshop also addressed possible strategies to be taken in to consideration to extend these activities beyond EUDAT2020 and address future work (the detailed notes of the Semantic services in EOSC workshop are provided in Annex B.3).

Conclusions reached

- The real strength of ArrayDB is the scalability possible for large data volumes.
- The data pre-processing step is a key, in terms of processing time and data quality aspects.
- The ArrayDB Working Group should continue within RDA.

4.4. Research data management: interoperability, collaboration, and the research library role

The workshop aimed at uncovering intersections between the activities of three organizations supporting research data management practices from different perspectives: RDA[10], as a standardization body; EUDAT and its Collaborative Data Infrastructure (CDI), providing research data management (RDM) services; and the Research Libraries community, helping researchers be successful in their RDM endeavors. The main points of discussion revolved around three topics: i) the need of repurposing in terms of collaboration, discussion amongst libraries on how to deliver services and the researchers view of the research libraries: there is also the need to join forces with other institutions, such as university computing units, like the Max Planck digital library, RDA and initiatives like Go Fair; ii) managing research objects and the expertise that libraries have - the role of libraries in vocabulary management, data typing, citations, how to involve libraries in the data process and in developing recommendations. Some of the results emerged from working groups, where library members participated or lead such as support for projects; demonstrators; proof of concepts; to show the added value. To find a way to define open and interoperable types of data, in order to preserve research objects, protocols and workflows efficiency should not become a taboo. Cost-efficiency is not always the issue (the detailed notes of the Semantic services in EOSC workshop are available in Annex B.4).

Conclusions reached

The main conclusion was the need to find a common ground between vision and further collaboration, in order to meet researchers' and librarians' expectations. Clear roles are needed. There are some budgeting issues that need to be addressed to take into consideration the thoughts that emerged from this co-located event and move forward. Another important aspect of the role of libraries is that of advocacy, raising awareness, training and decision-making.

4.5. SeaDataCloud workshop

The presentations of the SeaDataCloud workshop gave a nice overview of a number of activities underway to improve and expand the services of the SeaDataNet infrastructure for marine and ocean data management and highlighted where the cooperation and synergy with EUDAT and its services is planned and under development. The main discussion was around INSPIRE. As part of SeaDataCloud, a mapping has been made from the SeaDatanet standard formats (CDI metadata and ODV data) towards the INSPIRE data models. It is now planned to use this conceptual mapping to resolve a number of use cases, in particular for nutrients and contaminants which are highly relevant for the implementation of the EU Marine Strategy Framework Directive (MSFD). The use cases will provide the input for setting up Transformation Services from SeaDataNet towards INSPIRE compliance. This will be extremely relevant for Member States that are 'struggling' with INSPIRE compliance and that are or might become contributors to the SeaDataNet

[10] www.rd-alliance.org

infrastructure. However INSPIRE is never mentioned in the development of the EOSC which is more focused on FAIR, while scientific data should also be fit to support environmental management and its policies (the detailed notes of the Semantic services in EOSC workshop are available in Annex B.5).

Conclusions reached

The workshop was dedicated to making participants aware and informed on the ongoing activities in the SeaDataCloud project which is a joint activity between the marine data management community around SeaDataNet – EMODnet and EUDAT.

4.6. ENVRI workshop

The workshop focused on presentations on the H2020 project ENVRIplus[11] and its members & end user communities, and activities and outputs (specifically the portfolio of research data management services) of the work packages associated with the projects "Data for Science" theme. The original objective of the workshop was to provide a forum for different communities in ENVRIplus and EUDAT to discuss updated requirements for effective data management services and research support systems, to share development results and best practices, and to propose an agenda to move towards EOSC. In the end, due to the last-minute cancellation of several invited speakers, the agenda was shortened to focus on the development of services within ENVRIplus and on the on-going activities focusing on mapping the landscape of potential end users of ENVRI data products and other outputs (the detailed notes of the Semantic services in EOSC workshop are available in Annex B.6).

Conclusions

- The Horizon2020 cluster project ENVRIplus is one example of a collaboration between domain-specific Research Infrastructures (RIs) aimed at developing reusable solutions to address common challenges in managing research data. This approach is proving to be very successful, but cannot fully solve interoperability and sustainability issues.
- In the RI projects, IT service development and maintenance are known to be very important; however, their budgets are often very limited. In the meantime, existing software and tools developed by different RI communities are not yet fully exploited and reused.
- On the one hand, there are increased constraints on RIs and other "larger" (European) research projects to render whatever services they are developing to a larger audience − whether or not these research-related services are stand-alone, or rather layered on top of basic services provided by e-Infrastructures associated with the EOSC.
- On the other hand, duplication of efforts on similar topics can still be observed in RIs due to limited readiness, visibility or interoperability of existing software.
- A number of challenges have been identified, such as how to catalogue development results from different RI communities, assess their quality, identify gaps and promote standards for interoperability
- Who are the key actors that need to be involved in order to ensure that these RI-produced services are consistently and effectively evaluated and their quality assured?
- How can sustainability, in terms of both operations and competence, be guaranteed, at least in a the medium-long timeframe?

[11] http://www.envriplus.eu/

4.7. Federated AAI workshop

The main focus of the workshop was to report on work ongoing in the AARC[12] project, with particular emphasis on the AARC blueprint architecture (AARC BPA) and on its implementations.

There were a number of questions that were asked during the session. Most of the questions had a very technical focus. The AARC team has collected them and they will add a Q&A section to the AARC website (the detailed notes of the Semantic services in EOSC workshop are available in Annex B.7).

Conclusions reached

- One of most frequently asked questions is how AARC BPA and related work can inform the work in EOSC-hub. AARC has paved the way to show that it is possible to define a reference AAI for research collaboration. It is expected that the results of the AARC pilots concerning the deployment of the BPA in the production environment, as well as policy best practices and technical guidelines will be a valuable starting point for EOSC-hub. Luckily, there is a significant number of key partners that are involved in both AARC and EOSC-hub to ensure that continuity of work.

4.8. Piloting EOSC governance framework

EOSCpilot[13] is an EU project that supports the first phase in the development of the European Open Science Cloud (EOSC). EOSCpilot main objectives are:

- Facilitating access of researchers across all scientific disciplines to data;
- Proposing a governance framework and business model that sets the rules for the use of EOSC;
- Creating a cross-border and multi-disciplinary open innovation environment for research data, knowledge and services;
- Establishing global standards for interoperability of scientific data.

Establishing a governance for the EOSC is a challenging task, as it aims at constructing a framework allowing strong but disparate stakeholders to work together. Stakeholders include for instance: research communities, research institutions, research infrastructures, e-infrastructures, and research funding bodies. This framework also needs to address cultural challenges, encouraging the adoption of new ways of working and scientific practices. Overall, it will shape and oversee future development of the European Open Science Cloud.

At the end of its first year of activity, EOSCpilot has defined a first draft framework for governance, principles of engagement of stakeholders and sustainability. This initial framework has to be improved and tuned according to the inputs of the stakeholders in order to shape it in the most effective and broadly accepted way.

The session aimed at introducing an initial version of the governance framework, to discuss its main characteristics with stakeholders, highlighting possible limitations, incompleteness and problems, collecting suggestions and feedback in order bring this to its final implementation.

4.9. EOSC as a “skills commons” providing FAIR training for FAIR data stewardship

The EOSC family of projects each addresses skills, from EOSCpilot to EOSC-hub and OpenAIRE-Advance[14]. The overall context was identified in the first EOSC High Level Expert Group (HLEG) report, which highlighted a large gap in data expertise. So EOSC needs to be prepared to meet that challenge, using approaches that scale-up.

The main discussion points were the following:

[12] https://aarc-project.eu/
[13] https://eoscpilot.eu/
[14] https://www.openaire.eu/advance

- The concept of "train-the-trainer" is diffuse: it typically transfers knowledge about particular content, and may also − or mainly − aim at increasing training skills.
- EOSC could have a central coordinating role in "harvesting" and presenting information related to Open Science, Data Science, RDM. This may be more efficient and effective than several infrastructures and organisations doing this individually.
- A core set of curated materials is desirable, to support trainers' capacity to deal with topics where the good practice essentials change frequently, or lack consensus. These should be based on RDA outputs where possible.
- There is a need to keep track of and visualise the popularity of training resources: this partly indicates the value of the resource. Gradually, EOSC coordinating information on training etc. could grow into developing/ suggesting/ validating/ endorsing some quality measurement. However, absence of quality stamps does not necessarily imply no quality.
- EOSC could encourage proactive outreach by the Research Infrastructures towards institutional Research Data Management (RDM) services, helping both sides achieve their goals of broadening access to research communities and stimulating cross-disciplinary research. RI and e-Infrastructures could offer relevant materials and expertise to university research data services, in order to complement and enrich the cross-disciplinary training they offer, and fill gaps in disciplinary-focused materials. To add value, EOSC could for example organise "trainers/experts for hire" across the various projects and infrastructures. This would probably include some training of these trainers/experts.
- Large-scale and long-term research collaborations occupy a middle ground between data-intensive domains and the 'long-tail' of others. Collaboration partners have diverse practices and standards, so they have a strong need for mutual learning.
- e-Research Centres (e.g. Göttingen eResearch Alliance) have a role in coordinating local cross-institutional support, e.g. helping to build individual institutions' capacity to enlist 'data champions' who can address their needs for disciplinary-focused training.
- The National Open Access Desks (NOADs) are important as national ambassadors/intermediaries. The e-infrastructures and RIs often have national representatives, who may be an effective route for EOSC communication. National funding agencies can help with giving mandates, policies and guidance.
- Does EOSC also look beyond Europe? This should always be kept in mind.

The detailed notes of the workshop are available in Annex .

Conclusions reached

- When asked to identify their top 3 priorities for skill development in EOSC the workshop participants' top priority was 'support for training the trainer approaches'.
- The most feasible methods for making training materials FAIR were seen as 'adding identifiers and standard metadata' (findability) and 'non-restrictive licenses' (reuse).
- A marketplace of IT and soft services, offering information on training across Europe and a training catalogue are highly desirable, from user perspective
- National-level policies and funding are essential to be a good researcher and use EOSC in the optimal way.
- Significant willingness and funding from all member states is needed to make EOSC a reality. Fortunately, a lot can be done without much money − but through effort and by gaining consensus! − such as making data and training resources discoverable.

ANNEX A. DETAILED NOTES FROM THE BREAKOUT SESSIONS

A.1. Annex EUDAT what's next? CDI & EOSC

Rapporteur name: Damien Lecarpentier, CSC & Debora Testi, CINECA

EUDAT CDI Moving Forward: The session presented the recent developments of EUDAT in its effort to move from an EC project to a sustainable collaboration between partnering organisations, through the EUDAT CDI Collaboration Agreement. It also provided some information to organisations interested in engaging with the CDI and about the various opportunities to do so, including how to formally join the CDI.

Conclusions reached

The added value of the CDI is now becoming clearer to the audience and the fact that the CDI is an open endeavor welcoming new organisations to join, was seen as a very positive aspect of the collaboration and whole enterprise.

Research communities and the CDI: The session opened discussion on the role of the research communities within the CDI. How can EUDAT best support the involvement of research communities in the CDI? What are the appropriate mechanisms to be able to influence the EUDAT agenda? Those were the main questions discussed.

Conclusions reached

- The active involvement of research communities in the CDI is critical for the success of EUDAT. It is its trademark and should always be pursued.
- A users' board seems a good way to keep users involved in the future of EUDAT services.
- Definition of the exact role and responsibilities of the users' board should be achieved before defining precisely composition and membership of the board.

Q/A and discussion points

The discussion focused on the articulation between EUDAT and EOSC. Will EOSC have a user board (role, membership, terms of office)? And if so what are the complementarities between the different boards?

EOSC-Hub & OpenAIRE Advance: EOSC-hub and OpenAIRE-Advance are the two projects selected as part of the recent EC call e-Infra-12 to lay out the technical foundations of the European Open Science Cloud. This session provided an overview of the projects' work plans and the joint activities being discussed to address some of the upcoming challenges.

EOSC principles of engagement: This session presented work in progress within the EOSCpilot project to define some principles of engagement in the European Open Science Cloud. In particular, the session discussed how to engage as a service provider and presented the development of a set of guidelines to facilitate the findability and reuse of research data within EOSC.

Conclusions reached

Engaging the community on these topics is critical to get their ultimate buy in.

A.2. EUDAT Collaborative Data Infrastructure Services

Rapporteur: Shaun de Witt, CCFE

Presentation highlights: Approximately fifty attendees participated in the first part of the session, where all services were introduced, mentioning their history, current status and plans. Reports were given on the status of the documentation and training materials, while the Q&A part was very lively.

The second part focused on service demonstrations. There was clear interest from the audience, as well as many questions. There is still a string demand for the type of services being offered. One particularly

interesting recurring question was: “How to find and get a “good” long term agreement with an EUDAT service provider”.

Conclusions reached

Certain conclusions have been reached during this session:

- There is still a great deal of interest in the EUDAT services and, coincidentally, the B2 branding is very well recognized.
- The marketplace set up was hugely successful, with many very technical discussions. This seems like a good format for advertising services
- Plans for further service developments are much appreciated, even if it’s not sure how all of the plans will be funded.

Q/A and discussion points

The main discussion points included the following points:

- There should be possible to include search functionality by license type in B2FIND, but since values are textual, it is not clear how useful it is, unless we have controlled vocabulary.
- About the planned improvements for B2SHARE bridge, currently there is limited metadata support and a plan to include adding full set of metadata from within B2DROP.
- On download statistics and whether users have to agree to terms of use, it is already in the current terms of use, but information returned to the users is anonymized. User information is not stored, but what is used, is a hash of IP address. It may be possible to add statistics by country.
- In principle the file size limits on B2DROP can be extended, but the free version of NextCloud supports limited file size. It can be agreed to an extended contract with CSC (service maintainers) for a deployment based on commercial solution.
- B2SAFE adds specific items in PID interface, not technical enforcement. Enforcement is done within service.
- Regarding GEP, the suggestion was looking into DROOLS engine.
- Need to check about an official repository in docker hub for GEF images.
- End users used spark for data discovery with link to B2FIND, and processing is done in spark/Hadoop but data is stored in D2DROP. Could also make use of ArrayDB instead of Spark, butut at the moment there is no EUDAT SPARK or HADOOP interface.
- The official contact point for the CDI is the CDI secretariat via the EUDAT website.

Detailed Notes

The session was attended by 47 people. The EUDAT Service Update was comprised by the following information presented:

Evolution of CDI Architecture by Mark van der Sanden: Addressing cost effectiveness of services is common to many communities. Interoperability was the key to ensure services link to communities’ workflows. The Service Diagram was displayed. CDI links generic and thematic centres. The evolution of the CDI has been presented. There has been an HTTP API addition into B2SAFE in addition to gridFTP. HTTP API will be extended to include additional functionality later. Metadata management has been added into B2SAFE to support publication. Harvesting of B2SAFE metadata has been added into B2FIND vis B2SHARE instances associated with B2SAFE sites. B2NOTE has been added recently to B2SHARE and should be available from B2SAFE. There is also a plan to link B2NOTE to B2FIND. Workspace has also being developed. B2HOST and Generic Execution framework allow execution of workflows in containers. Demonstration implementation is prepared. Another point mentioned was linking B2SAFE to repositories such as Fedora Commons and DSPACE.

Data Access and reuse by Chris Ariyo: The presentation covered B2ACCES, B2DROP, B2FIND and B2SHARE.

- B2ACCESS: Support for personal certificate, ORCIS, etc; IdP support for SAML, OpenID, OAuth2 and X.509; Interoperability with PRACE, EGI, and working on Life Science AAI integration (in collaboration with Geant and EGI); Looking at distributed authentication and harmonization of user and community information inc. fine grained approach.
- B2DROP: Integrated with B2SHARE and B2ACCESS (inc. linking of metadata schema to B2SHARE); Further developments of B2SHARE bridge; Account movement from local B2DROP authentication to B2ACCESS.
- B2FIND: Extended harvesting with JSON and CSW; more than 20 sites already harvested and another 12 in preparation; FAIR principles adhered to; >600,000 entries; Supports faceted searches; Improved interoperability via common metadata schema; Intended to be maintained on EOSC-Hub; Improvements to user experience planned
- B2SHARE: Support of DOIs and PIDs; Installation via docker; Checksum support; Flexible metadata models; Versioning now supported; Records containing data files and associated metadata can be published for each community; Prototype integration to B2SAFE ready for support of larger data sets (planned Feb 2017 for production); Now includes download statistics, automated database migration, automatic fixity checking; Good uptake from community sites on running B2SHARE instances.

Data Preservation Service Area by Claudio Cacciari: The initial plans included consolidation of B2SAFE service and integration of HTTP API, PID service and data policy manager. After one year, the number of requirements exploded based on community feedback. The speaker summarized the new features in B2SAFE, Data Policy Manager, PID system. Lessons learned: Underestimated effort esp metadata management; Users requirements change so need a change management process; This led to very dynamic interaction between developers and users. The results of the comparison between start and end of project made very clear that: DPM can define high level policies; Documentation and training have been improved; Clear workflows have been defined; Integration with B2ACCESS. Regarding future development: Better integration with other services; Policy defined versioning; Metadata management; Better integration with service management framework and data management planning tools; To continue discussions with RDA.

Data Processing & Analysis by Tom Kirkham: The main points of the presentation were the following:

- On B2STAGE: Moving data around internally and externally. Started with gridFTP and now moved to HTTP. The main achievements: Service integration, HTTP API, data discovery service and standards support such as PID.
- On GEF: Builds service around docker; First release was in September 2017; Integrated with ESGF and EGO; Talking to climate for impact.
- On B2NOTE: Enriching metadata, support for text mining; Over 100 active users already; Integration with OpenAIRE is under discussion.
- On Big Data Analysis: Spark and Hadoop are enabled in EUDAT; Data subscription service; Integrated into EUROARGO project; Work in EOSC Hub and SeaDataCloud allows work to continue.

On user documentation and training, a lot of effort has been put into optimizing them. Thirty three documents exist under management framework, while three levels are dependent on users. Communities are active participants in developing documentation. There are forty maintained training modules. Hands on training environments are available, as well as other material on github.

On future service development, the CDI partnership agreement has been signed by organisations with commitment for 10 years support of services. There is also project based input such as EOSC-Pilot and -Hub, SeaDataCloud, and future H2020 calls. On the partnerships agreement, there are twenty two partners from generic and thematic service providers with two levels of membership. Membership fee applies

EUDAT Service MarketPlace and Demonstrations

On B2FIND by Heinrich Widmann: Different levels of granularity were discussed (provide metadata on the level of DOI references!). There are options to expose metadata to EUDAT: 1. via B2SHARE 2. direct harvesting

from community endpoints. Another point was how to improve interoperability of metadata exchange on technology level (used repository software, e.g. CKAN vs. Invenio etc. pp.).

On GEF by Asela Rajapakse: Deployment with B2HOST; orchestration of GEF jobs/Docker containers; trusted GEF services/Docker images with community repositories.

