Report of first EOSCpilot External Advisory Board meeting

Author(s)	Juan Bicarregui (editor)
Status	Final
Version	v10
Date	10 April 2018

Dissemination Level

☒ PU: Public

□ PP: Restricted to other programme participants (including the Commission)

□ RE: Restricted to a group specified by the consortium (including the Commission)

□ CO: Confidential, only for members of the consortium (including the Commission)

Abstract:

The first EOSCpilot External Advisory Board (EAB) was held on 29th November 2017 in Brussels immediately following the first EOSC Stakeholders form. Discussion was focused around the following four questions:

1. *What do research communities need from an "Open Data Science Environment"?*
2. *How can EOSC deliver integrated services that are relevant to community needs?*
3. *What changes are needed in capability and practices?*
4. *How should provision be overseen to maximize benefit?*

This report summarises the discussion and highlights a number of recommendations that arose from it.

The European Open Science Cloud for Research pilot project (EOSCpilot) is funded by the European Commission, DG Research & Innovation under contract no. 739563

Document identifier: EOSCpilot –WP8-D8.3	
Deliverable lead	N/A
Related work package	**WP1**
Author(s)	**Juan Bicarregui (editor)**
Contributor(s)	**Members of External Advisory Board, see Annex.**
Due date	**N/A**
Actual submission date	**N/A**
Reviewed by	**External Advisory Board**
Approved by	**N/A**
Start date of Project	**N/A**
Duration	**N/A**

Version	Date	Authors	Notes
10	**10 April 2018**	Juan Bicarregui and EAB	

***Disclaimer**:* The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EOSCpilot Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EOSCpilot Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EOSCpilot Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

TABLE OF CONTENT

EXECUTIVE SUMMARY

The first EOSCpilot External Advisory Board (EAB) was held on 29th November 2017 in Brussels immediately following the first EOSC Stakeholders forum. The EAB exists to advise the EOSCpilot project, however, as EOSCpilot is itself advising on EOSC, the board considered it reasonable to consider wider issues around EOSC in general. Discussion was focused around the four questions highlighted below and the following ten recommendations arose from the discussion.

Question 1. What do research communities need from an "Open Data Science Environment"?

Recommendation 1. *It is critical to strike the right balance between domain-specific and cross-domain provision of resources and services. The ability to find and access interdisciplinary services, and resources from other domains, will enable domain specific portals to provide access to wider range of services, whilst domain specific resources remain best accessed through domain specific portals.*

Question 2. How can EOSC deliver integrated services that are relevant to community needs?

Recommendation 2. *There is some confusion in the use of the term open. Open is sometimes used for federation of technology and data, and other times for transparent, collaborative, and sharable research artefacts such as papers, data and software. There is a need to clearly articulate the separate goals of (i) transparency and sharing of science, from (ii) federation of technology, and to share the rationale for the chosen direction broadly.*

Recommendation 3. *There are many barriers, both real and perceived, to interoperation. Barriers that are shared by many domains could be prioritised and tackled in a uniform way. In this approach, the EOSC is not so much about building new solutions but about removing common barriers to open science, one by one.*

Recommendation 4. *There is need to ensure accommodation of generic tools that are provided from inside and outside the EOSC and from international activities which are governed independently of EOSC. To this end, the parties who provide these tools need to be involved in global discussions around the Principles of Engagement.*

Recommendation 5. *In order to achieve a single EOSC infrastructure, there will need to be clearly defined complementary roles and close collaboration between forthcoming projects in the INFRAEOSC programme. The board recommends that such collaboration would be helped by having overlapping membership between the advisory boards for the projects, or even a single advisory board across all the projects.*

Question 3. What changes are needed in capability and practices?

Recommendation 6. *EOSC's skills provision needs to embrace and support training providers who operate independently of EOSC but who develop skills relevant to its use.*

Recommendation 7. *EOSC services should clearly describe the skills and organizational capabilities required for their use using a standard machine-readable template.*

Recommendation 8. *Sharing of research outputs should not be linked to judgment of their value. Consideration should be given to schemes that give recognition to research roles by means other than authorship of papers.*

Questions 4. How should provision be overseen to maximize benefit?

