D4.4: Consolidated Science Demonstrator evaluation report

Author(s)	Hermann Lederer, John Kennedy (MPG) Steven Newhouse (EMBL-EBI)
Status	**Final**
Version	V 1.0
Date	19/12/2018

Abstract:

The Science Demonstrators play an essential role as early adopters of EOSC from a range of science areas. Their input is used to drive and prioritize the integration of the EOSC services in a common homogeneous platform. Ten Science Demonstrators have been selected through Open Calls (see D4.2), after the pre-selection of five Science Demonstrators prior to the start of the EOSCpilot project. The fifteen Science Demonstrators have carried out their work within EOSCpilot according to the Engagement Model developed and described in D4.1.

The Science Demonstrators' project executions were guided by domain and IT aware experts, and these shepherding activities and project progress have been discussed and monitored in regular meetings. Feedback from the Science Demonstrators was collected in interim and final reports, as well as from participation in two Stakeholder meetings.

This Deliverable D4.4 now presents an evaluation of the findings, feedback, lessons learnt and recommendations of the Science Demonstrators as representative groups of customers of future EOSC services.

Dissemination Level

X	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)

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 –WP4-3	
Deliverable lead	**MPG**
Related work package	**WP4**
Author(s)	**John Kennedy, Hermann Lederer (MPG)** **Steven Newhouse (EMBL-EBI)**
Contributor(s)	
Due date	**31/01/2019**
Actual submission date	**29/01/2019**
Reviewed by	**Tiziana Ferrari (EGI)** **Brian Matthews (UKRI)**
Approved by	**Mark Thorley (UKRI)**
Start date of Project	**01/01/2017**
Duration	**28 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	19/10/2018	Hermann Lederer (MPG)	Structure
0.2	20/11/2018	Hermann Lederer (MPG)	First draft
0.3	06/12/2018	John Kennedy (MPG)	Details of Evaluation chapter
0.4	07/12/2018	Hermann Lederer + John Kennedy (MPG)	Polishing + Abstract + Conclusions
0.5	10/12/2018	Steven Newhouse (EMBL)	Corrections and improvements
0.6	11/12/2018	Hermann Lederer (MPG)	Ready for internal review
0.7	13/12/2018	Hermann Lederer + John Kennedy (MPG)	After review by Tiziana Ferrari (EGI), addressing her review comments and suggestions
1.0	19/12/2018	Hermann Lederer (MPG)	After review by Brian Matthews (STFC), Addressing his review comments and suggestions

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The Science Demonstrators play an essential role as early adopters of EOSC, and in providing feedback on its candidate service portfolio, from a range of science areas to stimulate the engagement of the science communities and stakeholders in Open Science, by building on the expertise of the research infrastructures and their service providers. Their input will be used to drive and prioritise the integration of the candidate EOSC services to meet the functional and non-functional needs of researchers, and to ensure that governance structures provide the guidelines needed by researchers.

To achieve this goal, in first steps structures and processes have been developed to enable the selection of best suitable Science Demonstrators through open calls and to execute Science Demonstrator projects along common rules of engagement. Accordingly, 15 Science Demonstrators after respective selection processes have carried out their work plans. In monthly meetings the Science Demonstrator and shepherd activities and respective individual Science Demonstrator progress has been monitored, and interim and final reports have been collected from Science Demonstrators.

Also dedicated Science Demonstrator sessions have been specifically prepared, organized and chaired during two Stakeholders meetings in Brussels (2017) and Vienna (2018).

The evaluation of the Science Demonstrator experiences and feedback is detailed in chapter 5 with conclusions in chapter 6. Essential findings and requirements include:

- Close and continuous collaboration between EOSC and the active research communities:

  The research communities wish to ensure that their needs are met by the EOSC and that they are adapted to as they evolve during the project. They also wish to ensure that they have a deep understanding of the EOSC and its future direction.

- Easy access to large scale resources and services:

  The research communities request that resources and services be well described, including functional descriptions, Quality of Services (QoS), Service Level Agreements (SLAs) and Technical Readiness Levels (TRLs). There is also a call to ensure that resource negotiations are completed in a timely manner.