On B2STAGE by Sri Harsha Vathsavayi: Using B2STAGE to move data between B2SAFE and B2SHARE and other EUDAT services; There are options to provide shared access tokens for sensitive data. So, the users belonging to a community can access the data using the shared access token; Development of B2STAGE after EUDAT-2020; installation requirements for B2STAGE.

On Data Policy Manager by Adil Hasan: It would be good to have some sanity checking of the policies that are created to avoid problems. This should be possible with better integration with the DPMT; There was a request to allow the DPM to infer the target (ie. result of the policy) from the source to reduce the typing users need to do; Being able to get the community and role from the DPMT would be good to have.

On B2SHARE by Emmanuel Dima: B2SHARE as metadata store for B2SAFE; Discussion about B2SHARE features for people interested in evaluating the service for an institutional installation (doi configuration, record versioning, etc.); EISCAT usage, with a way to automatically extract metadata for B2SHARE records; Possible integration between B2STAGE and B2SHARE (as we do for B2SAFE); Requests for improvement and new features, and the future of B2SHARE in EOSChub.

On Repository Integration by Jozef Misutka: How to find and get a "good" long-term agreement with an EUDAT service provider.

A.3. Harvesting results from the EUDAT Community: Demonstrators & Pilots

Rapporteur name: Leon du Toit, UiO

Presentation highlights

The main highlight of the presentations was the fact that all the demos show real examples of EUDAT services enabling research.

Conclusions reached

Certain conclusions have been reached during this session:

- In general, those communities that used APIs (in B2SHARE and B2DROP especially) have been successful in implementing solutions that meet their needs.
- The interdisciplinary nature of EUDAT tools is demonstrably useful.
- Implementations were difficult mostly because APIs were evolving throughout the project – bleeding edge.
- B2ACCESS is useful for enabling access and also for reporting on how data is actually used.
- Using EUDAT services provides incentives to standardize tools, data and metadata which saves time when experiments are repeated – sometimes resistance in the short term but this has long term benefits for the community.
- Integration with EUDAT services can foster the renewal of existing community tools.
- Communities are concerned about maintaining transparency during EOSC.
- EUDAT has become an influential player; It is important that EOSC fosters sustainability of services and APIs.
- Service stability before production release is paramount.
- Communities have a need for clear answers about what will happen to the underlying resources after EUDAT2020 (e.g. what will happen to the storage?).

Q/A and discussion points

The main discussion points included the following points:

- There are diverse communities with diverse needs and while common denominators exist, the difficulty is in tweaking the EUDAT services to be relevant, and to overcome the momentum inside communities, which favour using existing tools.
- Service stability and longevity is a still a concern.
- It is unclear what the possibilities are for communities to influence EOSC.
- Over the last 6 years there has been a shift in institutional structure and less transparency in H2020, e.g. the technical committee, this was a marked change relative to the first part of EUDAT. In this light we need to be careful during EOSC.
- Expectations management with the communities sometimes overplayed what EUDAT can deliver on.
- Thought experiment: what if EUDAT was no longer there? Then it becomes obvious that EUDAT has become relevant and influential.
- CLARIN experience: eduGain alone has issues for federated login – CLARIN SP integration did not work due to legal differences; DSpace plugin for B2SAFE – problematic experience with API stability.
- Communication bottlenecks; resource allocation is difficult to manage.

Detailed Notes

On Community demos: ELIXIR-EUDAT-EBI data distribution service; Datasets: Version, describe, automatically get updates.

On FAIR data pilot: To assess the FAIRness of B2SHARE; There is a problem in improving the FAIRness of B2SHARE – to make it behave as a FDP; To use REST API – keep it decoupled; To improve semantic interoperability; to provide mappings; Solution – proxy to B2SHARE API via B2SHARE-FAIR.

On Dendro: Project creation with B2DROP as storage backend.

On Pairqurs: Air quality, crowdsourced data collection; to use B2SHARE to make data discoverable; Longevity is important.

On ENES and EUDAT: B2SHARE (intended use case, single repo for multiple projects) − Multi-tenant hosting difficult; B2HANDLE − ESGF PID hierarchy; B2FIND − 1000+ records harvested − metadata records refer to DOIs to an ESM experiment; GEF.

On CLARIN: Tool switchboard – plugin for B2DROP; Extensibility of B2DROP via ownCloud a unique selling point – easy to integrate with external services; Switchboard also generic, can be used for many disciplines.

On LTER: Problem: ensure data availability after project end; CMS: B2SHARE-Drupal integration.

Conclusions:

B2SHARE very useful; v2.1.0 released to late, data versioning came too late and was in high demand; too many resources needed to develop this simple module (immature API at the beginning; lack of docs; server downtime; breaking changes in the API).

On Eiscat3d: Need for unified access to data; Different levels of data; Pilot: B2SAFE integration; We cannot get content related metadata out of B2SAFE; To use B2SHARE as a metadata repo for L2 and L3 data; To use PIDs; To develop their own metadata extraction client.

On EPOS: Seismology; Beta and in prod: b2access, safe handle, b2stage http api, gef; Non-prod: prov, version array DB; B2ACCESS fundamental ingredient – for access and for metrics.

On ICOS feedback: Need safe LT storage of observational data; Need to move this data, catalogue it; Expose metadata through other catalogues; Mixed experiences: some features were developed late relative to needs, good discussion and collaborations; did not have any impact on development of services because had no PMs there; concerns for the future: sustainability and availability.

On HerbaDrop:EUDAT as a preservation infrastructure; 500 million specimens still need to be digitized; herbarium images; use: b2safe, b2hande, b2find.

A.4. Cross e-infrastructure collaborations

Rapporteur names: Matthew Viljoen, EGI; Michaela Barth, KTH; Vasso Kalaitzi, LIBER Europe

Coupling data and HPC resources together to enable large scientific projects: the EUDAT-PRACE case

Making data and cloud resources interoperable using EUDAT and EGI services

Presentation highlights

- EGI Open Data Platform live demo (Matthew Viljoen)
- ICOS use case demo (Margareta Hellström)
- ENES use case live demo (Xavier Pivan)

Conclusions reached

- Recommendations given on how to improve documentation and further testing prior to tools being made available to early adopter communities
- Recommendations given on how to best on-board further user communities
- Work reported in
  - https://eudat.eu/deliverables:
  - EUDAT2020-DEL-WP7-D7.1. describing initial EGI-EUDAT pilot activity
  - EUDAT2020-DEL-WP7-D7.4. Final Report EUDAT/EGI)
  - Corresponding EGI Engage deliverables (with input from EUDAT):
  - EGI Engage Deliverable D4.8 “Cross-infrastructure case studies report”
  - https://documents.egi.eu/public/ShowDocument?docid=3026
  - EGI Engage Deliverable D4.9 “Open Data Platform: Demonstrator, Experience Report and Use Cases”
  - https://documents.egi.eu/public/ShowDocument?docid=3033
  - https://github.com/EUDAT-GEF/GEF for documentation on how to use the GEF backend

Q/A and discussion points

- Should the services provided be on a Google play or amazon shop level or is engagement from the RI expected to really contribute to the whole service portfolio?
  - A: Both, for onboarding new RI default FAQ and more guidance should be provided with the traditional e-infras providers’ services really being on a level that allows easy adoption from new communities, for other services a more collaborative and adapted approach might be preferable
- A general point was made on the role of the Research Infrastructures − not only as users, but also as service providers

Detailed Notes

ICOS STILT Footprint tool will make the sensitivity optimized position planning for measurement stations possible.

Test implementation of on-demand calculation and visualization of footprints and time series − still under development:

https://stilt.icos-cp.eu/worker/
https://stilt.icos-cp.eu/viewer/

Computing e-Infrastructure with extreme large datasets

Presentation highlights

On EUXDAT: is the European e-Infrastructure for extreme analytics in sustainable development, mainly focused on agriculture. Its main objective is to build a Large Data Analytics-as-a-service e-Infrastructure by

connecting extremely large and heterogeneous data sources expertise from various disciplines and results from past partners to provide analytics tools and services for these different kinds of users to support sustainable and productive agriculture. There are three pilots in this project: 1) Resources optimization, 2) Energy efficiency, 3) 3D farming. The platform, apart from data, will provide a data collector.

On DARE: Delivering Agile Research Excellence on European eInfrastructures. Its objectives are: 1) To deliver a new working environment for scientists and research developers (presenting methods in abstract terms, so that domain experts can understand, change and use them effectively and providing tools that visualize the runs of these methods in summary forms), 2) To improve further and integrate (tested programmatic dataflow specification APIs, big data technologies, provenance/data lineage solutions), 3) Work with two RIs (Epos and IS-ENES). Components in their working environment: dispel4py, S-ProvFlow, Exareme, Semagrow, BigDataEurope.

On DEEP-Hybrid-DataCloud: Designing and Enabling E-Infrastructures for intensive data Processing in a Hybrid DataCloud is a Platform-driven e-Infrastructure towards the EOSC. Its scope is a Computing e-Infrastructure with extreme large datasets. Its global objective is to promote the use of intensive computing services by different research communities and areas, and the support by the corresponding e-Infrastructure providers and open source projects. A powerful developer tester will be provided by the project partners.

On XDC: eXtreme DataCloud is about Foundations. It takes the move from the INDIGO Data management activity and the experience of the project partners on data-management. It aims to improve already existing, production quality, federated data management services, by adding missing functionalities requested by research communities. It must be coherently harmonized in the European e-Infrastructures. The New Functionalities: Intelligent and Automated Dataset distribution; data pre-processing during ingestion; data management based on access patterns; smart caching; metadata management; sensitive data handling.

On FREYA: Persistent Identifiers. It builds on THOR (which is built on ODIN). Three pillars of FREYA: PID Graph, PID Forum, PID Commons. There is close cooperation with RDA and own disciplinary pilot applications.

Conclusions reached

The main conclusion reached was that overlaps and complementarities need to be further discussed. Channels need to be kept open for further collaboration.

Q/A and discussion points

The main discussion points included the following:

- EUXDAT is about land-planning for agriculture with open data. There is a relationship between EUXDAT and EUDAT. There are services that can be used by the project. Compliance needs to be checked. One of the first tasks is to collect the requirements for the pilots. The channel needs to be kept open, so that there is no duplication of effort. EUXDAT went for Spark because the consortium is relying on existing expertise.
- The HPC resources in DARE are provided from partners involved (GRNET & Fraunhofer). There are containers in DARE partially.
- Regarding DEEP, the exploitation will be promoted through the EOSC framework – a channel to the EOSC-hub, in order to take the services and integrate them to the EOSC-hub. The Cooperation Board needs to be discussed with the DARE project as well. DEEP and EUDAT are under a different umbrella, that of the XDC (eXtreme DataCloud), so the collaboration is not obvious. However, the analysis of the existing overlap with EUDAT is still ongoing.
- Regarding FREYA, DIP Graph the linking between different types of objects with different persistent identifiers with CRIS as a basis. The services in EUDAT have been following a different direction. There is a need to keep open channels with the EUDAT CDI and the EOSC-hub. The scope of the Graph is still a little unclear, it's not a universal graph populated by harvesting everything.

Detailed Notes

The Scope of this event was to engage these developer projects in a dialogue with EUDAT. What follows, is a list of five projects and their key focus:

EUXDAT is the European e-Infrastructure for extreme analytics in Sustainable development, mainly focused on agriculture. Its partners come from different fields: Industry, University, Research, SMEs, End-users. The main objective is to build a Large Data Analytics-as-a-service e-Infrastructure, by connecting extremely large and heterogeneous data sources expertise from various disciplines and results from past partners to provide analytics tools and services for these different kinds of users to support sustainable and productive agriculture. The aims of the project are: 1. To manage data storage and movement & support heterogeneous data sources & configurable policies, 2. Adapt data processing tools for HPC & improved users' portal & new resources management (hybrid HPC & Cloud), 3. Service activities (access to the e-Infrastructure) & pilots implementation & access to data, 4. Networking activities & long-term sustainability & collaboration (i.e. PRACE, EGI). 5. To open this platform to external communities for them to benefit from it. There are three pilots in this project: 1. Resources optimization, 2. Energy efficiency, 3. 3D farming. For this purpose the project will look in the past and also use the data produced daily. Eight years in the past are needed in order to make the forecast. Drones are producing the biggest source of data (10-15 GBs per farm). Information produced from other initiatives should also be used. The platform, apart from data, will provide a data collector. The project is relying in OGC interfacing. External communities will have access to the data produced by the platform. The deployment orchestrator decides the processes. Tools: as much as possible open source: Jupyter/Spark/Hadoop/Geotrellis/geomesa/pCEP.

DARE is Delivering Agile Research Excellence on European e-Infrastructures. The motivation is the size and complexity of scientific data, the difficulty in the formulation domain-specific solutions in reproducible and reusable ways, the fact that big data technologies and analytics are often not taken advantage of and the advances in exploiting huge datasets, depend on synergy between all categories of experts involved. The objectives are: 1. To deliver a new working environment for scientists and research developers (presenting methods in abstract terms, so that domain experts can understand, change and use them effectively and providing tools that visualize the runs of these methods in summary forms), 2. To improve further and integrate (tested programmatic dataflow specification APIs, big data technologies, provenance/data lineage solutions), 3. Work with two RIs (Epos and IS-ENES). The project aims to empower those who develop methods and their encapsulations, so that they can help their communities and deliver a unified context for their work, in order to support innovative data-driven science and make transparent use of e-Infrastructures and other resources. Regarding EPOS community: 1. Interaction with EPOS to provide high level services, 2. Drive HPC simulations and data analytics (Boundary of Data and CPU intensive computation), 3. Ensembles of simulations, 4. Rapid assessment of earthquakeimpact for emergency coordination, 5. Rapid characterization of seismic sources. Regarding the IS-ENES community: 1. Ensembles of climatic simulations, 2. Understanding and communication of uncertainties, 3. Meaningful federation over heterogeneous data sources, 4. Streamline data lifecycle, 5. Support Climate4Impact (C4I) Services. The components that exist in their working environment are dispel4py, S-ProvFlow, Exareme, Semagrow, BigDataEurope.

DEEP-Hybrid-DataCloud is Designing and Enabling E-Infrastructures for intensive data Processing in a Hybrid DataCloud. This is a platform-driven e-Infrastructure towards the EOSC. Its scope is to be a computing e-Infrastructure with extreme large datasets. Its global objective is to promote the use of intensive computing services by different research communities and areas, and the support by the corresponding e-Infrastructure providers and open source projects. Its objectives are: 1. Focus on intensive computing techniques for the analysis of very large datasets considering demanding use cases; 2. To evolve up to production level intensive computing services exploiting specialized hardware; 3. To integrate intensive computing services under a hybrid cloud approach; 4. To define a "DEEP as a service" solution to offer an adequate integration path to developers of final applications; 5. To analyse the complementarity with other ongoing projects targeting added value services for the cloud. About the DEEP pilot use cases, there are three techniques of wide interest: deep learning models, post-processing and on-line analysis of data streams. The DEEP Consortium includes 9 academic partners, 1 industrial partner, 6 countries. During the presentation, a description of the

work packages and project governance took place, as well as a discussion on a collaborative Advisory Board. About the INDIGO components and evolution: INDIGO Orchestrator, Infrastructure Manager, uDocker, Cloud Information System, OpenStack/OpenNebula, PaaS layer, Docker, Ansible, INDIGO Virtual Router. Work Programme includes the following: Plan and requirements (11/17-01/18), Initial design (02/18-04/18), 1st prototype (05/18-10/18, 2nd prototype, full pilot testbed, promotion and exploitation (2020): To improve the support and final quality of the solutions and promote the exploitation in the EOSC framework following the integration activities. A powerful developer tester will be provided by the project partners.

XDC is the eXtreme DataCloud: Foundations. It takes the move from the INDIGO Data management activity and the experience of the project partners on data-management. It aims to improve already existing, production quality, federated data management services, by adding missing functionalities requested by research communities. It must be coherently harmonized in the European e-Infrastructures. The consortium includes 8 partners, 7 countries, 7 research communities represented and EGI. The New Functionalities are intelligent and automated dataset distribution, data pre-processing during ingestion, data management based on access patterns, smart caching, metadata management and sensitive data handling. Smart Caching aims to develop a global caching infrastructure supporting: 1. Dynamic integration of satellite sites by existing data centers; 2. Creation of standalone caches modelled on existing we solutions; 3. Federation of the above to create a large scale caching infrastructure. The Metadata handling use cases are those of LIFEWATCH, CTA, and ECRIN. ELG is responsible for maintaining active relationships with the infrastructure and technology providers, discussing synergies, strategies, roadmaps and requirements workflow for the software released by the project. The plan for the next couple of years and the main milestones include: research community requirements for new functionalities collected; research communities' requirements analysis performed, project architecture detailed, development schedule defined, Event with User Communities, XDC reference release 1, XDC reference release 2, Functionalities and scalability demonstrated.

FREYA is about Persistent Identifiers, builds on THOR (which is built on ODIN) and has three pillars: PID Graph, PID Forum and PID Commons. Datasets become extremely large due to: 1. Breadth, 2. Depth (taking more data over the same scope, more sensors for example, reference to SeaDataCloud,scaling-up of the dataset 3. Density: higher granularity and 4. Heterogeneity and interconnectedness (p. ex. Digitized recordings, material from old newspapers). The implications for the PID Graph include subsets of large datasets, provenance and attribution, possible new entities in research discourse e.g. "analysis", usage patterns: formulating and testing research questions, Borders between in-facility data management, in-community data handling, and cross-science pieces of scholarly communicable units. FREYA aims to have close cooperation with RDA and own disciplinary pilot applications.

A.5. The impact of the policy framework on EOSC

Rapporteur names: Heli Autere, CSC & Rob Baxter, EPCC

Presentation highlights

- A catchy introduction to the essentials of GDPR
- An up-to-the-minute summary of progress towards GDPR-compatible research Codes of Conduct.
- A report on data tagging, an approach to help repositories manage their sensitive data assets.
- An example of a safe haven service as one way to facilitate research on sensitive data.

Conclusions reached

- Data sharing for research just got a lot harder!
- Sending tools to data may be the best approach for now.
- Promoting standards and common approaches will help, even if sharing data is hard.
- The EOSC-hub collaboration of data hosts and data researchers might carry enough weight to lobby for easing restrictions for research data sharing.

Q/A and discussion points

- Impact of GDPR on individual organisations.
- Impact of GDPR on research.

Detailed Notes

The policy framework: GDPR and All That

The early session comprised four talks, beginning with the policy framework of the new General Data Protection Regulation, then taking a look at progress towards research codes of conduct in the life sciences, followed by a report on a pilot study applying the DataTags model to GDPR and concluding with an introduction to a data safe haven service as a possible type of future EUDAT service.