Recommendation 9. *The goal of service provision that is free at point of delivery is valuable aim and will be appropriate for many EOSC services. However, different models may be required for different resources and service provision.*

Recommendation 10. *Current project-based mechanisms are not well suited to developing and sustaining infrastructures. A means for funding infrastructure should be found that balances the need for coordination across providers with the transparency and creativity that arises from genuinely competitive calls.*

1. INTRODUCTION

The first EOSCpilot External Advisory Board (EAB) was held on 29th November 2017 in Brussels immediately following the first EOSC Stakeholders forum.

The EAB exists to advise the EOSCpilot project. However, as EOSCpilot is itself advising on the EOSC, the board considered it reasonable to undertake wider considerations around the EOSC in general. This report summarises the discussion and recommendations arising from the first meeting of the EAB. A second meeting of the Board is expected to take place in late 2018.

Discussion was focused around the following four questions which form the four headings in this report:

1. *What do research communities need from an "Open Data Science Environment"?*
2. *How can EOSC deliver integrated services that are relevant to community needs?*
3. *What changes are needed in capability and practices?*
4. *How should provision be overseen to maximize benefit?*

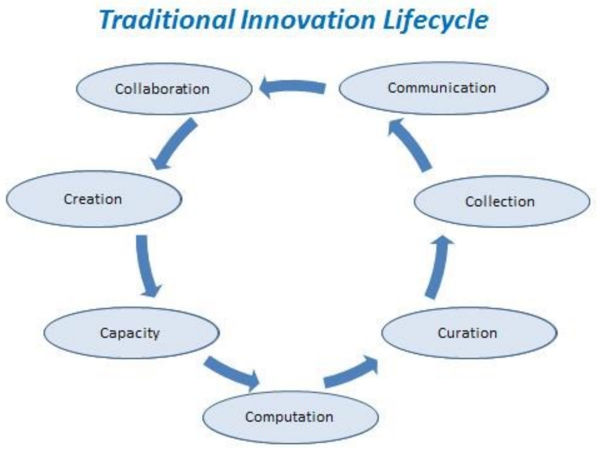
EOSCpilot is concerned with making the first steps "from vision to implementation" of the European Open Science Cloud. EOSCpilot is just a pilot: it is not "building the EOSC", nor is it a comprehensive design study. But it does provide input to future INFRAEOSC projects.

1.1. Context

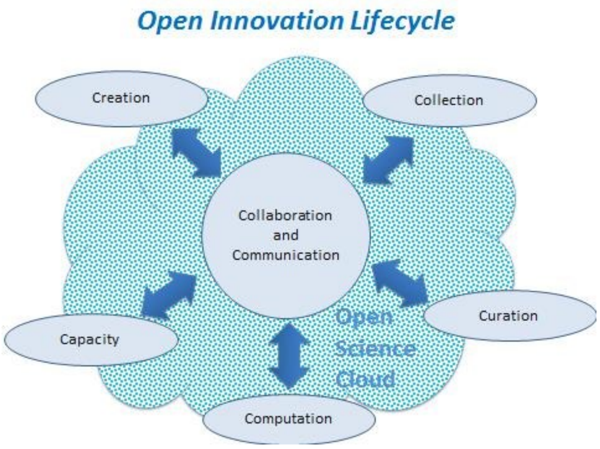
Collaboration and communication are central to open research and new forms of these are reforming the conventional cyclic research processes where communication has been primarily through papers. Open science involves:

- A shift from the practice of sharing results through articles in scholarly journals to sharing all available knowledge at each stage in the research process.
- A shift towards a scientific process that is based on wider cooperative work.

The diagram below shows how the research process can be made more effective through the provision of an "open-data science environment"[1]. On the left hand side, the conventional research lifecycle is shown where communication of results in provided at the end of the research in the form of papers. The right hand side shows how the open science cloud infrastructure enables communication and formation of new collaborations at any stage in the lifecycle. The EOSC will provide the e-Infrastructure to support these new processes in Europe.