- High-level services for research end users:

  The majority of research end users find low level services (such as managing VMs in a cloud) to be daunting. High level services developed by community Research Software Engineers and/or the EOSC should be deployed as a higher layer of abstraction, providing targeted services and solutions to the end users.

- Project enabling and support through IT and domain aware experts (EOSC shepherds and community Research Software Engineers):

  Within the EOSCpilot project the role of the shepherd in EOSC and their research community counterpart, the Research Software Engineer, lead to a very efficient project enabling phase. Many benefits can be gained from the continuation of this model.

- Support for cross infrastructure workflows:

  The need to run workflows which seamlessly access resources in different e-infrastructures was highlighted by numerous Science Demonstrators. Several aspects need to be addressed to enable this, including: AAI, Orchestrations tools and services, workflow languages and service interoperability.

- Addressing FAIR data principles on a community by community basis:

The different research communities face different challenges when aiming to make their data FAIR. Issues such as copyright and sensitive data being major obstacles for some. This leads to a need for a case by case assessment for FAIR data and a need for a range of services to ensure high levels of data FAIRness can be achieved in all cases.

2. INTRODUCTION

Science Demonstrators are early adopters of EOSC, selected from across a range of science areas.

Science Demonstrators dealing with societal challenges (i.e. from Life Sciences, Energy, Climate and Material Science) requiring access to and re-use of data and knowledge already developed by European Institutions are, of course, of particular interest within EOSC.

The purpose of this document is to provide the consolidated evaluation of the WP4 Science Demonstrator activities, outcomes and recommendations towards the goal of fostering the implementation of EOSC capable of catering for the needs of Open Science in Europe.

The deliverable is organized in six main sections and two annexes. Section 2 gives an introductory overview, Section 3 presents the 15 selected Science Demonstrators in EOSCpilot, Section 4 describes how the work and interactions with the Science Demonstrators was carried out, Section 5 contains the evaluations of the Science Demonstrator experiences and feedback as the major WP4 project result outcome, Section 6 provides conclusions and recommendations.

Annex A contains a glossary, while Annex B contains the final reports of the Science Demonstrators.

Through this structure, the essential activities from the three tasks in WP 4 Science Demonstrators are covered: Selection, coordination and evaluation of the Science Demonstrators (Task 4.1), Technical Requirements from the Science Demonstrators (Task 4.2), and Shepherding the Science Demonstrator (Task 4.3).

3. SELECTED SCIENCE DEMONSTRATORS IN EOSCpilot

The science areas targeted in EOSCpilot have been covered with the following representations:

First five Science Demonstrators (pre-selected)

Environmental & Earth Sciences – **ENVRI:** Radiative Forcing Integration to enable comparable data access across multiple research communities by working on data integration and harmonised access.
High Energy Physics – **DPHEP/WLCG**: large-scale, long-term data preservation and re-use of physics data through the deployment of HEP data in the EOSC open to other research communities.
Social Sciences – **TEXTCROWD**: Collaborative semantic enrichment of text-based datasets by developing new software to enable a semantic enrichment of text sources and make it available on the EOSC.
Life Sciences – **Pan-Cancer:** Analyses and Cloud Computing within the EOSC to accelerate genomic analysis on the EOSC and reuse solutions in other areas (e.g. for cardiovascular & neuro-degenerative diseases).
Physics (**including materials science**) – **Photon-neutron:** Improve the community's computing facilities by creating a virtual platform for all users (e.g., for users with no storage facilities at their home institutes).

Second five Science Demonstrators:

Energy Research – PROMINENCE: HPCaaS for Fusion - Access to HPC class nodes for the Fusion Research community through a cloud interface.
Earth Sciences – EPOS/VERCE: Virtual Earthquake and Computational Earth Science e-science environment in Europe.
Life Sciences / Genome Research – Life Sciences Datasets: Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability.
Life Sciences / Structural Biology – CryoEM Workflows: Linking distributed data and data analysis resources as workflows in Structural Biology with Cryo Electron Microscopy: Interoperability and reuse.
Physical Sciences / Astronomy – LOFAR Data: Easy access to LOFAR data and knowledge extraction through Open Science Cloud.