5 things you should know about data protection, David Foster, data protection practitioner, CERN

David gave a lively and very enjoyable talk from the practitioner's viewpoint of the new data protection regime we find ourselves in. His summary slide captures the essence of his talk (and his memorable limerick!):

- My personal data are mine: privacy notices should declare what, how and why data are processed.
- To abuse them is a crime: the definition of personal data is wide, the scope of processing is broad, this is complex to communicate inside an organization, it may help to consolidate processes and infrastructure, be vary of automated decision making and profiling.
- You cannot share: without safe guards because privacy travels with the data, impacting extra-territorial research. There may be a difficult culture change within organisations used to freely sharing personal data, complexity may increase with the incoming ePrivacy Directive.
- You must take care: you need to look after people's data; individuals have fundamental right to their data which you might be processing; there must be clear mechanisms to exercise the 8 basic rights, which should be in service's privacy notices and must be transparent to the 'owners'. Privacy by default and by design!
- Or risk a hefty fine: the new regime is all about managing risk. Fines can be large depending on the infraction, so mitigation of the risk of large fines is the name of the game.

He noted new advice from the Article 29 Working Party of European information commissioners: consent has now been interpreted quite strictly. As one consequence, consent is highly unlikely to be a legal basis for data processing at work, because consent must be specific and freely given (and workers are not entirely free to give consent if their job might be at risk).

As a fundamental principle, once consent expires, you must get the rid of the data, not retain them any longer than necessary. This will be a real headache for a lot of organisations!

Typical reactions this is just administration, so it does not concern me. This is unlikely to be true! Typical generic data collection, collecting too much data, use of unsecure mechanism, processing data without controls – all these practices will need to change.

Research data codes of conduct: status and roadmap, Petr Holub, BBMRI-ERIC data manager

Petr Holub, fresh from a code of conduct writing meeting, brought us up-to-date with current efforts to create consensus on personal data handling in life-science and biomedical.

Petr noted that most of the "GDPR panic" is at organizational level, not at researchers' level. Data in staff or student databases tends to be the immediate concern of most organisations right now.

In GDPR terms, a Code of Conduct (CoCo) is 'soft law' under GDPR article 40, with the flexibility to be updated more often than GDPR (although changing CoCo still requires approval of the European Data Protection Board [EDPB]).

Why do we need Codes of Conduct? To help to achieve and demonstrate compliance with GDPR in using personal data in research contexts.

GDPR has significant national components, and CoCos provide one mechanism to harmonise practices across European countries. There isn't just one CoCo, of course; your research might involve the use of multiple CoCos, for instance.

How the CoCo process works: the EDPB should be consulted as a part of writing process, the EDPB being a body defined under GDRP Art 68, although currently it's not clear who will be on it.

Petr briefed on the current focuses and timeline and gave some examples of CoCos, a good one being the GEANT CoCo led by Mikael Linden from CSC, which focuses on AAI data controllers (anyone who has to run an LDAP service, for instance). The GEANT CoCo is quite well advanced, with a public consultation period starting in February 2018. There are also other examples in cloud computing.

There was discussion on whether a university should do their own CoCo or follow some recognized international organization's CoCos? The recommendations from the discussion was that first to follow a national code of conduct. Antti Pursula from CSC noted that the Nordic countries have created a CoCo from the service providers' point of view.

DataTags for GDPR – a pilot, Heiko Tjalsma, DANS

DANS handle a lot of personal data as research objects – recorded interviews, for instance – and are interested in useful tools and methods to make management of data like this just a little easier.

Just to remind us: GDPR comes into force 25 May 2018, along with complimentary national laws for research and archiving, CoCos (national and/or European) etc. Heiko also noted the guidelines on consent and transparency from the EU Article 29 Working Party. This policy environment – privacy be design, data minimisation, rights to access, etc. – make support tools essential.

The DataTags project originated at Harvard in the US and was first developed as a way to mark data objects with a coloured "tag" which defines how it should be handled (encrypted, subject to access authorisation, etc.). Over summer 2017 DANS undertook a pilot study on applying the same approach to GDPR.

The outcome was a wizard-style questionnaire built from a decision tree and available online[15].

Data safe havens: a future EOSC service? Rob Baxter, EPCC, the University of Edinburgh

Rob presented the concept of a *safe haven*, a secure environment for research work on sensitive data. Could such services be brought into EOSC?

EPCC operate the Scottish National Safe Haven for the National Health Service as part of the Farr Institute for health informatics research. It's a secure compute and data environment – virtual desktops, locked-down environments, two-factor authentication, no outbound network access, managed channels for data ingress and egress – but this is the easy part. The service operates under the UK's Caldicott Guardianship framework for information governance in health research.

The big challenge for EUDAT and EOSC is matching national information governance mechanisms like Caldicott across borders, and putting in place the necessary people to manage the required approval processes. The safe haven service is much more than just a rack of kit with an Internet connection!

Restricted data in the EOSC: What are we going to do?

Francesca Iozzi gave an update from the EUDAT Sensitive Data Working Group which had met over the previous two days. Key highlights:

- Different communities, not surprisingly, have different new needs.
- There are a lot of challenges, both technical and ethical.
- The technical challenges are not the problem! (Echoing Rob's talk or earlier.)
- There are specific/regional solutions available, and one key issue is: should the data be moved or the tool?

[15] https://zingtree.com/host.php?tree_id=442670046

- How can we ensure reuse of data for research purposes?
- EU strategy and recommendation on principles, protocols and best practices are called for.
- We need more awareness-raising, and there is a clear need for cost-effective solutions: an EU-health information infrastructure?
- An inescapable conclusion is that we'll need to move the tools, not the data.

The session was opened up to the floor, with the earlier speakers forming an informal "embedded panel" of experts within an audience of experts.

There are clearly emerging national strategies to re(use) of sensitive data, though with different ideas on central vs distributed sensitive data storage. There is a need for proper processes for vetting, access etc.

Unfortunately, the Sensitive Data working group is not continuing under EOSC-hub. Some similar kind of forum was suggested as useful, perhaps at RDA?

The question was posed: how should EOSC support research with restricted data? How do we deal with information governance? Harmonization of national frameworks? What can EOSC implementers do to influence or support the agenda?

There are commercial aspects to be taken into consideration and user engagement was called for. There are already commercial secure clouds for banking, health care, but not research.

In the upcoming EOSC-hub project, what needs to demonstrated, and defined? Compliance can really only be tested in court (and nobody wants to be first!). We need effective risk assessments; GDPR is arguably overregulation for research data.

The status for international organisations is very difficult (e.g. CERN, EMBL, ESA). Life sciences, botanical garden, working groups have been working for many years on data sharing. Again, the Nordic countries have a lot of collaboration.

Possible approaches could be:

- Secure cloud solutions, safe havens or similar technical solutions.
- Pooling data on the fly, not storing in another country but just for analyzing as it "passes through". (This is still data processing under GDPR, of course, so might not help.)
- There are plans to test international pooling of medical data within the Nordics, but this will not probably scale.
  - How did the Nordic countries do that? By engaging the governments and EDPB.
- Europe-wide, who has the best of best practices? The Datatank data managing tool was mentioned

Some see the GDPR as just another cost item to organisations, but other organisations see the benefits. A quick straw poll across the audience showed that people generally see real benefits in GDPR legislation.

So, within EOSC-hub, service providers will need to plan carefully for their obligations under GDPR (tools like DataTags might help). In terms of enabling research with "sensitive" data:

- Tools to enable service providers to manage their data more easily could be the easy win for EOSC-hub; might even be possible to adopt standard approaches across disjoint services.
- Services like existing safe havens might be opened up to international researchers (this is already possible in places). This could be straightforward.
- Linking and combining data from multiple national sources (in, e.g., a safe haven) is technically straightforward but requires trans-national agreement on data transfers. Do we have right standards? (Probably not.) Could CoCos prove sufficient to enable this under GDPR rules?
- Enabling the broad sharing of "sensitive" data is probably just not feasible – which took us neatly back to David Foster's opening limerick:

My personal data are mine; to abuse them is a crime; you cannot share; you must take care; or risk a hefty fine!

A.6. User engagement & training

Rapporteur name: Ellen Leenarts, Marjan Grootveld (DANS)

Presentation highlights

The main highlights of the presentations included seven Hooray and seven Horror stories about training. A phrase that summarises the presentations is: "You have to enjoy the transfer of knowledge; otherwise get another job."

Conclusions reached

The conclusions reached include two main points:

- When we coordinate or provide training, building trust at all levels and targeting the right audience are essential – the human factor may be more important than content that perfectly matches a learning need.
- To allow us to organise and deliver training at the scale that the EOSC requires, a Community of Practice for training coordinators/training managers may be a valuable instrument for sharing experiences and resources.

Q/A and discussion points

The discussion that took place, focused on two main points:

- How to find or define the "right audience" for a specific training event or training resource – or vice versa?
- Would a Community of Practice be a good instrument to share experiences among training coordinators?

Detailed Notes

The two slots in this session had different topics: "Training experiences across Research Infrastructures" and "Building a trainers' Community of Practice (CoP)", respectively. The EUDAT training[16] team thanks all speakers and participants!

Training experiences across Research Infrastructures

The objective of this session was to investigate tools, techniques and topics for and in training which have proven to work well, or have been shown to be less successful. To this end the EUDAT training team had invited training experts from various e-infrastructures, research infrastructures, an ERIC and a regional research alliance. The speakers had been instructed that all talks should focus on one 'Hooray' story and one 'Horror' story and last no longer than 5 minutes. They should also briefly introduce the infrastructure they were representing and the primary audience for training, and close with a summary of lessons learned or practical advice.

Five-minute presentations were mixed with discussions:

- Olivier Rouchon, PRACE
- Elly Dijk, OpenAIRE
- Gergely Sipos, EGI
- Celia van Gelder, ELIXIR
- Ivana Versic, CESSDA
- Sven Bingert, Göttingen eResearch Alliance
- Alice Fremand, University of Strasbourg
- Michele De Rosa, Bonsai

[16] https://eudat.eu/training

- Shaun de Witt, EUDAT
- Discussion with the audience: which lessons can we draw?

Because both Göttingen eResearch Alliance and EUDAT had selected their respective 2017 summer schools for the Horror and Hooray stories, the organisers decided to cluster them in the same round. In addition, the session organisers had invited two former EUDAT summer school[17] attendants and asked them to feedback on this training event – they were explicitly encouraged to be critical.

Despite very interesting horror stories and good criticisms the team has intentionally drawn only positive conclusions from this first time slot. These are the main ones, formulated as recommendations:

- Target the right audience. This may sound obvious, but requires clear two-sided communication ahead of the training event.
- Make sure that the training topic and content is relevant for the audience. When possible, adapt content to for instance the research domain or the attendants' responsibilities (researchers versus research supporters). Make the content more easily understandable and applicable by means of use cases.
- Build on the existing knowledge of the participants; if necessary, ask ahead of the training. Unless you intend for advanced participants to guide the juniors, don't mix them.
- Be clear about prerequisite knowledge. Especially for larger events, distribute reading materials ahead of the training event. For instance, for a one-week or two-week summer school, allow a month time for reading and for installing the required software.
- In particular, when your training organisation is highly structured – e.g., covers several countries with their own trainers, affiliated with different organisations and/or research domains and/or national agendas – put effort into building trust on all levels. Be transparent.
- Provide clear guidelines for teachers and trainers you invite, on what is expected of them. A good lecturer is not per se a good trainer; the developer of an IT service is not per se a good instructor.
- Plan enough time for what you want to achieve during the event. Hands-on sessions may need more time due to unforeseen (untested?) circumstances. Plan also buffer time, for "debriefing", individual exploration of a tool, et cetera.
- In larger training events, plan social activities and allow time for impromptu meetings and one-to-one conversations. Much learning happens outside of the classroom.

Building a trainers' Community of Practice (CoP)

Sharing and reflecting on all the information provided by the speakers in the first time slot was very valuable. Actually, the EUDAT training team had expected this, and considers this kind of informal exchange among training experts crucial for the training challenges that the EOSC presents to all of us. For this reason, the team planned to explore, in the second time slot, how useful a so-called Community of Practice might be. After all, many of us are not professional, full-time trainers, and having a knowledgeable network is important.

A good introduction to CoPs is http://wenger-trayner.com/introduction-to-communities-of-practice/. As a working definition we said it is:

- A group of practitioners who share a common passion and want to do it better;
- There is no formal membership; people are free to join and leave when their situation changes;
- Informal – no chair, secretary, quorum, or voting required;
- Meetings are ad-hoc and minutes optional (but useful).

[17] https://eudat.eu/news/eudat-summer-school-where-data-management-meets-the-next-generation-of-researchers

In break-out groups the session participants gave several answers to our three questions:

- What keeps us as trainers motivated? E.g. enjoying the transfer of knowledge, learning from training participants and getting feedback from them; updating our own knowledge about the training topic; meeting other trainers. “If a trainer’s loss of motivation comes from routine, change that.”
- If we formed a CoP, what would you contribute and want to get out of it? E.g. pointers to training resources and to trainers; networking; sharing good practices (examples, strategies, etc).
- Does the CoP need a goal, and if yes, which? E.g. “I can’t defend something informal to my manager”; it doesn’t need a goal, but it needs a raison d’etre; no, a single point of entry to training experience and experienced trainers is valuable enough.

It was concluded that a CoP looks most promising for training coordinators and training managers. Through the EUDAT newsletter the session organisers will send an invitation to help explore what the passion and raison d’etre can be, and where to loosely embed it – maybe in the joint activities of the EOSC-hub and OpenAIRE Advance projects.

ANNEX B. CO-LOCATED EVENTS

B.1. Sensitive Data workshop

Session 0: Introduction

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Introducing the sensitive data working group
- Objectives: Collecting experiences, evaluate technical solutions, evaluate challenges connected with the use of sensitive data; the final aim is the facilitation of joint research with open and sensitive data
- Describing the Agenda

Session 1: Introduction: The SW working group and the sensitive data landscape

Rapporteur name: Abdulrahman Azab, Wolfgang Kuchinke

Presentation highlights

- What are EUDAT working groups?: Instrument of working groups to work on data topics
- Sensitive data working group:
  - EUDAT working groups are characterised by communita engagement and regard to requirements
  - First WG sept 2016 (Krakow), and Second Feb 2017 (Oslo), Third Jan 2018 (Porto). New communities that have to deal with sensitive data were invited. EUDAT WGs are channels for requirements and in this sense the Sensitive Data Group has raised awareness for the requirements for sensitive data. Now the constraints for EOSC are added.
  - Directions: B2share, B2Find, and B2Safe to be integrated on sensitive data. It was tried to initiate a project to enable B2SHARE to use sensitive data; this project was not supported by EUDAT, and it was not possible to find a right partner.
  - Results: During a Council meeting several PMs were approved that have been used, for example for the workshops and to analyse the landscape. A paper and a Memorandum will be written about the results.
  - Future: Try to influence the agenda of EOSC Hub; sensitive data belong into the hub. In general sensitive data management needs to have more attention; the group will contact service providers, policy makers and will engage in RDA for this purpose. The next group meeting will be at an RDA event in Berlin. A proposal for a sensitive data BoF has been made.

Conclusions reached

- Existing tools and services are sufficient, the level of activities must be policies and processes.

Session 2: Sensitive data and Linguistic research

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Presentation outline
  - Linguistic research:
  - What is linguistic research? It deals with any kind of language issues.
  - Why doing research in linguistic?
  - Different research methods are used in linguistics
  - Research aims: language processing, language acquirement, dealing with different languages
- Dimensions of linguistic data: protocol of data connections, medium used for recording data, aggregate data
- Metadata: Information about the data collections and the language users
- State of affairs in linguistics: raised awareness of FAIR data
- Collecting and sharing different types of data: Primary data, added data, aggregate data, metadata

- When is data considered sensitive?:
  - Permits identification of data subjects
  - Includes personal information
- RDM: there are current best practices in linguistics, but more to be done
- GDPR: It changes the rules for research (mainly keeping ethical standards). GDPR has made it not easier; e.g. what happens when someone wants that data is retracted, results may become unreproducible. In addition, often no anonymisation is possible; for example in the case of recording the signatures.
- Best practices in the light of GDPR:
  - Technical infrastructures need more work for secure data storage, encryption, authentications; technical and organisational measures for sensitive data are well under way
  - Raised awareness for FAIR; attention to data integrity,
- Recommendations: make a Research data management (RDM) portal, including the training of researchers; work on the consent forms to make it easily understandable; involve both experts and stockholders, continue working on GDPR, further develop the technical infrastructures.
- Research data management must play a role throughout the entire data life cycle
- Use data prompts for different types of data

Session 3: e-Infrastructure for Video Research

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Focus: eVIR project, dealing with sensitive video data
- Five key partners: NSD, USIT/TSD, Dept. of Psychology, Dept. pf Music, Dept. of teacher education, teaming with ICPSR, Uni Michigan, NordForsk
- Research focus and ambitions:
  - Safe storage, metadata standards, archive solutions, lab solutions for recording and training
  - National roadmap exists for an infrastructure for sensitive data
- Lab solutions, e.g. portable labs, virtual labs (on the cloud).
- Safe storage of videos
- Being flexible, generic, think about the entire workflow; very wide spectrum of users
- Main challenges: Ethical regulations and requirements, in Norway participation is optional
- For video storage – collaboration with Center of Sensitive Data;
- video capturing quality (both video and audio), free access to metadata (for metadata to be findable, accessible, and interoperable),
- Data Management Plan, should consider the whole workflow, aim for a national DMP
- storage solutions (remote/cloud solutions: BW, latency, and administration), coding (there are many software solutions, what to choose and certify).
- Types of data:
  - LISA study (linking instructions and students)
  - LISA Professional education
  - LISA Professional developments
- Metadata: Information about the data collections and the language users
- Governance challenges
  - Students have to be contacted for consent (vulnerable); aim for a dynamic consent
  - Right to withdraw; solution on a national level
  - Secondary use access to metadata
  - Consent bias (exclusion of certain groups
  - The “ethical requirements” challenge
    - Parents approval requirement for individuals below the age of 16 (in Norway)
    - Right to withdraw
- Data collection challenge

  - Plan for the entire workflow
  - Uae od Metatagging templates, linked to data
  - Quality: multiple activities are recorded, problem of bandwidth
- Storage solution challenge
  - Use of secured servers
  - Access to cloud services, virtual storage solutions when analysing data (TSD), this allows not to move data outside of Norway

Conclusions reached

- There is a wide range for using sensitive video data in research
- In this direction, we are moving forward
- Main issue: ethical requirements
- Technical challenges are there but solvable on the long term
- TLVLab Software
  - Use of video analysis
  - Working with sensitive data, faces are not anonymised

Q/A and discussion points

- Comment: B2share TSD pilot for publishing metadata of sensitive data in TSD
- In the US they may trust Amazon cloud for sensitive data
- How the "right to be forgotten" will be integrated with preserving the data integrity? TBD. The purpose of the consent is to describe for what data is being used. Allow for a informed choice. Aim is to have a national contact point for consent.