Traditional Innovation Lifecycle



Open Innovation Lifecycle

[1] The phrase "open-data science environment" is used in the 2015 Berlin Communiqué from G7 Science Ministers. https://www.bmbf.de/files/English_version.pdf

EOSCpilot's activities are concerned with removing any barriers that stand in the way of achieving the two changes described above. This brings three different types of challenge: scientific challenges, technical challenges and cultural challenges. The scientific challenges are really opportunities: to enable new forms of research through greater interdisciplinarity and more collaborative use of data and other resources. Interdisciplinarity is an important goal that has been emphasized in many funders' policies but it also introduces challenges as communities start from different baselines, with different practices in different fields, and with infrastructures that may be incompatible. These are some of the technical and cultural challenges that need to be overcome. These challenges will be addressed in the coming years through a number of projects contributing to the development of the EOSC at National and European level.

2. WHAT DO RESEARCH COMMUNITIES NEED FROM AN "OPEN DATA SCIENCE ENVIRONMENT"?

2.1. Providing access to services

In order to ground the discussion, the Board considered examples of service provision in the life sciences sector although many of the issues related to this question are general. A major challenge in life sciences, that is also present in other domains, is translating benefits from research to practice. To this end, the biological and medical Research Infrastructures work together through CORBEL[2]. Some have a dual role as both users and providers of services.

In the life sciences, there is a particular need to integrate areas of technology addressing areas such as genomics, proteomics and clinical data. This can be called "intra-domain integration". On the other hand, other forms of integration are truly "inter-domain". For example, services such as AAI should clearly be shared across all communities as it does not make long-term sense for any RI to build its own. The handling of personal data also requires common standards across all domains although the need is more critical in domains where personal data is central to the research.

A feature of many domains is the proliferation of data resources. In molecular biology alone there are at least 1800 data resources globally. Some of these are very mature whilst others have been created recently. There are many *de facto* domain standards such as those used by the PDB database[3]. Users access resources at all points of the research pipeline through domain specific portals that provide disciplinary targeted views of available services. There are also examples of generic services, such as HPC training, that are discoverable through domain specific portals such as TeSS and BioSchemas. It is hard to envisage how a single cross-disciplinary point of access could provide equivalent ease of use for all domains, although it is certainly possible that the functionality of domain-specific portals could be enhanced behind the scenes by integration with resources from other domains.

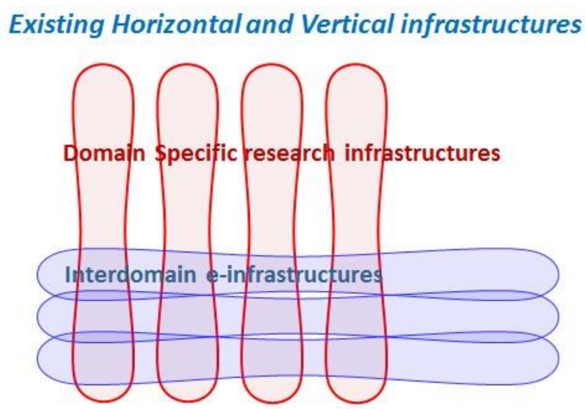
An inter-domain catalogue of services is therefore seen primarily as a machine-to-machine interface that will provide a basis for domain specific user facing portals. This is depicted in the diagram below[4]. The existing vertical and horizontal infrastructures shown in the left hand diagram are not replaced on the right hand side but are augmented to enable greater sharing at various levels through EOSC interfaces. The upper levels of the vertical infrastructures are still available to domain specific users who can now access a wider range of services and resources through their existing portals. This is enabled through a new inter-disciplinary layer that spans across horizontal and vertical provision and exposes services provided by the existing infrastructures. The lower level services in the vertical infrastructures are thereby made available