Third five Science Demonstrators:

Generic Technology – Frictionless Data Exchange: Across Research Data, Software and Scientific Paper Repositories.
Life Sciences / Genome Research – Bioimaging: Mining a large image repository to extract new biological knowledge about human gene function.
Astro Sciences – VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics.
Earth Sciences / Hydrology: Switching on the EOSC for Reproducible Computational Hydrology by FAIR-ifying eWaterCycle and SWITCH-ON.
Social Sciences and Humanities – VisualMedia: a service for sharing and visualizing visual media files on the web.

The selected Science Demonstrators (according to the selection process described in Deliverable D4.2) received central funding of 12 PMs for their typically 1 year lasting engagement with their various tasks (with the exception DPHEP by CERN which provided its contribution in kind) to help improving the service portfolio and processes essential for the EOSC. The engagement model of Science Demonstrators has been described in Deliverable D4.1.

4. WORK WITH THE SCIENCE DEMONSTRATORS

In total fifteen Science Demonstrators have been engaged with EOSCpilot during the two years duration of the project. The first five (pre-selected) carried out their engagement in the first project year (a few with a few months extension), the second five Science Demonstrators started their work in project month 7 until project month 18 (a few with a few months extension), and the third five Science Demonstrators started their work in project month 12, until project month 23.

With the fifteen Science Demonstrators (selected from over 80 applications in total) all major science fields targeted by EOSCpilot have been covered.

The necessary information flow and communication between Science Demonstrators and the project has been ensured by applying the engagement model devised in D4.1.

Accordingly, each Science Demonstrator shall:

- Work with their assigned shepherds to engage with the EOSCpilot in establishing the technical use cases, software tools, data models and scientific workflows that they use.
- Commit to adopting and using the EOSCpilot services (in WP5 and WP6) as they become available to meet the Demonstrator's technical requirements and providing feedback on the use and suitability of these services.
- Engage with the other WPs within the project through domain specific experts.

For each Science Demonstrator, a suitable shepherd has been assigned to act as the sole contact point for the Science Demonstrator during the engagement process. Additionally, a deputy-shepherd supported the main shepherd in their activities. Shepherd and deputy-shepherd have been selected among the staff belonging to the Partners taking part in Task 4.3 of WP4.

The shepherds closely interacted with their Science Demonstrators independently for carrying out the work plans and help with the identification of issues and solutions. Shepherds guided and moderated the interactions between Science Demonstrators and WP5 and WP6 and took part in respective meetings organized by WP5 and WP6.

The concept of shepherding has proven to be essential for the enabling of Science Demonstrators to establish and execute their domain specific work plans through setting up, testing of and engaging with EOSC type services.

Regular monthly meeting have been carried out with the Science Demonstrators and their shepherds to guide and monitor the progress of the activities according to the work plans, to discuss problems and provide support towards solutions.

During the first year, the EOSCpilot project organized the First Stakeholder Event in November 2017 (Brussels) where some of the Science Demonstrators provided public feedback and recommendations. A Second Stakeholder Event was organized in November 2018 (Vienna) where all Science Demonstrators shared their experiences and results, lessons learnt and recommendations, integrated in general panel discussions, with a large audience with high EOSC related relevance.

5. EVALUATION OF SCIENCE DEMONSTRATOR OUTCOMES

The Science Demonstrator achievements and outcomes have been evaluated, and the reporting is grouped into chapters with Technical Challenges, Policy Challenges, Cultural Challenges, Functional and Non-Functional Requirements. Feedback from each of the Science demonstrators was considered when compiling this section. The Interim and Final reports, as well as presentations at the Stakeholder Forums and the ensuing discussions, have helped us to highlight many areas of commonality with respect to the needs of researchers and the challenges they have faced.