Session 4: ECRIN Pilot

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- CORBEL project: Establish a collaborative framework of shared services for data sharing between ESFRI infrastructures
- Individual participant level data (IPD); this is the raw data of clinical trials participants
- Clinical trial services: architecture includes interconnection between external services (biobank), and ECRIN clinical trial (centered at the ECRIN management office).
- Community driven use cases; definition of a framework for opening access and reuse for clinical trials data
- Principles and recommendations: mainly sharing and reuse of IPD. Preparation of the development of an open clinical trial data repository
- EU prospective: A more EU prospective is required
- Consensus document on providing access to individual participant data from clinical trials [2017](10 participants, 50 recommendations)
- Requirements: Environmental, non-functional, and functional requirements were extracted from the consensus document
- Tools and services: a list of tools and services was extracted from the document (e.g. Data Sharing Plan, metadata schema, de-identificaton service, data repository, boards for overseeing data sharing…
- Findings in context (what is needed):
  - Need to develop support systems planning and preparation of data sharing
- What needs to be further investigated: Level of Individual Participant Data and document sharing, cost and time for preparing data for sharing.
- Attitudes of different stakeholders; reasons why researchers don't share data
- Types and quality of different research output
- Comparison of different access regimes
- Incidence of any misuse of data in research

- Questions about Repos: should we have fewer but larger repos? involve specialists in data storage?
- Central portal for identification of trial data stored in different repositories / registries
- Future role of data repos:
  - More Useful tools for sharing in many domains
  - Further development necessary for clinical data
  - Further harmonisation necessary on the level of scope and policies and business models for long term sustainability

Conclusions reached

- The principles and recommendations are globally relevant for data sharing
- Some recommendations may not be applicable globally (e.g. role of de-identified data in the US)
- Applicability is possible in countries outside Europe while in EU it is more demanding.

Q/A and discussion points

- What are the technical limits? Were principles evaluated? Any proof of concept?
  - There is a prototype of a repository under development in CORBEL project. Clinical trials data storage and sharing is tested as part of an EUDAT pilot with B2SAFE, B2SHARE and B2FIND, including testing in TSD.
- Should be One repo or several?
  - Several repos that can talk together is a better approach, especially if one considers that ECRIN clinical trials data is international data.

Session 5: Balancing public access to sensitive data

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Researchers can use instruments for measuring sensitive data
- Roaming a city within 30 months with 200 volunteers, 13 million measurements
- The PAIRQURS case: Citisens are measuring air pollution; loads of personal automated data (200 volunteers, 20000 trips, 77000 hours and 13 Mio. measurements) biomedical data, location data are personal data. Results are shared with B2SHARE.
- Sensitivity types:
  - data centric (e.g. fitness for use, transitive sensitivity),
  - person centric (locating individual), release presents new data (KPK), person-centric sensitivity
- Classical solutions:
  - data centric: document with metadata, …
  - person centric: anonymization, cloaking, and aggregation
- Trade offs: what to hold and what to release?
- PAIRQURS sensitivity landscape: what vs who (what to release/hold to/from whom).
- PAIRQURS initial design had issues (too many restrictions for data access).
- Issue: Aggregating data may still give means to identify the person by watching the pattern.

Conclusions reached

- Sensitive data can be data centric or person centric
- It is still an issue to totally anonymize aggregated data. Can massive data ever by anonymised? There is always the possibility to disclose data, for example by spacial hotspots, matching of time to space, use of equipment.

Q/A and discussion points

- Why do you need to keep all those data IDs on your data center?
  - For some datasets, we need to trace back to individuals, and some are not (where skipping the identifier is possible)

- What if you randomize the use of the sensor instrument?
  - This is possible. People became attached to a certain box; this was a problem.
  - Calibration drift, with large enough data this disappears and is above noise level

Session 6: EU data infrastructure for health monitoring, research, and governance: why, who and how?

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Why health information is used in Europe?: broader access the information critical to researchers, avoid drawing conclusion from random variations, best practices to avoid common mistakes, benchmark the effect of local policies. Large amount of health data is in Europe shared between repositories (EUBIROD project)
- There exists a universe of data users with different interests.
- Sharing of health care data sets to evaluate the performance of different healthcare systems in Europe. These are routine data sets not created for research. Policy makers must be made accountable for their policies and for the results EU legislation
- OECD Health at a Glance 2017: is about the quality of care between different countries. Countries are assessed on the basis of available data that can be hardly accessed on a routine basis. 2 years were needed alone for data collection; this was too long.
  - Council recommendations require to go back to the single country.
- Why a EU health data infrastructure needed?: compare results, monitor actions, plan new prevention and care policies. Example: Diabetes. Here we have multidimensional data about Diabetes prevalence. Summary results are used as indicators. Micro-aggregated data is possible for privacy protection.
- Issue: service providers and data centers may exchange aggregate data but not identified data.
- Key data governance mechanisms: Health information system, legal framework, public communication plan, data de-identification process, data security practices. (details in the Organisation for Economic Co-operation and Development − OECD, web-site).
- EU health information infrastructure has never been created (discussed, but never created).
- Content domains to achieve this:
  - General principles of policy goals
  - Subjects of public health monitoring
  - Statistical methods and information technology
- Action plan with different implementation levels:
  - Action level1 (the EU): EU members need to agree on common priorities and indicators.
  - Action level 2 (national): national practices in privacy and data protection shall be freely implemented. Countries have to fund their own infrastructure. A metadata infrastructure is needed.
  - Action level 3: Cross border sharing; here the EU Commission comes into play. A potential ERIC for health data is needed. BrifgeHealth is concerned only with population-based data.
  - Action level 4: Central infrastructure. Metadata should be shared. For each analysis one has to negotiate the risk and what can be done. The data should stay in the local database.
- Blueprint of a general platform for population-based data: leaving the data locally and share metadata

Conclusions reached

- The EU health information infrastructure is still an aim that is needed.
- The implementation levels (above) should be taken into account
- We should present Best Practices; make specific examples with benefit/risk assessment
- Possibility in the GPRD to have waivers for research

Q/A and discussion points

- Why How many local repositories to construct?
  - Depends on how many data providers willing to work together.

- Where the computation happens?
  - On the local data that is not shared
- The "Privacy by Design" approach was positively received.

Session 7: NSD: Tools and strategies for sharing and preserving sensitive data

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- NORD: Norwegian Open Research Data Infrastructure
- RAIRD: Remote Access Infrastructure for Registry Data
- NSD: Norwegian center for research data (national center of expertise on data protection)
- Data protection on NSD:
  - Data Protection Official for Research (DPO) of 137 institutions
  - Training and lecturing
  - Pre/post control of projects
  - Tracking the administration of individual cases
  - Has to deal with GPRD impact
- Key elements: Data protection officers, right to access and to be forgotten, privacy by design (inherit in systems from the onset)
- Norwegian Ministry for Education and Research's "National strategy for access to and sharing of research data" (12/2017) has 3 principles: data must be as open as possible, data should be prepared and handled properly, data should be FAIR
- NSD-NORDi: Norwegian open research data infrastructure: the goal is to develop new e-infrastructure for research data for data curation, preservation, discovery, access, training. Is compliant with GDPR. Researcher has a "MyPage" site at the portal as entry point with a data deposit service.
- RAIRD: Remote access Infrastructure for register data:
  - Pro: Full population registers can be used for research. Reduces barriers for researchers. Contains data with personal ID and sensitive information, like tax and income.
  - Pro: Need-to-know access
  - Con: Approval process is complicated involving registry owners and an inspectorate; it takes about 6-12 months. In addition, the data preparation step is expensive, the researcher's institution has to pay for it
  - Con: researchers get their own copies of data
- Balance between disclosure risk and data utility is important to keep in focus.

Conclusions reached

- This platform (RAIRD) is used only for registered data
- Privacy preserving interaction is much larger, involves also safe output
- Privacy protection by introducing uncertainty; e.g. by using Winsorisation of continuous variables, removal of outliers, suppression of tables with low frequency variables

Q/A and discussion points

- Can be used for unregistered data?
  - No
- Everything will be stored in the RAIRD platform?
- Analysis through metadata, data is enriched with metadata
  - There can be many instances of RAIRD

Session 8: Services for sensitive data in the Nordics

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Service provider point of view
- Sensitive data handling in the Nordic is mainly focused on the use of human data in the medical field.
- Main concern (and challenge) is to protect the privacy of individuals (utilize the use of data in research together with preventing unauthorized access).
- 3 different approaches: ePouta secure cloud, Tryggve secure collaboration and local EGA
- CSC ePouta secure cloud:
  - IaaS model, ISO27001 certified.
  - Combine HPC resources as part of the customers subnet through MPLS using L2VPN
  - Authorization management system REMS; remote desktop
  - Example use case: Sequence initiative Suomi (SISu). One wants to bring sequencing data from US back to Finland. At university of Helsinki, anonymised patient data together with health data
- Nordic sensitive data collaboration (Tryggve):
  - Composed and mainly funded by ELIXIR nodes, and under the administration of NeIC. For Tryggve 2, 6 Mio. Euro available, with sensitive biomedical data. Vision is to give access to a secure infrastructure for sensitive data.
  - Infrastructures: TSD (Norway), Mosler(Sweden), ePouta (Finland), and Computerome (Denmark)
  - Website: Neic.no/tryggve
  - Example use case (supported): Generic risk score and family history as predictors of schizophrenia in Nordic registers (data of 10000 subjects)
  - There are other use cases to be supported
- Sensitive data archiving in the Nordics – Local EGA:
  - Publish metadata on central EGA while the data is archived in the local EGA
  - There can be many geographically distributed local EGA
  - Local EGA activities are connected to control national genomic data
  - Data Access Committee mechanisms are supported; one has to apply for access
  - Tryggve developers are taking the main role in the development of this project to support ELIXIR EU
- Integration of ePouta with TSD; user interface is TSD, backend resources are ePouta. This service will be available for EU, based on user fees and a binding contract.

Conclusions reached

- How to provide sensitive data services is a challenge
- Nordic partners are well connected; involving EUDAT, ELIXIR, EOSC
- There are many service providers who have the experience to resolve those challenges
- Collaboration is useful between Tryggve and EUDAT

Q/A and discussion points

- Can the EGA project handle non-biological data?
  - It can be, but it currently for biological data
- Why are separate EGAs implemented? The local EGAs are for sensitive data (e.g. genomic data), B2SHARE is for open data. Aim is the extension to whole of Europe, enabling data linkage in a secure way. In contrast to the rest of Europe, the Nordic countries are homogenous, regulations are close. TSD/ePouta are prepared for EOSC. The aim is to generate trust.
- Comment: Would be nice to have a publication for local EGA as a EU project.

Session 9: Level of assurance and attribute metadata management (Willem Elbers)

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Introduction of B2ACCES the AAI of EUDAT
- Level of assurance is needed in order to trust that information about an entity belongs to the entity.
- Level of assurance Profiles IGTF: ASPEN, BIRCH, CEDAR, DOGWOOD (different levels of assurance)
- AARC: project that introduced two new IGTF profiles
  - Cappuccino: Access to compute and research data
  - Espresso: suitable for processing personal research data
- Main topic is the Identity Assurance; this means to determine with some level of certainty that an electronic credential is really representing an entity
- AARC allows for differential assurance with two levels of users: Cappuccino (easy access to research data) and Espresso (identity verification). The requirements for user identification are unique tracing to an identity of identifier
- To release identity LoA: follow REFEDS recommendations (unique, assignable, re-assignable)
- Granularity: Use of AAI proxy solutions for authentication. A proxy assigns a LoA based on a policy (question: to what level can one trust the released information?)
- Attribute metadata: Includes information about the origin, and the verification state of an attribute.
- Use of multi-factor authentication can be used to increase the level of security. Single factor authentication can be used in the initial login, then multi-factor can be used when attempting to access sensitive data.
- Step-up authentication: multi-factor authentication, need to specify what factors are needed
- B2ACCESS: Support user authentication based on a personal certificate. Supports eduGAIN. It can aggregate information from external attribute providers or use its own IdP. Decision is based on used metadata of upstream IdP. There is a pilot for life science AAI employing groups of IdP, differential levels of assurance for EGI and GEANT.

Conclusions reached

- Identity assurance is a good start but might not be sufficient
- Federated identity assurance may not be sufficient, more fine grained solution needed
- B2ACCESS is a flexible product for different levels of assurance, but it relays coarse grained identity assurance information
- Life science use cases may require additional feature, combination of different concepts

Q/A and discussion points

- Does the life science pilot include sensitive data as well?
  - This is not yet tested with real life use cases, but this should be possible.
- Comment: Having a EU level LoA is a challenge
- Does B2ACCESS support multi-factor authentication?
  - Not so far.

Session 10: Danish Life science e-Infrastructure

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- It is about national supercomputing for the Life Sciences based on public-private partnerships.
- Computerome: Large user community (1200 – 1500 users), central large infrastructure (8 PB, +70 GB/s BW, 160048 CPUs).
- Computerome infrastructure is design initially to deal with sensitive data.
- Cloud support: currently +1100 users, 9 active projects running 28 private clouds, 8-10 TB data, a large part of it is sensitive data, including data of all main Danish hospitals and research institutes

- Issue: Biological data is not self-explained in many cases, and has to be combined with other data. The combination of patient records with other information has not yet been allowed. Distributed data production is the case, 24 hours globally, but the aim is to put more analytics there where the data is (=cloud bursting) the delivery of secure services across borders
- Secure private cloud:
  - Scaling resources in the cloud
  - provide exclusive control of resources, users can bring their own images
  - Elastic compute: scalability in many dimensions.
  - Use case (IPSYCH) deals with data for psychiatric research (autism, schizophrenia, bipolar disorder, …) combining patient records, genetic data, environmental data from 80000 Danes. Used are encryption (end-to-end), de-identification, 2-fold authentication, secure environment.
  - Clinical use case: Identification of diseases causing genetic variance. Research with cancer markers. The transfer of data goes automatically to cloud for analysis; this requires suitable SLA
- Ambition: build virtual systems that has all the features of physical infrastructures.
- Current project: allowing users to pull data from different sources into the private cloud. Use of a secure data lake to semistructure data using automatic translators. Sensitive data sets are managed in different copies.
- It is possible to pool data on the fly; this is used for precision medicine. The curation is done automatically
- Support for multi-site infrastructure for the private cloud: distributing the compute and centralize the data
- Computerome 2.0 (late 2018)
- Readmap 2017-2022: support for cognitive computing and AI machine learning.

Conclusions reached

- Computerome is a central infrastructure that supports cloud and physical solutions
- Computerome is not designed for performance, but for dealing with sensitive data
- Further development is needed to support better scalability in Computerome 2.0
- Danish national HPC infrastructure consists of 3 points, allowing closed networks for precision medicine

Q/A and discussion points

- How to access computerome?
  - All the information is available on the wiki
- What is the status of the multi-site infrastructure?
  - Still under development.
- When is pooling data on the fly useful?
  - It is kind of virtualisation, not moving the data. Cases where some sites don't have enough support for data management
- How do you have access to sensitive patient data? What sort of consent exists?
  - In Denmark opt-out is the case. The rule is that patients need to state that they don't allow access to their data if they want to. In a use case, data from DK Statistics and clinical data is combined in a fully automated cloud.

Session 11: Genomics research data management in Inserm (Isabelle Perseil)

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Is part of the French Genomic Medicine Plan 2025.
- Inserm is a nutshell for genomics data management. It is the 1st ranked center for genomics research in EU in terms of publications.

- Support infrastructure (BIOSPHERE)s: French ELIXIR node, and French Genomics. Strategy to manage research data including genetic DMP, security strategy (PRIVGEN project), Big Data analytics, deep learning and training.
- Strategy: Use advanced training, genomic research mapping, community developments
- The French research cloud: large scale with wide flexibility (+50 VM types)
- Case study: GWAS analysis (require large amount of resources in case of full genome sequencing).
- Issue: in many use cases, a lot of compute and storage resources are needed together with cloud support but not on public clouds.
- Inserm project: Next generation sequencing use case: processed on the Lille cluster with +1PB of data. Main issue for such use cases is the storage capacity.
- Integrating next-generation sequencing into the healthcare pathway: central national database with different access platforms for different parties that need it for different purposes (e.g. clinic, research, industry)
- >100 GB Ram, WES pipeline for sequences of 1000 individuals. The adapted solution has 1PB of storage (Lille 1), 224 CPUs, 1792 GB Ram for shared computing nodes
- Inserm project: Cross-disciplinary research program: Develop new AI based methods for analysis, large set of tools, new features for data integration and data sharing.

Conclusions reached

- Processing large data is challenging both in terms of security and storage
- New methods based on machine learning are useful for genomics research
- Challenges for sensitive data in France: to strict rules

Q/A and discussion points

- Since it is difficult to get access to health data in France, is it possible to modify the rules?
  - The process of being granted access for a specific party to health data is very long and complex.
- 12 sequencing centers and 2 HPC centers with associations of hospitals; no transfer of patient data is necessary

Session 12: TSD services for sensitive (Maria Francesca Iozzi)

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- Started 2009. It moved from being a project into production service. 2014 in full production. 20 pilots.
- Main use cases are life science use cases.
- Developing TSD we needed to both support usability and security which is hard to achieve (security has always priority), in addition to data sharing.
- 4 main aims: Integrity, sharing, confidentiality, and usability
- Architecture: Remote access solution that is accessible through two-factor authentication where users get access to both compute and storage resources.
- PaaS and SaaS. High usability by using BYOD, one can connect from everywhere using Linux or Win.
- Numbers +500 VMs, 410 projects, +1700 users (Important customer is the Norwegian cancer genomics consortium NCGC for personalised medicine)
- TSD now has the Dragon bio-informatics processor for the genomic workflow that is based on FPGA (30 hrs to 26 mins)
- Current supported use case in cognitive neuroscience: the life-Brain project imaging studies (involve 7 EU partners)
- Data collection tools: Web + API interfaces for patient data with web based questionnaire. One use case for those tools: PreventADALL (preventing atopic dermatitis and allergies), Collecting environmental and life style data from pregnant women; reached over 60000 answers. And MOBA (Norwegian Mother and Child Cohort study)

- Use case for data transfer using smart devices: MinMat for monitoring dietetic routines. Nurses get messages to visit elderly patients at home.
- Future use case: support for pain sensors.
- Currently supported: Video sensitive data.
- There is new regulation from the ministry of research regarding data protection, sharing and integrity stating that research data should be as open as possible, policies should exist to make data available
- eHELSE (12/2016): In Norway there are 300 data registries, biobanks, and repositories in Norway, but with little integration with registries in Scandinavia, 17 months to get access. Goal is better access and better quality of health data usage.
- International collaboration: part of Tryggve, collaboration with EUDAT and EOSC Hub WP 6.6

Conclusions reached

- TSD is the national infrastructure for storing and processing sensitive data in Norway.
- Aim is to facilitate the access to health data
- Better technical solutions are needed and the development is ongoing. Collaboration in EOSC Hub, B2SHARE and TSD (publish metadata, harvested by B2FIND and share anonymised data with B2SHARE), Sigma 2 and universities.
- In Norway, there are so many data registries, getting access is a very slow and complex process.