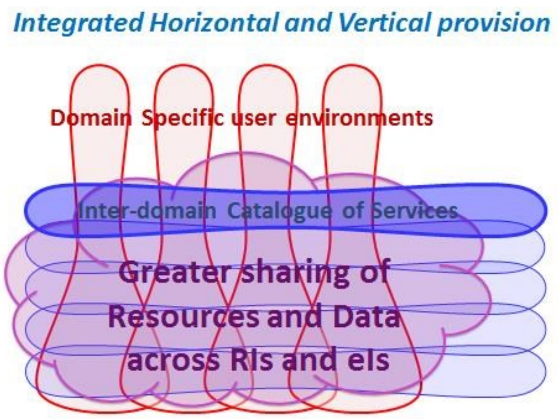
[2] CORBEL - Coordinated Research Infrastructures Building Enduring Life-science Services. http://www.corbel-project.eu/home.html

[3] The world wide Protein Data Bank see http://www.wwpdb.org/ or https://doi.org/10.1007/978-1-4939-7000-1_26

[4] Diagram derived from eIRG concept of e-Infrastructure Commons. eg http://e-irg.eu/documents/10920/363494/2017-Supportdocument.pdf

for wider use where appropriate whilst remaining primarily focused on supporting the requirements of their own communities.


Existing Horizontal and Vertical infrastructures
Domain Specific research infrastructures
Interdomain e-infrastructures


Integrated Horizontal and Vertical provision
Domain Specific user environments
Inter-domain Catalogue of Services
Greater sharing of Resources and Data across RIs and eIs

<u>**Recommendation 1.**</u> *It is critical to strike the right balance between domain-specific and cross-domain provision of resources and services. The ability to find and access interdisciplinary services and resources from other domains will enable domain specific portals to provide access to wider range of services, whilst domain specific resources remain best accessed through domain specific portals.*

3. HOW CAN EOSC DELIVER INTEGRATED SERVICES THAT ARE RELEVANT TO COMMUNITY NEEDS?

The EOSC will combine the offerings of the many current research infrastructures, e-infrastructures, data repositories etc. In order to do this, existing infrastructures may need to adapt their services to support standards that enable interoperation. The EOSC will then emerge as a “system of systems” that provides seamless access to resources across disciplines where this requirement is driven by scientific needs.

Four attributes of such a system of systems are:

- operational and management independence,
- evolutionary development,
- geographical distribution, and
- heterogeneity of constituent systems.

EOSCpilot is defining a framework to enable services to connect and defining agreements on how data and services can work together across domains. There is a need for mechanisms to “hand off” responsibility for services to particular domains. As well as the services provided by current research infrastructures and e-infrastructures, many commonly used externally provided services and tools (such as GitHub) would continue have a place in the EOSC. Such services should interoperate with EOSC rather than be duplicated within it.

The board reflected on the fact that two aims of the EOSC were to improve openness and to improve federation and observed that these two aims were to some degree independent. Openness does not require federation, nor does federation require openness, although some barriers are common to both, and can be addressed together.

At this point it is not clear how the proposed concepts of federation of services as a system of systems, and the federation of governance that comes from this will remove barriers to openness. Is data portability best addressed through federation? Is the system-of-systems model robust to service ‘drop-outs’, especially with regard to the core enabling services? What *is* clear however is that achieving the intended federation will be very complex, both technically and socially. In particular it is likely that some people will question whether the investment required to integrate with the proposed EOSC federation will yield added value in their domain. These are strong cultural barriers that will need to be addressed if the EOSC is to be successful.

A domain-centered activity where existing infrastructures create inventories of the barriers that each faces with regard to the two aspects of open science described above might be helpful. These inventories should consider both intra-domain and inter-domain cooperative science, as well as collaboration at a global scale. Barriers that are shared by many domains could be prioritised and tackled in a uniform way. In this approach, the EOSC is not so much about building new solutions but about removing common barriers to open science, one by one.

Some candidates for immediate consideration seem to be:

- AAI (Authentication and Authorization Infrastructure)
- High throughput networks (cf. US Science DMZ),
- Easy access to commercial cloud compute services,
- Data portability so data from one community can be used and understood by tools of another community (e.g. US DataONE project)
- Operation with externally provide common tools (e.g. GIThub)
- Collaboration at a global scale (e.g. through RDA) , and
- Knowledge representation to enable semantic interoperability.