5.1. Technical Challenges and Functional Requirements

Each Science demonstrator set out to realize a domain specific use-case by using EOSC services and resources. In this section, we will present an overview of the challenges which were faced by the demonstrators. We have broken this down into two main areas of interest, namely, Services and Resources.

5.1.1. Services

Possibly the largest common requirement was the wish to have tools and services to aid with orchestration and the deployment of Cross-Infrastructure workflows/pipelines. Once this requirement is broken down it is clear that there are several underlying components which need to be fulfilled to enable this. Many of the demonstrators have taken a similar path to achieve this goal.

The core components of these solutions are:

- Orchestration tools/services.
- Workflow Languages and tools.
- Containers.
- AAI.
- Interoperability.

Orchestration: While some demonstrators developed their own orchestration services others took advantage of solutions provided by EOSC. In all cases, a solution was found to allow the SDs to deploy workflows across EOSC resources. However, there does appear to be room for improvement both at the technical level and at the level of documentation and support. It is plain from the responses of the SDs that EOSC services and tools for workflows are required but there will always remain a need to allow communities themselves to layer their Orchestration tools upon EOSC resources. Both models are needed; Community "*Bring your own orchestration*" and EOSC "*Orchestration out of the box*".

Workflow Languages: Several benefits of using workflow languages were highlighted by the Science Demonstrators, including standardization, reproducibility and ease of use. Two main solutions which were explored are the Common Workflow Language (CWL) and NextFlow. In addition some SDs defined their own workflows using JSON or community specific solutions.

Containers: The current trend of using containers to house software and in some cases data was strongly present in the EOSCpilot Science Demonstrators. Containers are seen as an ideal solution to allow portable and reproducible workflows to be executed across varying resources (e-infrastructure). Both Docker and Singularity have been identified as suitable technologies (with a slight preference for Singularity). Support for containers in workflows and also for container repositories is one of the requests coming from the SD communities.

AAI: In general the Science Demonstrators took a pragmatic approach to AAI as it was clear that at the beginning of the project no common (cross-infrastructure) AAI solution was available. There is, however, a need for solutions to allow workflows to access services from different infrastructure

providers in a seamless manner. In addition, communities have stressed that this needs to be as easy as possible for the end users. Researchers will simply not use services if there is a steep learning curve just to access them.

Interoperability: The ability for services to interoperate is seen as a 'must-have' for workflows to be able to function in a cross-infrastructure environment. Although this is strongly linked to the issue of AAI there are also other aspects which were brought up by the Science Demonstrators. The use of standards, as well as clear and well defined APIs to access services is equally important. Additionally, it has become clear that solution providers within research communities wish to be able to pick off-the-shelf services without being forced to use other dependent services. In short, communities prefer a loose-coupling approach as opposed to strongly integrated services.

In addition to the components needed to enable workflows the science demonstrators identified several other technical challenges. These include data services and solutions, Jupyter notebooks and high-level services.

Data Services and Solutions: In general the Science Demonstrators were performed as a proof of concept and large-scale data resources were not requested. However, some stand-out cases do exist and these are highlighted below. Several demonstrators made use of data services or were able to highlight areas where services need to be deployed and/or improved. These services include the need to:

- Operate a Long Term Trusted Digital Repository (TDR).
- Address FAIR principles on a community-by-community basis and enable anonymous data access when requested.
- Provide data transfer services.

A request for high performance filesystems within the cloud environments was made. OneData was solution which was tested the most by the Science Demonstrators. Experience with using OneData was initially poor but improved through the lifetime of the EOSCpilot project.

The ability to stage data in a long-term archive as part of a workflow, along with data publication services and intermediate repositories where data and associated metadata can be stored (including workflow description), were also highlighted.

The core data services and solutions needing focus are therefore:

- Long-Term TDR (more clarity about the service and testing needed).
- Improvements to OneData (both functionality and improved stability).
- Discuss FAIR data on a community or case by case basis.
- High-performance filesystems (NFS like) for cloud and container jobs.
- Data Publication services (More evaluations needed, engagement with/from communities).
- Services for reliable large-scale data transfers (including networks between data centers).