Q/A and discussion points

- How about getting access to anonymized datasets through B2Share?
  - May be possible when we trust the anonymization process.
- Should a data warehouse structure be built? Architecture is not yet defined for this.

Session 13: Can sensitive data be FAIR? (Luiz Bonino)

Rapporteur name: Abdulrahman Azab (UiO), Wolfgang Kuchinke (HHU)

Presentation highlights

- FAIR Data: Finable, Accessible, Interoperable, and Reusable
- The GO FAIR Initiative aims to make research data findable, accessible, interoperable and reusable across national and disciplinary boundaries
- Accessible: means access under well-defined conditions. Metadata is retrievable and accessible when the data is no longer available.
- The Internet of FAIR data and services IFDS: Data, compute, and tools
- Two initiatives:
  - Personal health train (constraints are heterogeneous data, security and privacy), data must be locally accessible; the train leaves the data station only with results
  - Delegated access (data in hospital system), access is only for a specific purpose; third party staging and execution: because hospitals cannot run extensive algorithms locally, data is temporarily stages for execution
- First initiatives of the IFDS: Personal health train, and Farm data train
- Data access and execution: different scenarios based on who has the rights to access the data and where the execution can take place
- Meta-data approach: metadata use a formal accessible shared and applicable language for knowledge representation. This is necessary because a single central storage is not being to happen, we have to deal with a distributed infrastructure that works together
- Data set linked to metadata specifying the distribution license, type of data, where to access and download
- Meta-data layers (top-down): Data Repo (FDP), dataset catalogue, dataset, distribution, data record (sematic knowledge about the dataset).
- Metadata model extension: add more annotations on the same entity (extend the ontology)

- Bioschemas.org, organisation of Metadata using bioschemas elements for landing page, then metadata is harvested by Google
- FAIR-IFIER. The FAIR-IFIER can speed-up the process of data FAIRification of data sets; ORKA is used for annotation. The Data FAIRport initiative addresses service interoperability using a "locker" for sensitive data (Person health data locker)
- FAIR data ecosystem: integrated ecosystem for supporting FAIR access to data. Data access through personal data lockers between data users and data owners.

Conclusions reached

- Sharing and accessibility of metadata is more important than it is for data.
- The metadata layers architecture may not be sufficient, and needs to be extended with other vocabularies
- Important principle: data should not move!
- Several FAIR tools exists that ensure the fairness of data and services

Q/A and discussion points

- If the data owner gives access, how to lock what is being done with the data?
  - There are mechanisms to assure that only authorized actions are made on the data when access is granted.
- ORKA for annotation. Annotation is used when data is used for different purposes, annotation of sensitivity is possible. Can the annotated data be corrected e.g. modifying some metadata fields
  - Yes
- "Passengers" are missing in the train, consent should be included

Session 14: General discussion

Rapporteur name: Abdulrahman Azab (UiO)

Main points

- National strategies to make reuse of sensitive data
- Central vs distributed sensitive data storage approaches
- Strategies to make research jointly on both sensitive and open data, across disciplines and across borders
- Sensitive data and FAIR principles

Conclusions reached

- Taking the burden of composing EU strategies for sensitive data reuse is cumbersome
- There is a need for making the data available, otherwise it will lose its value by time
- If the required technology and competence needed is available somewhere, it should be easy to share it among other interested sites. Joining forces is much more useful than reinventing the wheel. Having a Nordic cloud infrastructure for sensitive data can be a good example.
- The discussion was too centered on the success of Nordic countries; we need a European dimension. This means political discussions and different views on the exchange of data. But one can learn from the Nordic countries, e.g. Best practices, overlap of care and research, quality of care. We must show the advantages of EU collaboration, use of data effectively and more publicly, generation of trust
- There are obstacles in the way of sharing data across countries mainly due to the very different security rules and regulations.
- Important is that different domains join; this done best by joining the infrastructures. The technology is there to analyse large international cohorts
- The trust mechanism between service provides and data owners should include guaranteeing that the data owner has full control on the data management and access workflow.
- Good solution is the "Pooling data on the fly". Data is stored in national repositories/clouds and pooled upon approval for specific analysis only

- Data will not leave the hospital, the algorithm is sent. Citizens should be able to send their train. Evaluation necessary, how sensitive is the data, is data OK for a specific purpose, then decide if data is anonymised, or treated otherwise.
- It is very important to make publications in English other than e.g. Scandinavian languages. Because for policy makers, the technical infrastructure is difficult to understand
- There is not going to be one solution that resolves all issues. There can be different platforms to resolve different issues regarding data security and integrity. The question is, where will metadata be stored in Europe? A single repository is unrealistic. But one should concentrate on free access to metadata of sensitive data
- Due to social and cultural aspects, data protection in France is very strict.
- Encryption of data is not enough. One method is fragmentation, to access only the part that is necessary for analysis
- Sensitive data cloud: Dream that should be possible to achieve.

B.2. Semantic services in EOSC

Rapporteur name: Barbara Magagna (Umweltbundesamt)

Presentation highlights

Yann le Franc − Introduction: Three main aims of the workshop: overview on EUDAT semantic services, overview on semantic activities in various research communities and discussion about needs for semantic services in EOSC-

Aymeric Rodriguez − B2NOTE: B2Note, integrated in B2Share has been released Jan 2018. It is based on the Semantic Web Annotation data model, allows 3 types of annotation: semantic tag, free-text keyword, comment. Semantic tags are keywords derived from indexed ontologies or controlled vocabularies.

Julia Karimova - Using B2NOTE − The U.Porto pilot: is a use case of B2Note at the end of the RDM Workflow. Data is organized and described in Dendro Platform; published on B2SHARE and annotated using the B2NOTE service.

Yann Le Franc − Semantic Lookup Service: The B2NOTE Service builds on a semantic index (SolR) to support term/concept discovery for semantic annotations. But semantic resources (ontologies, thesauri..) are distributed over the web with limited visibility, different formats and metadata descriptions, different repository data models and different APIs. A Semantic Lookup Service was developed to aggregate semantic resources from three repositories (BioPortal, EBI-OLS and AgroPortal).

Yann Le Franc − Modeling Data Life Cycles with PROV: EUDAT services can be linked to different stage of the UK data archive Data Life Cycle. The DLC is considered as Data Management Workflows and the PROV model is extended to build a declarative representation of data management workflows. A service to create DLC descriptions using EUDAT services and store the associated provenance template is under development.

Clement Jonquet – FAIR data requires FAIR ontologies, how do we do?: Ontology repositories help to make ontologies FAIR. Users can publish, download, browse, search, comment, align ontologies and use them for annotations both online and via a web services API. Overview on repositories and collaborative projects reusing NCBO technology (e.g. AgroPortal). A new standard for ontology metadata with the VSSIG is being developed.

Nicola Fiore – Toward LifeWatch ERIC FAIR: For sharing information harmonisation and standardisation of the heterogeneous data is needed. LifeWatch Thesauri are used to insert keywords in the LW Metadata and for data annotation, LifeWatch Ontology (based on OBOE) comes with a SPARQL endpoint. In collaboration with LTER a biodiversity community portal as a central registry for semantic resources is being developed.

Paul Martin − ENVRIplus − OIL-E and ENVRI Knowledge Base: ENVRI RM (Reference Model) is used to help standardise RI descriptions and identify compatible interfaces and points of overlap. OIL-E describes RIs and

their components to produce a semantic landscape of RIs. To reuse technologies, the ENVRI Knowledge Base will provide a layer for use by services that associates technology profiles with different workflows.

Barbara Magagna − LTER-Europe – EnvThes: LTER is a network of institutes and sites with heterogeneous data management and IT capabilities organized with DEIMS as metadata catalogue, EnvThes and Data Integration Portal. EnvThes is used as the harmonized vocabulary for observation and measurement and to select keywords for annotation of datasets, sites and data products in DEIMS.

Nuno Freire – Europeana: is the platform for Europe's Digital Cultural Heritage Community aiming at the aggregation of metadata from libraries and museums within Europe. It developed Europeana Data Model (EDM) as semantic layer on top of the cultural heritage objects and a Europeana linked data strategy to apply automatic enrichment by linking source data to reference data building an Entity Collection.

Christian Pichot – ANAEE: is an EU infrastructure for analysis and experimentation on ecosystems including data management infrastructure. Two level architecture: discovery level using catalogue compliant with INSPIRE and ISO 19115 standard; access portal to access the resources. Work on interoperability using semantics (ANAEETHES and Ontology) and Semantic Linking service.

Doron Goldfarb – Semantic Look Up Service: A centralized index of concept extracts from multiple resources in multiple repositories enables central access to multi-disciplinary concepts/terminology for semantic services, identification/cross-repo search of unique and overlapping resources in different repositories supporting development of own semantic resources. Open challenge are common metadata standards / APIs.

Alexandra Kokkinaki – SeaDataCloud: Host 239 vocabularies at NERC vocabulary server. NVS is a SKOS-vocabulary including internal and external mappings. It offers search in vocabularies, vocabulary builder, submission of new terms, has to pass vocabulary management group, RosettaStone translation European marine datasets served by SeaDataNet are annotated with terms from controlled vocabularies.

Vangelis Karkaletsis – DARE Delivering Agile Research Excellence on European e-Infrastructure: started exploiting and extending SemaGrow and Big Data Europe (BDE) technologies for annotation and semantic enrichment of datasets, methods, and tools, including provenance information, discovering mappings between schemas and vocabularies, building a Semantic Registry for DARE assets (data, methods, tools, workflows)

Maggie Hellström – ICOS: The major mission of ICOS is to collect and make available high-quality observational data from its state-of-the-art measurement stations operated with a long-term perspective. It operates three networks: atmosphere, ecosystem, ocean. A semantic web & ontology-based approach for metadata storage with own ontology is being applied, persistent identifiers for optimizing interoperability for all data assets.

Menzo Windhouwer – CLARIN: Research Infrastructure for the humanities and social sciences, provides easy access for scholars to digital language and discovery and annotation tools. CMDI architecture; Concept Registry as SKOS-based registry containing metadata concepts and Vocabulary Registry (CLAVAS) containing vocabularies for metadata value ranges.

Pedro Principe – OpenAire: To reach interoperability OpenAire collects content from different providers, literature repositories (publications), content data, datasets, research outputs, from registries (projects), collecting software from Zenodo. Aim is to clean metadata, de-duplicate (one source many links), to provide guidelines for literature repositories, for data repositories and for Open Science Content Providers.

Mark Greenslade – IS-ENES: Infrastructure for Earth System Modelling with very large scale simulations and long term archives. Scientists need metadata describing relationships between model inter-comparison projects, experiments & requested variables. Centralized citation management links datasets to citations via PIs. Provides documentation of dataset errata, ES-DOC ontology and vocabularies used, but not OWL/SKOS.

Conclusions reached

The main conclusions reached, are described as follows:

- There is a clear need for semantic services with multi-disciplinary coverage.
- Semantic resources are first class citizen in the implementation of FAIR data management system.
- Needs of communities like Biodiversity and Ecology for aggregating data from multiple disciplines.
- Needs for Knowledge Engineers/Ontologists to access, aggregate and analyze semantic resources.
- Needs for SKOS vocabularies lookup service.
- Needs for provenance of mappings and versioning of concepts.
- A central hub for ontologist and knowledge engineers to access multi-disciplinary semantic resources and analytics.
- A multi-disciplinary curated semantic index that can be used by semantic tools (annotators, text miner, etc.).
- A set of recommendations and standards to ease the interoperability of semantic resources.
- Potential business applications that would guarantee the sustainability of the resources.
- B2services are used by many infrastructures.
- B2NOTE and Semantic Look Up service are the EUDAT Semantic Services of interest for most RIs.
- To enhance user experience when searching the ICOS data catalogue, it would be useful to have the possibility to connect to an external “synonym service”, of the type proposed by e.g. AnaEE.
- Annotation exchange service would be a needed extension of B2NOTE.
- Annotation of time series datasets to indicate regions of unusual activity, indicating an interesting event or a possible instrument failure would be needed, in general annotate content of files would be a necessary feature.
- Integrate automated text document annotation functionality in B2NOTE as part of an expert annotation workflow.
- A tool embedded in an ontology and thesaurus editor to comply with a metadata standard for semantic resources would be good to have.
- People should be able to vote or comment on a annotation stored in the B2NOTE database by adding some social media functionality.
- Visualisation of the annotation must be improved, better integration of B2SHARE.
- Need for a marketplace for ontologies.
- Semantic lookup service should be integrated with the Semagrow.
- Increase metadata quality in repositories; towards a new standard for ontology metadata with the VSSIG.
- Standardization of Ontology Repositories.
- Collaboration between LifeWatch/LTER-Europe and AgroPortal to build a Biodiversity Community Portal.

Q/A and discussion points

The discussion focused on the following points:

- How to deal with semantic drift? Refer to a concept/term in a specific (archived) version of an ontology/vocabulary?
- How to deal with emerging semantics? How to suggest new concepts/terms to existing ontologies/vocabularies?
- How long are you storing these annotations in B2Note? In B2Note central database is stored what you have in the annotation, the semantic annotation and the URL of the data.

B.3. Array databases for research communities

Rapporteur name: Christian Pagé (CERFACS)

Presentation highlights

On RASDAMAN: Array Database: SQL + n-D arrays; Mature, operational; Service Model: OGC standardized. WCS: Core + Extensions. OGC Web Coverage Processing Service (WCPS).

On Earth Server: Provide agile analytics on PB of data. Federation of services. EO, Marine, Climate, Planetary. H2020 project. eodataservice.org; @eodatacube. Dice -> Stack -> Use. Regional Data Cubes with different data sets and different domains, different formats. The conclusion is: One user interface per user category. One interface is not enough, users need customised views (web portals).

On the EUDAT Generic Execution Framework: GEF frontend is used to build docker images with scientific code/tools previously containerised. GEF helps to move computation nearer the data; Docker-based solution.

On Ophidia: Support end-to-end large experiments; Architecture Datacube OPHIDIA -- framework exploiting parallelism. Ophidia and RDA project BARRACUDA exploits the PID recommendation by RDA.

On SkyArrays: experiment with Rasdaman and lessons learned. On the positive aspects: ArrayDB helps make life easier; Standardized access to data; Less error prone for subsetting/assembling; Flexible access to relevant data partition. On the negative aspects: Input data have to be perfect, not always in real life, unfortunately; Still careful in the DB query when many processes ask (distributed installation might solve this).

On the EUDAT CDI: The EUDAT CDI has a partnership agreement to sustain its services in the long term (10 year); B2Host to support GEF; EUDAT is working on a workspace area for not registered data; The EuroArgo use case might fit an arrayDB, they are currently based on Spark.

Conclusions reached

The main conclusions reached were the following:

- Big strength of ArrayDB is scalability for large data volumes.
- Data pre-processing step is a key aspect, in terms of processing time and data quality aspects.
- ArrayDB WG should continue within RDA.

Summarising the lessons learned, there some positive aspects (ArrayDB helps make your life easier; Standardized access to data; Less error prone for subsetting/assembling; Flexible access to relevant data partition) and some negative aspects (Input data have to be perfect, not always in real life, unfortunately; Still careful in the DB query when many processes ask (distributed installation might solve this).

Q/A and discussion points

The discussion focused on the following points:

- ArrayDB application scope e.g. geospatial, additional?
- Desired features e.g. data access, data processing, standard APIs, big data analytics
- From ArrayDB to DataCubes
- Hosted or dedicated service, onsite support?
- Target communities? Research Infrastructures? Researchers? Commercial?
- Accounting model? Pay per use/space
- Added-value services offered by e-infrastructure providers e.g. data preparation
- Others?

Detailed Notes

EUDAT is bringing together tech experts, research infra, to assess new technologies that could be used as a base to future services. There are parallel activities in RDA and EOSC. The goals are: a meeting point, tech review, demos, strategies and recommendations.

The Presentations by technology experts included the following points:

Vlad Merticariu − RASDAMAN: Rasdaman: Array Database: SQL + n-D arrays; Mature, operational.; OGC & INSPIRE WCS implementation; Array Data Model. Query Language. New: Polygon Clipping; Architecture sketch. Support different types of grids; Data Model: CIS 1.1 Coverage Definition. Example of a Simple Coverage in GML; Encoding Coverages: Single File encoding or Multipart container; Service Model: OGC standardized. WCS: Core + Extensions. OGC Web Coverage Processing Service (WCPS); Data Import: Data Files + Ingredient + WCST Import. Ingredient is in json. Result is a coverage = datacube. Operations: WCS; WCPS. Connect Applications. For example, NASA World Wind. Questions: No authentication/authorization yet. Authentication managed at DB level; Services on top: is interpolation available? Only Nearest neighbour; NetCDF is supported in the output; Supports time within layers? WMS support is limited but for WCS is ok.

Simone Mantovani – Earth Server: H2020 project. eodataservice.org; Provide agile analytics on PB of data. Federation of services. EO, Marine, Climate, Planetary; KO May 2015. 3 years. 6 EU 1 US 1 AUS; AUS is willing to evaluate RASDAMAN; Cross-domain platform. Discovery Exploration, Visu + processing; Effective data subsetting application scenario: main outcome; Enabling technology targeting a variety of data; Interoperability Layer. @eodatacube. Dice -> Stack -> Use. Regional Data Cubes with different data sets and different domains, different formats. User Interface: web portal with several data sets from several scientific domains; User Interface: Jupyter. Conclusion: One user interface per user category. Exploiting data across infra. One interface is not enough, users need customised views (web portals). Questions: Data Ingestion, Sentinel2: 100000 prods per day and about 3 TB per day. No problem in ingestion and data synchronizing. Is it sensitive to tiling? How different data formats are ingested?

Asela Rajapakse – Generic Execution Framework: Presentation of EUDAT and its Service Suite. GEF is not as mature. GEF helps to move computation nearer the data. Docker-based solution. GEF Architecture. GEF frontend used to build docker images with scientific code/tools previously containerised. DKRZ will take over the GEF and push towards TRL7. Several Use Cases are available.

Christian Page − OPHIDIA: Collaboration CMCC; Support end-to-end large experiments; Architecture Datacube OPHIDIA − framework exploiting parallelism; Big data analysis combining data and metadata including provenance, native support for formats like NetCDF; Extensible to support new primitives, multiple interfaces and programmatic access; Many operators available MPI and openMP based; PyOphidia: client class, cube class; Use cases; Large documentation available with practical examples; Ophidia and RDA project BARRACUDA exploits the PID recommendation by RDA; Ophidia and EOSC − ECAS. Questions: Ophidia training? Not planned but certainly needed; Ophidia supporting only NetCDF? Not sure but it should support more formats; What is offering more respect to other systems? Parallel processing and remote interaction. Reuse existing operators.

Andrea Pagani − SkyArrays : experiment with Rasdaman: Sky view factor applications: heath stress, fog formation, road temperature; Old workflow vs New workflow; Initial dataset problematic solved acquired a new cleaner dataset from a different provider; Computation, the new way: Using a rasdaman server. Divide work among slaves. Comparison results between the old and new workflow approach.

Lessons learned

In this use case the following lessons have become clear as positive and negative aspects.