The Board agreed that strong cooperation between projects will be essential to address the many challenges inherent in developing the EOSC to benefit science. In particular, it must be clear that the projects in the INFRAEOSC programme are working towards a common vision of the EOSC, albeit by providing complementary aspects of it, and that this vision is consistent with the needs of the research communities. A mechanism for encouraging this might be to have overlaps between the projects' advisory boards or even a single advisory board across several projects.

Recommendation 2. *There is some confusion in the use of the term open. Open is sometimes used for federation of technology and data, and other times for transparent, collaborative, and sharable research artefacts such as papers, data and software. There is a need to clearly articulate the separate goals of (i) transparency and sharing of science, from (ii) federation of technology, and to share the rationale for the chosen direction broadly.*

Recommendation 3. *There are many barriers, both real and perceived, to interoperation. Barriers that are shared by many domains could be prioritised and tackled in a uniform way. In this approach, the EOSC is not so much about building new solutions but about removing common barriers to open science, one by one.*

Recommendation 4. *There is need to ensure accommodation of generic tools that are provided from inside and outside the EOSC and from international activities which are governed independently of EOSC. To this end, the parties who provide these tools need to be involved in global discussions around the Principles of Engagement.*

Recommendation 5. *In order to achieve a single EOSC infrastructure, there will need to be clearly defined complementary roles and close collaboration between forthcoming projects in the INFRAEOSC programme. Collaboration would be helped by having overlapping membership between the advisory boards for the projects, or even a single advisory board across all the projects.*

4. WHAT CHANGES ARE NEEDED IN CAPABILITY AND PRACTICES?

EOSCpilot has a workpackage addressing the dual issues of skills and capabilities. As with many areas of the project, this workpackage is not intended to produce fully implemented solutions but will deliver clear, tested pathways to those solutions and test them for feasibility where possible. It is already clear that, whatever form EOSC takes, some of the solutions to the challenges in skills will come from outside EOSC itself and its strategy and delivery model must take account of this.

4.1. Skills

For skills, the project has done work to map the skills landscape and build on frameworks for skills produced by others such as by the EDISON project. Attention is particularly focused on the skills required by researchers and those who support them, such as data stewards, when using EOSC services and also the skills of data scientists. It has produced a competence framework (D7.1) and is now considering the issues of how people best acquire particular skills, what provision already exists and what gaps need filling. As a consequence, the project will propose a strategy for the sustainable development of skills.

Skills should be distinguished from training: training is one way to acquire skills. Furthermore, there are already some excellent existing initiatives, such as software/data carpentry, which EOSC needs to acknowledge, perhaps support, but not compete with or duplicate. What is also clear is that there is excellent work taking place within research infrastructures that should be maintained and supported. The gaps identified so far relate to cross-infrastructural requirements and most existing provision is focused on single e-infrastructures and research infrastructures, with a few notable exceptions which consider tasks more holistically. The EOSC vision requires ease of integration of data and services from multiple providers, which requires technical interoperability but also skills which cross domains. Some of the skills requirements need to be dealt with within educational institutions and to that end the workpackage is engaging with universities on curriculum issues, particularly amongst those championing data science.

Recommendation 6. *EOSC's skills provision needs to embrace and support training providers who operate independently of EOSC but who develop skills relevant to its use.*

4.2. Capability

A complementary area of work concerns capabilities that organisations, such as research groups, universities, institutes, companies, need to have in order to allow their skilled staff to utilize EOSC effectively. One requirement identified so far is that EOSC's services need to articulate more clearly what skills and capabilities are required to make use of them and to do so in a consistent way.

Recommendation 7. *EOSC services should clearly describe the skills and organizational capabilities required for their use using a standard machine-readable template.*

4.3. Practices

It was pointed out that many perceive strong reasons for not publishing data, fearing loss of credit and the extra work needed to prepare data for publication, noting that the second worry might actually be a symptom of the first since few researchers complain about the work needed to write papers. On the other hand, care needs to be taken to avoid pressuring researchers into premature release of data and prevent researchers "gaming the system" to advance their reputation by promoting unvalidated research outputs.