Jupyter notebooks: Many science demonstrators used Jupyter notebooks and requests have been made for better integration of notebooks and EOSC services. For example, the ability to bind storage into notebooks. Remote notebook access to enable visualization and interactive analysis close to the data is also good.

High-Level Services: The service catalogs currently offered were considered by many to be too low-level to be presented to end users. The communities are interested in higher-level services which in some cases they are developing themselves. For example, VREs (D4Science).

Others areas: Visualization and virtual networks were mentioned by some Science Demonstrators as areas they would like to see addressed. Visualization tools would allow researchers to visualize remote data in real time and also perform interactive analysis (Jupyter notebooks have been used in such use

cases but other tools may be requested by communities). Virtual networks would allow researchers to configure remote resources at different sites in such a way that they communicated over a secure/dedicated network. This could lead to simplified configuration with respect to security and also firewall port openings.

5.1.2. Resources

Easy and straightforward access to compute, storage and other resources is an important factor for EOSC users. EOSC resources need to be well described and easy to access. This is equally important when encouraging new users groups to begin using EOSC and when facilitating the scaling of existing users. Through the engagement and feedback from the Science Demonstrators several areas of focus have been identified, namely:

- Ease of access.
- Scale.
- Stability.
- Resources for sensitive data.
- Long Term Digital Archives.

Ease of access: In the initial phase of the project access to resources was often time consuming and lead to several delays in projects and in some cases projects needed to be down scaled. This is partially due to the proof-of-concept nature of the EOSCpilot project and improved during the project lifetime. However, problems still remain when projects request resources. The resource request process needs to be simple and transparent to the communities requesting resources. A two-track approach may also be of benefit. One track supporting individual users (likely to be small scale users) and a second track supporting whole projects (likely to be larger in scale).

Scale & Stability: Although the pilot projects have focused more on proof-of-concept rather than large-scale deployments, several Science Demonstrators attempted to scale their workflows into regions of hundreds of VMs. This was problematic both due to the need to negotiate with resource providers, in some cases leading to several months of delays, and also functionally with issues of stability being encountered as well as scaling issues. The Science Demonstrators stated clearly that EOSC resources must be prepared to operate at much larger scales if they are to be seen as attractive by research communities.

Sensitive Data: Resources that are suitable for use with sensitive data, both compute and storage resources were not found. This either identifies a gap in the service and resource catalog or highlights a need to better advertise resources which are appropriate for sensitive data.

Long Term Digital Archives: Providing resources for a long term digital archive means planning for a large-scale and long-term deployment. The lifetime of such a service is to be measured in decades and needs to be set out clearly in SLAs and QoS descriptions. Moreover, the services need to be described in detail and the roles of the service users (communities) and service providers need to be well defined. It's clear that currently the interested community and the resource provider need to discuss more to agree on a clear set of responsibilities and roles. Through the EOSCpilot a resource provider was identified to provide a trusted digital repository but the search took eight months and the supported Science Demonstrator was coming to an end as real negotiations started. These discussions between community and provider need to be continued to ensure a service can be realized that suits community needs.

5.2. Policy and Governance Challenges and Non-Functional Requirements

The pragmatic, bottom up approach of the EOSC Science Demonstrators meant that not many 'political' challenges were encountered.

The user communities would clearly like to have a strong influence on EOSC with a voice equal to that of the infrastructure and service providers. As the end users, they wish to be involved in decision making and steering to ensure that the EOSC solutions are appropriate for their communities. They want a strong voice to allow them to provide feedback about the existing services and resources as well as to identify needs and priorities for future development.

Management of sensitive and/or copyright data in a cloud environment is one area that needs to be addressed by EOSC and the user communities together. This is both an issue of services and policies.