Positive aspects:

- ArrayDB helps make your life easier
- Standardized access to data

- Less error prone for subsetting/assembling
- Flexible access to relevant data partition

Negative aspects:

- Input data have to be perfect, not always in real life, unfortunately
- Still careful in the DB query when many processes ask (distributed installation might solve this)

NB: of course a necessary condition is a positive interaction with engineer to share knowledge for a working solution.

Questions: The tuning, optimisation and interaction with the arrayDB engineers required suggest that this is a very specific solution and not general purpose. Did you have prior experience with traditional DB? Tiling schemes might differ depending on use case, some data duplication might exist. Not perfect Data issue exists also without arrayDB.

Christian Page demos: Video on OphidiaBigData github; Data download is usually the most expensive step but there is a caching system that reuses data when available; Results can be accessed via opendap and http and can be in png or nc; The cluster can be monitored during the computation. Questions: Graphical interface is used just for monitoring not to define the wf; Ophidia don't need to run on every cluster node; It exist only 1 operational installation. Technical discussion: Distributed data need to move to the same place to be analysed; Federations of data cubes with different technologies; Landsat mirror data in data infrastructure.

Technical Discussion main questions: Distributed calculations with current systems. Is it possible with Ophidia? Example of ESGF computing nodes accessing the data nodes locally to perform data reduction and subsetting before handing back the data to the user. Rasdaman errors, there should be more doc about configuration parameters depending on the number of users and their habit.

EUDAT CDI − Mark van de Sanden, SURFsara: History of the EUDAT CDI. EUDAT CDI has a partnership agreement to sustain its services in the long term (10 year). CDI Architecture and Service Diagram with services in production. GEF is not there yet because still in development and in prototyping. Evolution of the CDI. B2SAFE is using GridFTP and PIDs are automatically generated using B2HANDLE. Extension has been done to provide the access through the HTTP API instead of relying on GridFTP. The HTTP API will also be able to eventually launch processing using the future GEF Service. B2SAFE is also extended to B2SHARE to provide metadata information, and access through B2ACCESS. Not fully integrated yet (B2ACCESS). Extension to B2FIND so that all data in the CDI is searchable. EUDAT is working on a workspace area for not registered data. B2Host to support GEF. EuroArgo use case might fit an arrayDB, they are currently based on Spark. SeaDataCloud − data hubs.

Beyond EUDAT2020

EUDAT CDI PA 09-2016+ ; EOSCpilot (no dev) end 01-2019; EOSC-hub 01/2018-01/2021 then EOSC? SeaDataCloud also mentioned. Future H2020 calls? 01-2019+. For ArrayDB future dev work, what would H2020 calls be an appropriate place? EOSC-hub: no new dev of services possible. DIscussions within the EUDAT CDI on how to pursue further developments and implementing new requirements and developing new services. In the EUDAT Service Diagram, where an ArrayDB data would be located?

Recommendations

- ArrayDB application scope e.g. geospatial data, others?
- Desired Features e.g. data access, data processing, standard API
- Data preparation: ~80% of the time in the preparation phase. How is the datacube helping in the data process. Processing is already done beforehand before having it in the datacubes (rasdaman).
- From ArrayDB to DataCubes
- Hosted vs dedicated service, onsite support?
- Target communities? Research Infrastructures? Researchers? Commercial?

- Accounting model? Pay per user / storage space / …
- Possible added-value services offered by e-infrastructure providers e.g. data preparation
- Data preparation: interesting wrt the GEF. Prepare the data to ingest somewhere. E.g. metadata extraction.
- Others?
- Need to go to a 24/7 service, more generic, not project or data set specific.

Big strength of ArrayDB is scalability for large data volumes. What is missing is a WCS attached to B2SHARE.

B.4. Research data management: interoperability, collaboration, and the research library role

Rapporteur name: Vasso Kalaitzi, LIBER Europe

Presentation highlights

Astrid Verheusen talked about LIBER, as Europe's largest research library network, focusing on scholarly communication, digital skills and research infrastructure, helping libraries support world-class research. The vision for 2022 includes the following aspects: Open Access is the main form of publishing; FAIR research data; digital skills underpin a more open and transparent research life cycle; RI is participatory, tailored and scaled to the needs of diverse disciplines; tomorrow's cultural heritage is built on today's digital information. Advocating/raising awareness/community engagement regarding Open Science; contribution to RDM; training and support; bringing in the "human component" besides the infrastructural one. Research libraries are uniquely placed at the intersection of knowledge, research, tools and services and education.

Daan Broeder talked about the EUDAT initiative, starting already in 2011. The EUDAT CDI members are: generic, integrated service providers; generic, interoperable service providers; thematic, integrated service providers; thematic, interoperable service providers, while there is collaboration between Service providers and Research communities. Within the framework of consultancy and training, there is a network of best practice, co-design and technology exchange. The EUDAT Services are covering the whole DLC, reuse existing technologies, existing practices & recommendations, are a standard service, while there is also a need for legal entity and minimal maintenance through member fees. The EUDAT community partners include thematic research centres, organisations representing community RIs, while at the same time EUDAT is participating in community projects. The vision for the EUDAT CDI collaborations is to work towards cooperation and inclusiveness, building on results from cross-community discussion processes, making use of existing knowledge infrastructures. EUDAT should be part of librarians' toolkits.

Raphael Ritz spoke in his first presentation about the RDA vision: researchers & innovators openly share data across technologies, disciplines and countries to address the grand challenges of society. RDA builds the social and technical bridges that enable open sharing of data. RDA is member based. Its principles are: openness, consensus, balance, harmonization, community-driven, non-profit & technology neutral. RDA provides recommendations that make data work: "Create-adopt-use". About libraries for research data, RDA provides an overview of practical, free, online resources and tools that users can immediately take advantage of to incorporate research data management into the practice of librarianship. About EOSC & FAIR, it was stated that RDA is FAIR. While FAIR is about principles, RDA is about realization.

Giannis Tsakonas spoke about a common vision of a FAIR condition for Research Data. European libraries are implementing mostly "soft" services, such as development of DMPs, citation practices, etc. Technical services not fully established and research data management is not featured in all institutions. Determining what to discard or dispose of items can be harder decisions than those at the time of acquisition [Borgman]. There are "Collection development" issues. Technical heterogeneity and metadata quality need to be addressed. Other issues addressed by this speech were: Libraries as organizational interfaces; Front ends to communicate with various stakeholders; combat defragmentation; Collection, process and alignment of user requirements; To assist short/long term data preservation.

Leon du Toit spoke about collaboration principles and needs, barriers for adoption and the removal of the barriers: libraries can alleviate some barriers to e-Infrastructure service adoption via roles that are linked to their outreach function, being point of contact and their role in decision support. There are opportunities for libraries in research data management, as well as certain complications: differences in maturity and capacity, training needs, budgeting in projects. The goals include: common understanding of research data management among libraries; managing cooperation; explicitly in budget processes. Where do libraries fit in: outreach; consulting; training; brokering; coordinating; technical services such as DMPs. Researchers rely on collaboration between libraries and research and e-Infrastructures.

Raphael Ritz also spoke about ICT technical specifications and definitions and the identification of ICT specifications procedure. The European Multi Stakeholder Platform is an expert advisory group on ICT standardization. There is RDA compliance with requirements for ICT technical specifications. The specifications approved include: basic vocabulary of foundational terminology and query tool; the data type model and registry; the machine actionable policy templates; the persistent identifier type registry. Next five under evaluation.

Conclusions reached

The main conclusion was about the need to find the common ground between visions and further collaboration, in order to meet researchers' and librarians' expectations. Clear roles are needed. There are some budgeting issues that need to be addressed. To take into consideration the thoughts that emerged from this co-located event and move forward. Another important aspect of the role of libraries is that of advocacy, raising awareness, training and decision-making.

Q/A and discussion points

The main points of discussion are described as follows:

- The need of repurposing in terms of collaboration, discussion amongst libraries on how to deliver services and view of the researchers on the research libraries: There is also the need to join forces with other institutions, for example university computing units, Max Planck digital library, RDA and initiatives like Go Fair.
- Managing research objects and the expertise that libraries have: The role of libraries in vocabulary management, data typing, citations, how to involve libraries in the data process and in developing recommendations. Some of the results emerge from working groups, where library members participate or lead; support for projects; demonstrators; proof of concepts; to show the added value. To find a way to define open and interoperable types of data, in order to preserve research objects, protocols and workflows.
- Efficiency should not become a taboo. We are not always talking about cost-efficiency.

Detailed Notes

During the first interactive part, the following questions were discussed:

- Based on the presentations, is there anything you feel is missing from the role libraries/service provides/data infrastructure communities/research communities can play within EOSC? What is the role your type of organisation/community can play?
- Is overlap in scope, activities, and roles a waste of resources or are there also benefits? What are the pros and cons of multiple organization conducting similar efforts?
- How can we organise efficient collaborations between such organisations (personal unions, organisational interactions, etc.)? Could your organisation/community actively participate in this effort?
- What could be the long-term vision of a strong, synergetic collaboration within and for EOSC? What kind of approach would you like to see being followed?

The main discussion points by the four participant groups were the following:

- Where do we take the expertise to do whatever we want to do? Where do we get these people? Representation of a European effort. How to you make sure that these people are collaborating? How do you know what are the roles and the responsibility of whom? What is the interaction of the libraries with research institutions, for example? The alternative is to have an organization doing it all, instead of multiple organisations. Different people responsible working with the communities they are involved the most. We need to let go of the word "efficiency". To choose these people wisely – how to build a synergetic collaboration: rely on trust. Similar amongst different stakeholders – building trust is what we need to do.
- First thing mentioned as missing is support − the researchers need to have projects and funding. How to engage researchers to RDM. Not every country/society has data policies / support / incentives / rewards. These are important but not always covered. Sometimes it is not only about the carrot, but funders need to also use the whip. The discipline-specific guidance is missing. What does it mean in my field? It is different from one discipline to another. Overlap can be a problem when different organisations talk about the same thing, but it can be productive in terms of advocacy. Not only is service provisioning but a lot of information is missing, especially at national level. However, we do know a lot more about each other but is more about talking than doing. European initiatives need to work together. Efficiency: sometimes it may be at international level – directives might help but who is doing that at European level? Europe might be regulating our lives too much – should we care for more directives? Librarians know their field, but it's not very common – maybe more collaborations are needed.
- The roles libraries are taking is not very focused or well defined. Libraries evolving into research managers, but this interaction doesn't happen often. Roles that can be played: advocating. Pro and a con: you have competing resources and you hope the best one will get to the top. That is the bad aspect. These communities should work towards interoperability. Efficiency: RDA is a good example about this. Synergetic vision: EOSC crystalises a good vision but it's a difficult thing to adopt. Again, it goes to interoperability and competition.
- More focused about training and exchanging knowledge – many different organisations sometimes cause confusion. Training is needed in different levels. Train how to train. How far do we see, for example commercial service providers? What different organisations can bring in? Pros and cons: more cons: difficult to find what are the different approaches and the benefits we can get out of them? How can we reach them? How is all this information transformed to what is needed? How is it transferred to the local organization? To librarians and IT people? Gaps and how do we organize it? For example, we have EOSC, which is a European initiative, but how do you bring this at a national level? How do you want to organise this? We don't have to look at the efficiency first? Different levels of interaction need to be organized, especially at a local level? Takes a lot of effort to use these platforms at a local level. They have to seek more collaboration. Website or registry to find what is really going on? Getting more information on trends.

During the second interactive part, the following questions were discussed:

- What are the services needed by your organisation/community in terms of RDM? What are the services already used by your organisation? What are the further needs/challenges LIBER/EUDAT/RDA could address though this collaboration?
- Are there training needs in your organisation? Are you already tackling them? What would be the best training scheme that your organisation/community could follow? What kind of professional categories could attend such training sessions? What are the skills needed and what the available learning materials/tools and standards into integrating them?
- Do you feel there is a gap between organisations of the same type, in terms of integrating the skills needed and the implementation of Data Management Plans? What are the main problems causing these gaps, if any? How would we bridge the divide?

- What could be the long-term vision of a "package deal" for services/training/advocacy? What is the support your organisation/community needs, that is not available at this moment? What are the communities that should be involved further in this discussion?

The main discussion points by the four participant groups were the following:

- The main problem identified is that without funding data management gets second. Long-term support for data is needed. Training can be community depended, which is part of the issues. Different communities have different training needs. Skills in implementing DMPs: in services where people bring their data for processing and experiments they need to introduce the data management issues in the applications – need to know what happens when the funding ends. We need to identify workflows within the communities; what are the steps we need to follow to make sure data management activities are in place – data policies needed.
- Finish experience: they are organizing services at national level to collect and preserve data. Advice on opportunities, standards, collection, organization, storage, to make them available and FAIR, it's emerging but not everyone can afford it. Opportunity to have a central training in place. Local support from communities needed – another layer of support – awareness needed on how to organize data, where to put them, how to describe them – cooperation is needed. National centre for training calls local people to support in Finland – librarians are relevant. Problem of scalability in multiple levels.
- Services needed: discovery tools are needed – Need for training: lack of funds and staff that is educated to help the researchers – need to train the trainer for library staff to further help researchers – state of mind of the researchers – they need an overview of services and guidance; the state of mind goes before that – difference between countries.
- Assessment of price issues – expertise – lessons learnt – how to move these things to a national and European level – to have some kind of routine for that. Group mostly of service providers – Training needs: train the trainer, making the material available; institutional level: they don't have the means or the (example: instead of long documentation: 6 mins of YouTube videos) DPM skills: maturity of the institution plays a key role. Clear incentives for doing this exercise. Are there the tools available? Package deal: difficult to understand – different things are needed at national and European levels.

B.5. SeaDataCloud workshop

Rapporteur name: Dick M.A. Schaap – MARIS

Presentation highlights

- Presentations gave a nice overview of a number of activities that are underway for improving and expanding the services of the SeaDataNet infrastructure for marine and ocean data management;
- Presenters had undertaken good efforts to explain the background, the marine context and the reasons for the planned upgrading of services as well as the present status of developments
- Presenters also highlighted where the cooperation and synergy with EUDAT and its services is planned and under development

Conclusions reached

- Not applicable as the Workshop was more dedicated to making participants aware and informed about the ongoing activities in the SeaDataCloud project which is a joint activity between the marine data management community around SeaDataNet – EMODnet and EUDAT

Q/A and discussion points

- Most discussion was around INSPIRE. As part of SeaDataCloud a mapping has been made from the SeaDatanet standard formats (CDI metadata and ODV data) towards the INSPIRE data models. It is now planned to use this conceptual mapping for working out a number of use cases, in particular for nutrients and contaminants which are highly relevant for the implementation of the EU Marine Strategy Framework Directive (MSFD). The use cases will provide the input for setting up Transformation Services from SeaDataNet towards INSPIRE compliance. This will be very relevant for

Member States that are 'struggling' with INSPIRE compliance and that already are or might become contributors to the SeaDataNet infrastructure.

- INSPIRE is an EU Directive for implementing a European Spatial Data Infrastructure and is led by EU DG Environment as it is considered as an instrument for supporting environmental management.
- However, INSPIRE is never mentioned in the development of the EOSC which is more looking at FAIR, while scientific data should also be fit for supporting environmental management and its policies.
- There weres also some questions in how far SeaDatacloud is looking into 1) provenance and 2) annotation of data. These are no subjects in the SeaDataCloud project. However, there is interest from the FAIR and EOSC perspective and there are experiences in the marine domain that can contribute in a later stage.

B.6. ENVRI workshop

Rapporteur name: Maggie Hellström, ICOS

Overall, 13 people took part in the event, listening to overview presentations on the H2020 project ENVRIplus and its members & end user communities, and activities and outputs (specifically the portfolio of research data management services) of the work packages associated with the projects "Data for Science" theme. The original objective of the workshop was to provide a forum for different communities in ENVRIplus and EUDAT to discuss updated requirements for effective data management services and research support systems, to share development results and best practices, and to propose an agenda to move towards EOSC. In the end, due to the last-minute cancellation of several invited speakers, the agenda was shortened to focus on the development of services within ENVRIplus and on the on-going activities focusing on mapping the landscape of potential end users of ENVRI data products and other outputs.

Background

The Horizon2020 cluster project ENVRIplus is one example of a collaboration between domain-specific Research Infrastructures (RIs) aimed at developing reusable solutions to address common challenges in managing research data. This approach is proving to be very successful, but cannot fully solve interoperability and sustainability issues.

As the European Open Science Cloud (EOSC) is now truly taking off, it is of great interest to gather relevant experts from the ENVRIplus community, e-Infrastructures and other technology service providers to discuss of how the services and other products being developed, for example in ENVRIplus will fit in this new EOSC landscape.

In the RI projects, IT service development and maintenance are known to be very important; however, their budgets are often very limited. In the meantime, existing software and tools developed by different RI communities are not yet fully exploited and reused. On the one hand, there are increased constraints on RIs and other "larger" (European) research projects to render whatever services they are developing themselves available to a larger audience − whether or not these research-related services are stand-alone or rather layered on top of basic services provided by e-Infrastructures associated with the EOSC. On the other hand, duplication of efforts on similar topics can still be observed in RIs due to limited readiness, visibility or interoperability of existing software.

A number of challenges have been identified, such as how to catalogue development results from different RI communities, assess their quality, identify gaps and promote standards for interoperability? Who are the key actors that need to be involved in order to ensure that these RI-produced services are consistently and effectively evaluated and their quality assured? How can sustainability, in terms of both operations and competence, be guaranteed, at least in a medium-long timeframe?

Contributions & presentations

- Paul Martin, University of Amsterdam: Introduction to ENVRIplus Theme 2
- Maggie Hellström, Lund University: The ENVRI landscape
- Paul Martin, University of Amsterdam: The ENVRIplus Theme 2 Service Portfolio

The presentations were followed by a general discussion on the topics covered, with special focus on interactions with end users, and how to map their requirements and needs for both data products and research data-related services.

B.7. Federated AAI workshop

Rapporteur name: Licia Florio, Christos Kanellopoulos (GÉANT)

Presentation highlights

Christos Kanellopoulos and Licia Florio presented during the Federated AAI session organised during the EUDAT Conference. The main focus of the presentations was on reporting on the ongoing work in the AARC project, with particular emphasis on the AARC blueprint architecture (AARC BPA) and on its implementations.

Q/A and discussion points

There were a number of questions that were asked during the session. Most of the questions had a very technical focus. The AARC team has collected them and we'll add a Q&A section to the AARC website.

Conclusions reached

One of most frequently asked questions is how AARC BPA and related work can inform the work in EOSC Hub. AARC has paved the way to show that it is possible to define a reference AAI for research collaboration. It is expected that the results of the AARC pilots concerning the deployment of the BPA in production environment as well as policy best practices and technical guidelines will be a valuable starting point for EOSC-hub.

Luckily there is a significant number of key partners that are involved in both AARC and EOSC-hub to ensure that continuity of work.