How can one deal with such issues? One way to address these concerns might be to highlight examples where positive change has already happened, such as in astronomy. It was noted that making data sharing possible and giving credit for it are two separate steps. Models of scientific communication are emerging where distribution of results has been separated from value judgment of the significance. (e.g. Wellcome Open Research[5]) and where credit is given for different roles in research (e.g. CRediT[6]) rather than a single model of paper authorship. Such measures need to be adopted by research-performing organisations who are the bodies closest to researchers. The question to consider is how EOSC can effectively support these organisations to achieve the necessary changes in practice.

[5] https://wellcomeopenresearch.org/

[6] http://docs.casrai.org/CRediT

Recommendation 8. *Sharing of research outputs should not be linked to judgment of their value. Consideration should be given to schemes that give recognition to research roles by means other than authorship of papers.*

5. HOW SHOULD PROVISION BE OVERSEEN TO MAXIMIZE BENEFIT?

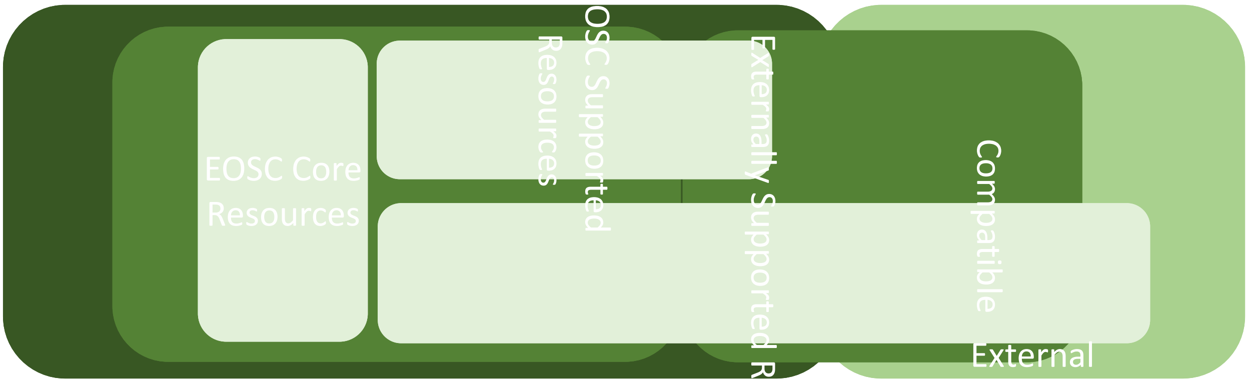
5.1. Overseeing heterogeneous provision

EOSCpilot has produced a draft framework for governance of the EOSC[7]. The framework outlines a three-layer governance model consisting of: a Strategic Layer that forms the strategic vision and objectives of the EOSC; an Executive Layer that commissions core resources and ensures supported services are properly compensated to meet the strategic objectives; and a Steering layer that represents the views of user communities through, for example, a Stakeholder Forum.

The steering layer has elsewhere also been called an advisory layer, which reflects that its role should be both advisory and steering. It was noted that whilst the strategic layer might play a strong steering role during the initial development of the EOSC, the steering layer should have an increasing prominent role as the EOSC matures.

The draft governance framework outlines a number of "strategic requirements" and "stakeholder requirements" for the governance of the EOSC, but it was noted that these "requirements" are not absolute necessities, but rather more like general goals

It was noted that there is considerable work to be done to define the optimum business model for EOSC services and resources. Although services may be 'free at the point of use'[8], service provision will need to be compensated in some way. EOSCpilot is currently working on analysing possible business and funding models for EOSC services and some initial sketches of models are provided in the document. It is likely that different funding models will be appropriate for different types of service. For example, the EOSC core services that make the EOSC function may need a different model of provision than services targeted at specific communities. For EOSC to gain community acceptance, it is important that the rationale for these different models is clearly explained and universally agreed within the community.