Policies need to be defined which define how community services can be integrated into the EOSC and possibly advertised as EOSC fringe services. A clear procedure needs to be defined and a clear division of responsibilities is needed for services which are run by a community in the EOSC environment. In general, communities have called for clarity and openness regarding EOSC policies.

Non-functional requirements include areas such as the interaction between the communities and the EOSC, service stability, the need for training and a need for better service descriptions and documentation.

We will first focus on the following areas:

- Requesting resources and finding services.
- Support - community solution developers.
- Service stability (clear SLAs and TRLs for services).
- Stronger collaboration between EOSC and communities (improved clarity).

Resources and Services: In general resource negotiation was seen as far too time consuming, and identifying services was seen as challenging due to a lack of description of the services and/or lack of adequate advertising of the services. Improvements were seen as the project progressed; resource negotiation became easier, the services became more visible and service catalogs were produced. However, it is clear that communities feel more clarity is needed in both areas and clear SLAs need to be advertised.

Support: The role of the shepherd, as a single point of contact and also as someone who could aid with designing the architecture of a solution on EOSC, was seen as essential by the Science Demonstrators. The communities have highlighted the need for high level services to be developed for EOSC. These services are often community specific and often layer on top of several core EOSC services. Close collaboration is essential when developing them.

Stability of services and resources: Several Science Demonstrators called for improvements in stability of both the services and resources. A stable infrastructure is a must if we want to encourage user communities to adopt solutions within the EOSC. The Science Demonstrators highlighted the need to have clarity when services and resources are offered through the EOSC. This would require SLAs to be advertised from service and resource providers and also for the services themselves to state their Technical Readiness Level (TRL). Many communities are happy to participate in the early uptake of services with low TRLs, this helps them evaluate new solutions. However, they would like to be able to build their core solutions on services with high TRLs, thus ensuring stability for their end users.

Collaboration and interaction: In general more clarity was requested from the communities when they communicate with EOSC. The role of the shepherd was very much appreciated for the individual

projects but for higher level discussions the communities see a need for their voice to be heard directly (and not via a proxy).

5.3. Cultural Challenges

Many Science Demonstrators highlighted the need to hide the complexity of EOSC from the end users by building higher-level services on top of EOSC base-level services. These high-level services would be presented to the community end users. The base-level EOSC services would be used by EOSC and communities alike to build the high-level services. The research end users should see the absolute minimum level of IT complexity.

Many Science Demonstrators also called for a split of the community users into two main categories, Service enablers and End users. Once this distinction has been made the cultural challenges being faced can be addressed on two fronts.

1. Service enablers:

Service enablers are community experts who can build community specific, high-level, services on top of the EOSC services. To aid these service enablers in developing these high-level services, specific training events and documentation should be prepared which is targeted at them. They would also benefit strongly from close engagement with EOSC service experts both when designing the architecture of solutions and debugging problems. These service enablers can be seen as the bridge between the research communities and the EOSC. Their counterparts, within the EOSCpilot project, were the shepherds and this proved to be a very effective model.

2. End users:

A lack of end user IT skills was identified by several demonstrators as a challenge to EOSC usage. This can be address in numerous ways for instance ensuring the end users only see high-level services but also through training and improved documentation. Best practices and example use-cases would also help highlight how EOSC resources and services can be used by research end users. Some Science Demonstrators highlighted this as a more general issue. The current developments in IT with distributed/cloud computing and container solutions are occurring at a very high pace. These developments are very different to the 'tried and tested' systems which many users have experience with. For example, HPC/HTC systems with software deployed by local administrators. Training and documentation are needed in general to address these development trends in IT and make them accessible to research end users and have them seen as an opportunity rather than an obstacle.

In addition to the need for user-friendly, high-level, services, some demonstrators simply stated that performing the use-case on their own 'in house' resources would have been much simpler and quicker. They highlight the availability of face-to-face support and quick turn around when problems occur. This highlights a need to address areas such as communication and the speed at which issues are addressed. It also highlights a need to better explain the benefits of the EOSC and to put better support processes in place.