Detailed Notes

The AARC BPA is a reference architecture to help research collaboration to implement an AAI without starting from scratch. The BPA offers defines an architecture and proposes different building blocks to support its implementation.

To date the AARC BPA is implemented by the three main e-infrastructures and by several research infrastructures, namely:

- EGI Check-In
- EUDAT B2Access
- GEANT eduTEAMS
- Elixir AAI
- Indigo Data Cloud
- DARIAH-EU

More pilots are ongoing in AARC to enable more research infrastructures to build an AAI that is compliant with AARC BPA. Christos reported in particular on the Life Science AAI pilot, which is now ongoing in AARC. The second phase of the pilot will end at the end of May 2018 when a production-like AAI will be available.

Why is this pilot so special? For two main reasons:

- The Life Science AAI will serve all 13 projects (but there may be more in the future) that operate in the life science field. The life science projects came to the conclusion that one single AAI would be a more sustainable and cost-effective way to enable users' access to life science services. For the first time there will be one AAI for a thematic community.
- EGI, EUDAT and GÉANT are working together to deliver the pilot, whilst AARC provide the necessary support.

B.8. EOSC as a "skills commons" providing FAIR training for FAIR data stewardship

Session Chairs: Angus Whyte (Digital Curation Centre), Gergely Sipos (EGI) and Ellen Leenarts (DANS)

Rapporteur name: Marjan Grootveld (DANS/EUDAT)*, with additions from others.*

Presentation highlights

The EOSC family of projects each addresses skills, from EOSCpilot to EOSC-hub and OpenAIRE Advance. The overall context was identified in the first EOSC HLEG report, which highlighted a large gap in data expertise. So EOSC needs to be prepared to meet that challenge, using approaches that scale-up.

EOSC-hub includes training as an integral part of the hub. It will be both domain-specific and generic; the domain-specific aspects will relate to thematic services, competence centres, and business pilots. Generic training will cover federated and collaborative services, common services including Data Management Planning (with OpenAIRE Advance, and service management (FitSM).

OpenAIRE Advance includes a RDM 'task force'. This will establish capacity and increase knowledge primarily among NOADs (National Open Access Desks) in order to support RDM activities, and support the Open Data Pilot/FAIR data. This will feed into a training WP which develops practical how-to style resources for putting OS principles into practice.

EOSCpilot is identifying a skills framework, training infrastructure and strategies to recommend for training in EOSC. A 'top 10 gaps' list ranges from FAIR policy development, 'soft' project and service management skills, through to more technical aspects of workflow description and cloud computing.

EOSCpilot skills framework aims to help organisations and individuals plan their skills development to make effective use of the EOSC. The framework is in two parts; firstly, competences that researchers and professionals need to acquire for open science and data science, and secondly the capabilities that EOSC services will help their research teams and organisations acquire.

The FAIRness of training resources, i.e. materials and information about events, has been an issue for EOSCpilot. The project has drawn on the exemplary approach offered by ELIXIR to consider how FAIRness might be applied to the skills and training resources made available through EOSC, especially to help users find them.

Conclusions reached

Questions were put to the workshop using an interactive polling tool. They were derived from previous discussion in EOSCpilot, e.g. at EOSC Stakeholders Forum. The responses indicated:

- When asked to identify their top 3 priorities for skill development in EOSC the workshop participants' top priority was 'support for train the trainer approaches'.
- The most feasible methods for making training materials FAIR were seen as 'adding identifiers and standard metadata' (findability) and 'non-restrictive licenses' (reuse).

Q/A and discussion points

The main discussion points were the following:

- The concept of "train-the-trainer" is diffuse: it typically transfers knowledge about particular content, and may also − or mainly − aim at increasing training skills.
- EOSC could have a central coordinating role in "harvesting" and presenting information related to Open Science, Data Science, RDM. This may be more efficient and effective than several infrastructures and organisations doing this individually.
- A core set of curated materials is desirable, to support trainers' capacity to deal with topics where the good practice essentials change frequently, or lack consensus. These should be based on RDA outputs where possible.

- There is a need to keep track of and visualise the popularity of training resources: this indicates partly the value of the resource. Gradually, EOSC coordinating information on training etc. could grow into developing/ suggesting/ validating/ endorsing some quality measurement. However, absence of quality stamps does not mean there is no quality.
- EOSC could encourage proactive outreach by the Research Infrastructures towards institutional RDM services, helping both sides to achieve their goals of broadening access to research communities and stimulating cross-disciplinary research. RI and e-Infrastructures could offer relevant materials and expertise to university research data services, in order to complement and enrich the cross-disciplinary training they offer, and fill gaps in disciplinary-focused materials. To add value, EOSC could e.g. organise "trainers/experts for hire" across the various projects and infrastructures. This would probably include some training of these trainers/experts.
- Large-scale and long-term research collaborations occupy a middle ground between data-intensive domains and the 'long-tail' of others. Collaboration partners have diverse practices and standards, so they have a strong need for mutual learning.
- e-Research Centres (e.g. Göttingen eResearch Alliance) have a role in coordinating local cross-institutional support, e.g. helping to build individual institutions' capacity to enlist 'data champions' who can address their needs for disciplinary-focused training.
- NOADs are important as national ambassadors/intermediaries. The e-infratructures and RIs often have national representatives, who may be an effective route for EOSC communication. National funding agencies can help with giving mandates, policies and guidance.
- Does EOSC also look beyond Europe? We should remind ourselves to do that.

Detailed Notes

Fifty-four conference attendants registered for this co-located session. The session started with about twenty attendants, which grew to twenty-five, but early departures for travel reduced this.

Introduction: aims and structure of the event (Kevin Ashley, DCC): Kevin introduced EOSCpilot and in particular the questions addressing skills and capabilities needed to use the EOSC services. The HLEG suggested we need a huge number of data-skilled people − how are we going to train and educate them? EOSCPilot has made progress on answering some questions − we use today's workshop to validate our approach to answering them. The questions are still open − we use today's workshop to help us move towards the answers.

Evolution of training provision in the EOSC projects: EOSCpilot (Angus Whyte), EOSC-hub (Gergely Sipos) and OpenAire Advance (Ellen Leenarts): First, Angus introduced the EOSCpilot skills work package. The focus was on data stewardship, defined as the "formalisation of roles and responsibilities to ensure that research objects are managed in accordance with FAIR principles and for long-term reuse". EOSCpilot is learning about skills gaps, e.g. with regard to data policy requirements, tool and domain standards, and workflows for cloud resource utilisation. These gaps are identified by desk research to analyse the landscape of skills resources at various infrastructures, institutions, service providers, either at national or international level. The so-called "science demonstrators" in EOSCpilot are a further source, as these are experimenting with service provision across a range of disciplines, and coming across infrastructure integration challenges. Currently the WP is planning surveys to engage with the Research infrastructures about how EOSC may amplify training and what resources are FAIR or have lasting value. Next, Gergely broke down the mission from EOSC-hub into four aspects: services (regarding e.g. data and tools), federation services (e.g. AAI, a marketplace set up and run according to principles of engagement), processes and policies (e.g. security regulations), and federated operations (e.g. lightweight certification of providers). The scope of training in EOSC-hub contains the nearly 50 services that are already in part of the project, as well as generic topics FitSM and data management planning. Ellen finally presented the two tasks in OpenAIRE Advance that relate to training: first, a task force Research Data Management that will increase the RDM knowledge of the National Open Access Desks, and help them to spread the message and tools to researchers. Second, the Open Science

Helpdesk will take a multiplier approach too, to assist researchers in the transition to open data by default and open science practices.

EOSCpilot Skills Framework- Mapping competences to service capabilities for data-intensive research (Angus Whyte): The EOSCpilot skills framework is based on existing competence frameworks from sources like the EDISON project. The framework also distinguishes between individual skills and organisational capabilities, and aims to show how services relate to both of these. The skills can also be phrased as user stories, to articulate the skills requirements for services. The capabilities can be expressed using language found in job descriptions for the professional groups involved in delivering open science and data science. These specify the organisational context that competences are applied in, and the levels of competence and responsibility expected. The next steps for the Skills Framework are is to compile examples of skills relevant to the services being defined for EOSC, and further define the capabilities these services offer. One possibility would be to compile examples of capabilities and competences relevant to EOSC from relevant job vacancies advertised by EOSC partners and stakeholder organisations. This could help other organisations plan their recruitment and staff development.

−FAIR training - applying FAIR principles to training resources (Ellen Leenarts): The first question is why one would make not just research data, but also training resources FAIR: users/researchers and trainers should be able to find and reuse relevant materials. FAIRness also fits the approach of Training as a Service (TaaS). Some considerations: updates of resources need versioning of PIDs (F1); what metadata are relevant (F2) − see for instance https://tess.elixir-europe.org/materials/, Celia van Gelder (ELIXIR) explains that TESS contains ELIXIR content as well as metadata scraped from various training portals (TESS might be a good practice for a similar EOSC-hub portal, as described in EOSCpilot D7.2). She refers to http://schema.org/ for work going on to standardise metadata.

Live poll and breakout groups

Responses to the interactive poll questions are indicated below. There were 16 respondents to these:

- About role: half of those present would rather have picked more than 1 role.
- About feasibility of FAIRness: why do 2 people think that licensing is hard?
- About kinds of examples: nearly even scores. “Something else”: there may be more needs than the three presented.
- About priorities: TtT and cataloguing what’s there (2 options) get the top 55%. “Something else”: advocacy for the needs and opportunity for training.
- About carrots and sticks: “other” scores high: both options; should just be part of one’s education as a researcher.

Three breakout groups took forward discussion begun at the EOSC Stakeholders Forum in November 2017. Groups 1 and 3 joined together, with 9 participants contributing. Group 2 had 7 participants.

Group 1. How can EOSC support research training providers to contribute to international level training infrastructure?

- Should FAIR principles be extended to training resources, and if so what kinds of resources are worth the effort to make reusable?
- Keep track of & visualise the popularity of training resources: this indicates partly the value of the resource
- How to define and measure the added value of training?
- Short-term and long-term feedback > reuse what’s been agreed by the working group on benchmarking the quality of RDM training. See the list of mandatory and optional questions for feedback forms: https://osf.io/6au9b/
- In general: adopt/endorse output from RDA working groups and don’t come up with new things
- What can EOSC offer that adds value. E.g. organise “Trainers/experts for hire” across the various projects and infrastructures. Would probably include some training of these trainers/experts.

- Can this be related to the envisioned Marketplace in EOSC-hub? Linking services and training (events, trainers, resources).
- NB: define "Service": merely IT services or also/mainly soft services?
- EOSC could have a central coordinating role in "harvesting" and presenting information related to Open Science, Data Science, RDM, ... That's a lot of work, but maybe more efficient and probably more effective than several infrastructures and organisations doing this themselves time and again. EOSC as an umbrella.
- Should EOSC perform quality assurance, certification of providers, or badging of content in a central catalogue of training materials and events harvested from participating organisations?
  - Gradually, EOSC coordinating information on training etc. could grow into developing/ suggesting/ validating/ endorsing some quality measurement. At least some requirements: principles of engagement (EOSCpilot) – not about the content.
  - Putting labels and quality stamps is difficult. If they are missing it does not mean there is not quality.
- Should EOSC monitor what is being provided and attempt to fill gaps either in the content or mode of delivery? [not discussed]

Group 2. How can EOSC assist research performing organisations to develop the competences and capabilities for open data science? How can EOSC assist institutions to plan the skills required to deliver their strategies and services for implementing FAIR principles, open science and data science? Should EOSC broker the supply and demand for disciplinary-focused training across institutions and research infrastructures?

- There are missed opportunities for RI and e-Infrastructures to offer relevant materials and expertise to university research data services, in order to complement and enrich the cross-disciplinary training they offer, and fill gaps in disciplinary-focused materials.
- EOSC must enhance the findability of skills resources and expertise across the research communities and infrastructures.
- e-Research Centres have a role in coordinating local cross-institutional support, for example helping to build individual institutions' capacity to enlist 'data champions' who can address their needs for disciplinary-focused training.
- EOSC could encourage proactive outreach by the Research Infrastructures towards institutional RDM services, helping both sides to achieve their goals of broadening access to research communities and stimulating cross-disciplinary research.
- Large-scale and long-term research collaborations occupy a middle ground between data-intensive domains and the 'long-tail' of others. Collaboration partners have diverse practices and standards, so they have a strong need for mutual learning.
- The EC has a welcome role in promoting standards. The 'blunt instrument' of directives seems effective, for example in the case of INSPIRE. The revision of the Charter and Code for Researchers suggested in the OSPP report on Rewards could also help promote use of standards and contribution to their development.
- A core set of high quality and up-to-date EOSC training resources would support the train-the-trainer approach. Train-the-trainer is an effective strategy only so far as the subject material is up-to-date and relevant, and for generalist trainers this can be challenging in topics where the scope and content quickly change, such as domain standards.
- What are the biggest gaps in cross-disciplinary skills for data stewardship?
- Suitable topics for a core tier of training resources would include: Citation, persistent identifiers, data protection, dealing with commercial sensitivities around data access, licensing, long-term preservation including migration and software or service dependency issues.

Group 3. How can EOSC coordinate national-level policies, strategies and reward mechanisms to stimulate open research data practices? What information could EOSC collect and publish to inform national-level policies, strategies and mechanisms?

- A lot of funding is expected from the countries, and the policies will be national-level.
- NOADs are important as national ambassadors/intermediaries.
- The e-infrastructures and RIs often have national representatives, so maybe (for the time being?) EOSC can better communicate with the e-infrastructures and RIs than with the member states.
- National funding agencies can help with giving mandates, policies and guidance on the Why and (generic!) How of Open Science and research data management.
- What can EOSC do to encourage and amplify efforts of funding bodies, institutions and other stakeholders to recognise researchers' skills for data stewardship and open research practices?
- What can EOSC do to nurture the career structures and rewards for professional support staff who contribute to open research practices?

Conclusions from breakouts

Topic 1:

- A marketplace of IT and soft services, offering information on training across Europe. Training catalogue highly desirable, from user perspective
- We need an explicit definition of "service"! Is training a service?
- Principles of engagement if you want your training to be in the "catalogue": Tick metadata boxes.
- Two scenario's: user perspective, as a researcher you want to find the most relevant courses, and a 'machine-interoperable perspective', so training can be found by IT services looking for it.

Topic 3:

- National-level policies and funding are essential to be a good researcher and use EOSC in the optimal way.
- New HLEG Expert group for EOSC that is talking to national ministries
- NOADs and other national intermediaries are very valuable.
- Let's not forget that "EOSC" is a metaphor and look beyond Europe.
- To enforce to work towards EOSC, perhaps create an EOSC website instead of EOSpilot and EOSChub and OpenAIRE Advance. This is currently discussed under the EOSC-hub − OpenaAIRE-Advance collaboration.
- Carrot approach is seen as way forward (rewarding data and application producers similarly to paper authors). Whose responsibility is to engage with publishers in the EOSC context?

Closing discussion: How can the EOSC help Research Institutions, Libraries and Infrastructures work to close skills gaps?

Significant willingness and funding from all member states is needed to make EOSC a reality. Fortunately a lot can be done without much money − but with effort and gaining consensus! − such as making data and training resources discoverable.

ANNEX C. PROGRAMME

The programme of the EUDAT conference is available at https://eudat.eu/eudat-conference-2018-programme

Monday 22 January 2018			
Co-located Events			
14:00 - 18:00	Sensitive Data Workshop (Part I)	Semantic Services in EOSC (Part I)	Array Databases for Research Communities

Tuesday 23 January 2018				
Co-located Events				
09:00 - 12:30	Sensitive Data Workshop (Part II)	Semantic Services in EOSC (Part II)	Research data management: interoperability, collaboration, and the research library role	SeaDataCloud workshop
12:30 -13:30	Lunch break			
EUDAT Conference starts				
13:30 - 13:45	EUDAT and the European Open Science Cloud Ecosystem – Kimmo Koski, Managing Director at CSC & EUDAT Coordinator (Presentation) Plenary Session 1: The European Open Science Cloud – Putting the Vision into Practice 13.45 - 14.15: The European Open Science Cloud: from principle to practical implementation – Augusto Burgueño Arjona, Head of Unit "eInfrastructure", Directorate General for Communications Networks, Content and Technology (DG CONNECT) (Presentation) 14.15 - 14.45: EOSC-Hub: first steps towards realising EOSC vision – Per Öster, EOSC-hub Project Director (Presentation) 14.45 - 15.30: Interactive panel discussion *Moderator:* Annabel Grant, Senior Stakeholder Engagement Manager, GÉANT *Panel members:* ▪ Augusto Burgueño Arjona, Head of Unit "eInfrastructure", DG CONNECT, European Commission ▪ Françoise Genova, Researcher, Centre de Donnees astronomiques de Strasbourg (CDS) ▪ Per Öster, Director, CSC & EOSC-hub Project Director ▪ Grazia Pavoncello, ministerial representative, Italian Ministry of Education, University and Research (MIUR) ▪ Alex Vermeulen, Carbon Portal Director, ICOS ERIC			
15:30 - 16:00	Coffee & Networking			

16:00 – 18:15	Plenary Session 2: The European Data Infrastructure (EDI) and the Data Challenge 16.00 - 16.45: SKA Regional Science Centers: A Platform for Global Astronomy – Michael Wise, Head of Astronomy, ASTRON – the Netherlands Institute for Radio Astronomy (Presentation) 16.45 - 17.30: Progress and latest updates on the EDI initiative – Serge Bogaerts, Managing Director, PRACE (Presentation) 17.30 - 18.15: Interactive panel discussion *Moderator:* Rob Baxter, Software Development Group Manager, EPCC, University of Edinburgh (Presentation) *Panel members:* ▪ Serge Bogaerts, Managing Director, PRACE ▪ Giuseppe Fiameni, Leader of the "Middleware for HPC services" group of the SuperComputing, Applications and Innovation department at CINECA ▪ Kimmo Koski, Managing Director at CSC & EUDAT Coordinator ▪ Sinead Ryan, Chair of Theoretical High Energy Physics, School of Mathematics, Trinity College Dublin Michael Wise, Head of Astronomy, ASTRON – the Netherlands Institute for Radio Astronomy
18:15 - 19:00	Poster session 1 minute-madness & launch of poster competition (Presentation) *Poster applicants will be given the opportunity to present their poster in one minute.*

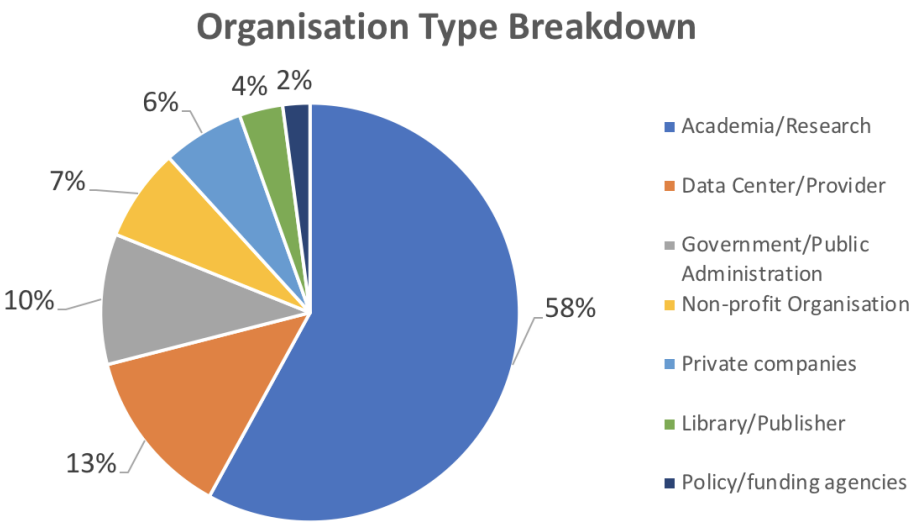
Wednesday 24 January 2018 | TOPIC: Researchers & research communities: Benefits, Opportunities and Open Issues				
EUDAT conference				
09:00 - 12:30	*EUDAT what's next?: CDI, EOSC & EDI*	*EUDAT Collaborative Data Infrastructure services*	*Cross e-Infrastructure collaborations*	*The impact of the policy framework on EOSC*
09:00 - 10:30	EUDAT CDI: Moving Forward & Next Steps	CDI services: building blocks and the way forward	Coupling data and HPC resources together to enable large scientific projects: the EUDAT-PRACE case	The Policy Framework: GDPR and All That
10:30 - 11:00	Coffee & Networking			
11:00 - 12:30	Research Infrastructures & the CDI	An in depth look at the future CDI services	Making data and cloud resources interoperable using	Restricted Data in the EOSC: What Are We Going to Do?