The EOSC ecosystem will include resources that are supported through different mechanisms

[7] https://eoscpilot.eu/content/d22-draft-governance-framework-european-open-science-cloud

[8] *"The European Open Science Cloud aims to give Europe a global lead in scientific data infrastructures, to ensure that European scientists reap the full benefits of data-driven science. Practically, it will offer 1.7 million European researchers and 70 million professionals in science and technology a virtual environment **with free at the point of use**, open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines".* [European Cloud Initiative - Building a competitive data and knowledge economy in Europe, EC COM(2016) 178]

Recommendation 9. *The goal of service provision that is free at point of delivery is valuable aim and will be appropriate for many EOSC services. However, different models may be required for different resources and service provision.*

5.2. Developing a single infrastructure through project-based funding

It has been observed elsewhere that project-based funding is not well suited to infrastructure development[9]. It was noted that in order to ameliorate the heterogeneity that inevitably arises from project based funding, some calls aimed at infrastructure development are being envisaged for delivery through single projects with size equal to the whole budget for the call. Often only one credible proposal is submitted against these calls.

The Board considered the pros and cons of these "single project" calls. On the positive side, this can be a way to involve many interested parties in a single project and so deliver a more homogenous result. On the other hand, "monopoly" proposals can militate against transparency by, in effect, placing a great deal of control in the hands of the proposal coordinator.

An alternative model, that is used in some calls, it to deliver a project through a small core project team and include broader community engagement through the "cascading call" mechanism that provides a relatively easily way to transparently widen participation.

Although this mechanism was not available for EOSCpilot, the need to involve additional participants during the project has been met by expanding the consortium through open calls for further science demonstrators that selected 10 additional demonstrators through a formal review procedure. However the grant amendment procedure used to implement this expansion carries a relatively high administrative burden.

Care is required when developing infrastructure through project based funding to balance the need for coordination across service providers against the risks of delivery through monopolistic projects. Where it is envisaged that the goals of a particular call would be best delivered through a single project, a mechanism should be found that reintroduces competitiveness and creativity into responses. It should be possible for funders and prospective providers to discuss and adjust work plans during the review process or even for different aspects of the call to be addressed by separate proposals that are then brought together, post review, into a single project.

Recommendation 10. *Current project-based mechanisms are not well suited to developing and sustaining infrastructures. A means for funding infrastructure should be found that balances the need for coordination across providers with the transparency and creativity that arises from genuinely competitive calls.*

[9] For example "*There is an intrinsic tension between the structural character of data networks and the time-limited, project-based funding models available for many research programmes.*" [Co-ordination and support of international research data networks, OECD, December 2017, http://dx.doi.org/10.1787/e92fa89e-en]

Annex. Membership of External Advisory Board

Membership of EAB can be found at: https://eoscpilot.eu/about/external-advisory-board

The following EAB members were present at meeting:

Ron Appel (Swiss Institute of Bioinformatics)
Julian Bauer (European University Association) for Lidia Borrell-Damian
Francoise Genova (Research Data Alliance)
Tony Hey (STFC) - Chair
Kate Keahey (Argonne National Laboratory)
Robert Kiley (Wellcome Trust)
Karel Luyben (Technical University of Delft)
Daphne Raban (University of Haifa)
Herbert van de Sompel (Los Alamos National Lab)
John Womersley (European Spallation Source)

The following EAB members not able to be present at meeting:

Bennie Fanaroff (Former Director SKA South Africa)
Robert Grossman (Genomic data commons)
Simon Hodson (CODATA)
Frank Jenko(Max Plank Institute Plasma Physics)
Ross Wilkinson (Australian National Data Service)

The EOSCpilot team members present at meeting were:

Juan Bicarregui (STFC, EOSCpilot coordinator)
Kevin Ashley (DCC)
Matthew Dovey (Jisc)
Simon Lambert (STFC)
Brian Matthews (STFC)
Andrew Smith (ELIXIR)
Prodromos Tsiavos (Onassis Cultural Centre)