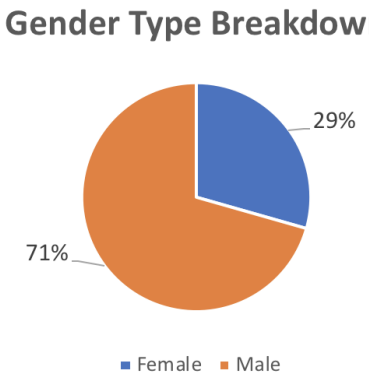
			EUDAT and EGI services	
12:30 - 13:30	Lunch Break			
13:30 - 17:00	*EOSC: what's in it for researchers & service providers?*	*Harvesting results from the EUDAT Community: Demonstrators & Pilots*	*Cross e-Infrastructure collaborations*	*User Engagement & Training*
13:30 - 15:00	EOSC-Hub & OpenAIRE Advance	Communities adoption of EUDAT services and Lessons Learned	Computing e-infrastructure with extreme large datasets	Training Experiences across Research Infrastructures
15:00 - 15:30	Coffee & Networking			
15:30 - 17:00	EOSC Principles of Engagement	Communities adoption of EUDAT services and Lessons Learned	Collaboration with SMEs and commercial stakeholders	Building a Trainers Community of Practice
17:00 - 17:30	**Plenary Session 3: Wrap-up from the parallel sessions**			
17:45 - 19:00	EUDAT2020 Council (closed meeting)			

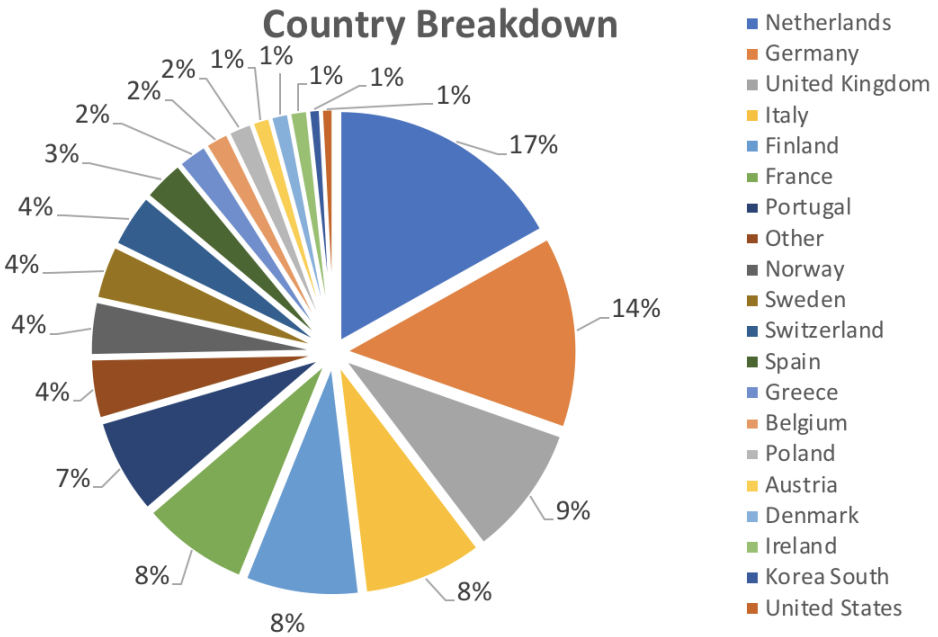
Thursday 25 January 2018				
Co-located Events				
09:00 - 13:00	ENVRI Workshop	Federated AAI Workshop	Piloting EOSC Governance Framework	EOSC as a 'skills commons' providing FAIR training for FAIR data stewardship

ANNEX D. PARTICIPANTS

238 stakeholders attended the EUDAT conference.





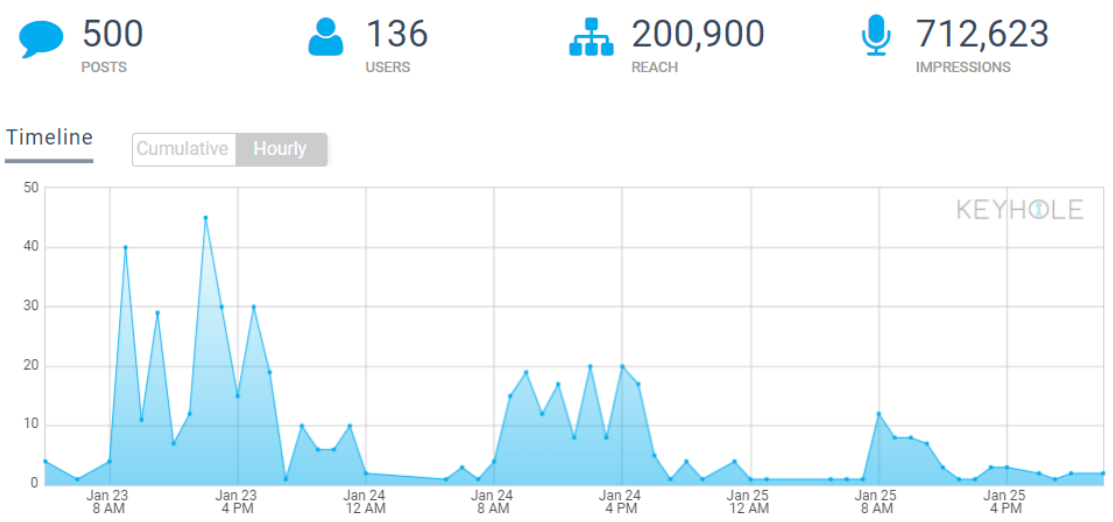


ANNEX E. IMPACT ON SOCIAL MEDIA

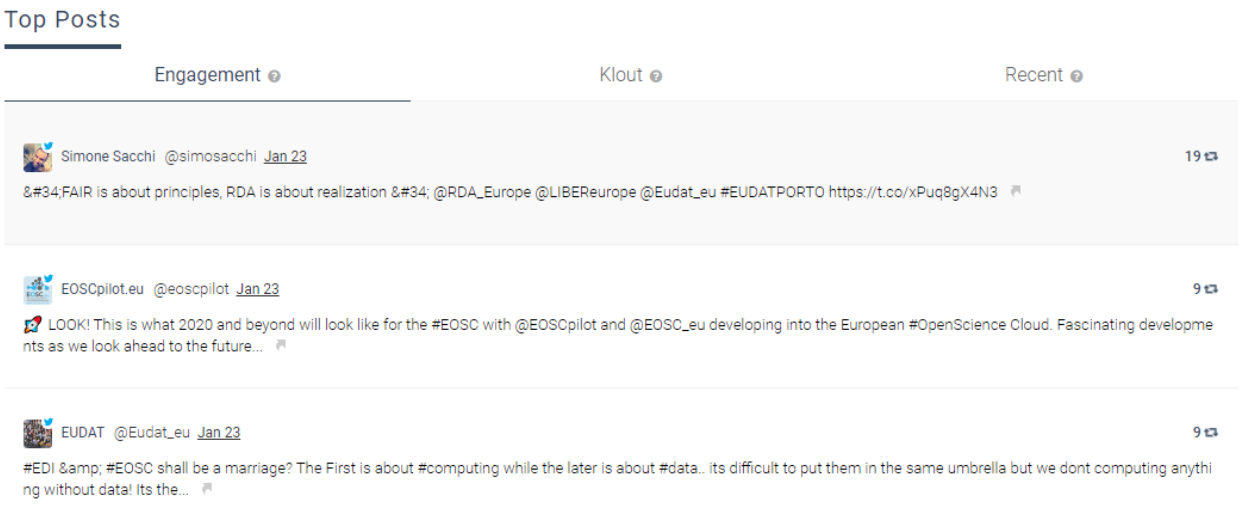
Before the EUDAT conference a set of communication and dissemination activities were performed to promote the conference and engage participants. The results were successful as over 230 relevant stakeholders attended the conference. A hashtag #EUDATPORTO was also created to engage remote participants during the conference and in this annex the usage statistics of the hashtag are reported.

#EUDATPORTO Usage Statistics from 22nd January 2018 until 25th January 2018 from www.keyhole.co

The image below shows the number of posts, number of users, total users reached and total number of impressions. The hashtag was mostly used on 23rd January, the first official day of the EUDAT Conference.



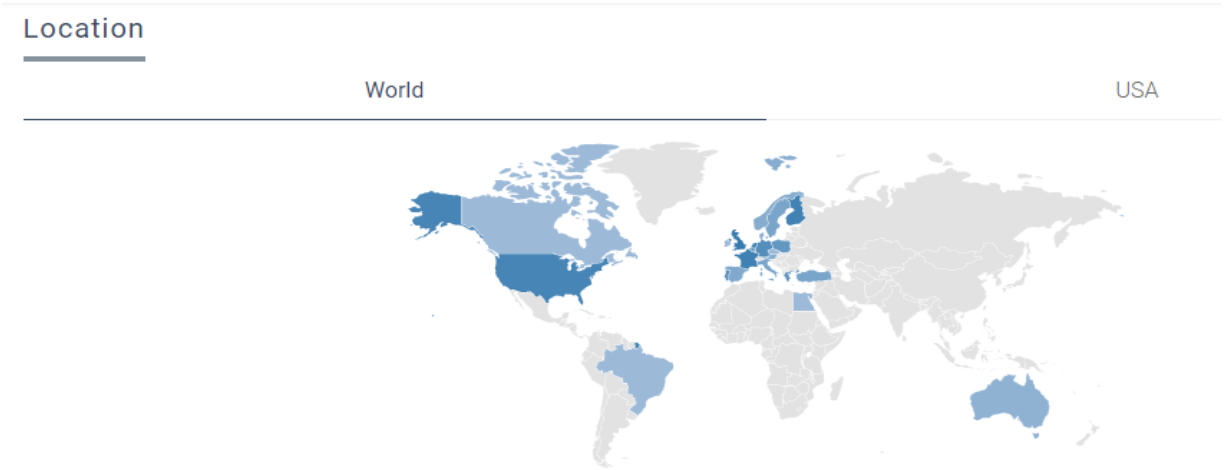
In image below reports the top 3 posts featuring the hashtag.



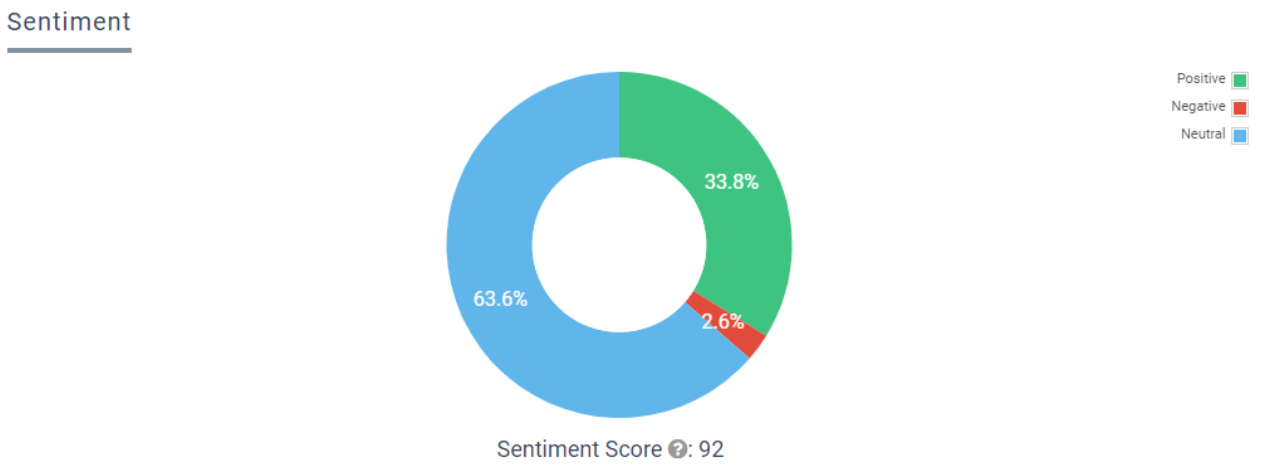
In the following picture a word cloud with the most used hashtags related to the #EUDATPORTO tweets is shown. EOSC, Open Science and Data Management are the top 3 hashtags mentioned.

The picture below, on the other hand, shows a word cloud with the most used keywords related to the #EUDATPORTO tweets. Data, Research and conference are the top 3 keywords mentioned.

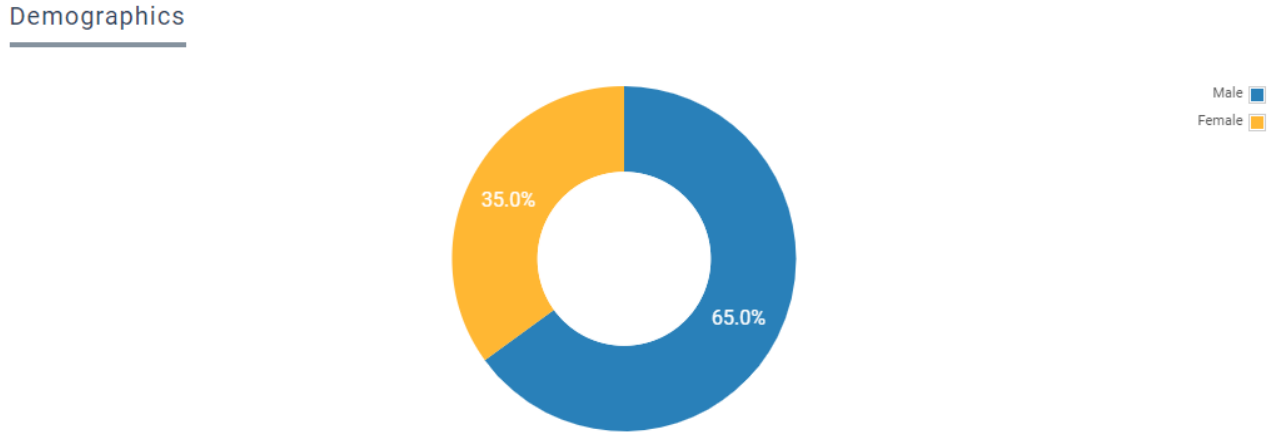
The picture below shows where the tweets originated: mostly from Europe (Finland, France, UK), but also from the United States.



In the image below the reaction to the tweets is also reported: 33.8% positive, 63.6% neutral, 2.6% negative.



Finally, the image below shows the demographics of the Twitter users related to the hashtag #EUDATPORTO, which were mostly male (65% males, 35% females).



ANNEX F. PARTICIPANT FEEDBACK

The conference participants were asked to complete the following feedback form to assess the conference. 95% of the respondents to the survey confirmed that the EUDAT conference met their expectations. 59% of the respondents rated the 2 plenary sessions between good and excellent. Positive feedback was also given on the different breakout sessions and co-located events. The sessions on EOSC and EUDAT services were particularly appreciated. 40% of the respondents assessed the social events and the poster session as Excellent. Points raised on things that could have been improved were: more presentations on practical experiences and that an extra half day would have eased the time pressure (the programme was a bit too packed).

Please rate the co-located workshops: 22-23 January 2018

- Sensitive Data Workshop
- Semantic Services in EOSC
- Array Databases for Research Communities
- Research data management: interoperability, collaboration, and the research library role
- SeaDataCloud workshop

Please rate the Conference Sessions – 23 January 2018

- Plenary Session 1: The European Open Science Cloud – Putting the Vision into Practice
- Plenary Session 2: The European Data Infrastructure (EDI) and the Data Challenge
- Poster session 1 minute-madness & launch of poster competition

Please rate the Conference Sessions – 24 January 2018

- EUDAT what's next?: CDI, EOSC & EDI
- EUDAT Collaborative Data Infrastructure services
- Cross e-Infrastructure collaborations
- The impact of the policy framework on EOSC
- EOSC: what's in it for researchers & service providers?
- Harvesting results from the EUDAT Community: Demonstrators & Pilots
- User Engagement & Training

Please rate the co-located workshops – 25 January 2018

- ENVRI Workshop
- Federated AAI Workshop
- Piloting EOSC Governance Framework
- EOSC as a 'skills commons' providing FAIR training for FAIR data stewardship

Conference overall assessment

Did the EUDAT Conference meet your expectations? Yes/No. Please elaborate your answer

How would you rate the networking opportunities & social dinner at the EUDAT Conference?

Please elaborate your answer

Was the EUDAT Porto Conference the first EUDAT event you attended? Yes/No?

Your suggestions: Please use this space to provide your suggestions, ideas and feedback on the EUDAT Porto Conference as well as on future events.

ANNEX G. GLOSSARY

Term	Explanation
AAI	Authentication and Authorization Infrastructure
API	Application Programming Interface
CDI	Collaborative Data Infrastructure
CoP	Community of Practice
DMP	Data Management Plan
DPMT	Data Policy Management Tool
DOI	Digital Object Identifier
DPO	Data Protection Official
EDI	European Data Infrastructure
EGA	European Genome-phenome Archive
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
GEF	Generic Execution Framework
HLEG	High Level Expert Group
HPC	High-Performance Computing
ICT	Information & Communications Technology
NOAD	National Open Access Desk
OGC	Open Geospatial Consortium
PID	Persistent Identifier
PRACE	Partnership for Advanced Computing in Europe
RDA	Research Data Alliance
RDM	Research Data Management
SKA	Square Kilometer Array
SKOS	Simple Knowledge Organization System
SME	Small to Medium Enterprise