5.4. FAIR Principles and Open Science

FAIR data is one of the core principles upon which the EOSC is founded. However, the acceptance and pursuit of FAIR data principles is still not very high in some communities. This is partially due to some differences in opinion about what FAIR data actually means. It is now becoming clear that there are degrees of fairness and the EOSC will need to address FAIR data principles on a community by community basis. A generic, 'one size fits all' collection of services cannot address the concerns of all communities, and the EOSC will need to engage with communities to define the services which meet their needs and make their data as FAIR as possible. It is equally clear that services to enable FAIR data

need to be simple for the end users and that the benefits of FAIR data need to be highlighted and exploited.

Science Demonstrators followed Open Science principles as far as possible, providing transparent and accessible knowledge to both their communities and EOSC in general - including data generated by the Science Demonstrators, services and tools developed, and lessons learnt.

6. CONCLUSIONS

The concept of Science Demonstrator activities in EOSCpilot was very successful. Science Demonstrators from all major scientific disciplines evaluated numerous use cases. Their experiences and recommendations are detailed in chapter 5.

Accordingly, there is a strong need to ensure that the EOSC works closely and continuously with the research communities. Through this collaboration the EOSC will remain aware of community needs, and communities will stay aware of EOSC developments and solutions.

The concept of the shepherd, as introduced in EOSCpilot, who acts as a first point of contact and aids with designing solution architectures in EOSC, has proven to be invaluable and should be continued in EOSC. A well working interface between the research communities and the EOSC is probably the most important non-functional aspect that should be focused upon.

In summary, the Science Demonstrators have provided an essential contribution to the objectives of EOSCpilot by providing relevant input and recommendations for the further development of EOSC to meet the functional and non-functional needs of the science communities, researchers and end users.

ANNEX A. GLOSSARY

Term	Explanation
Cloud computing	The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
Consortium	The EOSCpilot project consortium.
Data analysis	The process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.
Data integration	To combine data from disparate sources into meaningful and valuable information.
Data interoperability	To work with other data systems and exchange information while preserving the meaning and relationships of the data exchanged.
Data management	Development and execution of architectures, policies, practices and procedures in order to manage the information lifecycle needs in an effective manner.
EOSC	European Open Science Cloud.
FAIR	Findable, Accessible, Interoperable, Reusable.
Grid computing	A distributed computing architecture that combines computer resources from various domains to reach a main objective. In grid computing, the computers on the network can be orchestrated to work on individual tasks concurrently together, thus functioning as a much larger computer.
HPC	High-Performance Computing. Implies the use of parallel processing for running advanced application programs efficiently, reliably and quickly.
HTC	High-Throughput Computing. Implies the use of many computing resources over long periods of time to accomplish a computational task.

Term	Explanation
Network resources	Forms of data, information and hardware devices that can be accessed by a group of computers through the use of a shared connection.
Open Science	The movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.
Science Demonstrators	High-profile pilots that integrate services and infrastructures to show the usefulness of the EOSC Services and to drive the further development of EOSC.
Science Demonstrator Representative	Contact person from a specific Science Demonstrator. They will work together with the Shepherd assigned to the Science Demonstrator in order to develop the proposed project.
Shepherd	Staff who supports the main contact of a Science Demonstrator in order to facilitate the engagement with the EOSCpilot project in establishing their technical use case, software tools, data models and scientific workflow to be used.

ANNEX B. FINAL REPORTS OF SCIENCE DEMONSTRATORS

This annex contains the final reports of the Science Demonstrators in EOSCpilot.

EOSCpilot: Science Demonstrator Final Report	
Date	**2018-Nov-09**
Science Demonstrator Title	ENVRI Radiative Forcing Integration
Representative name, affiliation and email from proposing organisation(s)	Werner L. Kutsch ICOS ERIC, werner.kutsch@icos-ri.eu Alex Vermeulen ICOS ERIC, alex.vermeulen@icos-ri.eu Ville Kasurinen ICOS ERIC, ville.kasurinen@icos-ri.eu Stephan Kindermann IS-ENES2, kindermann@dkrz.de Sylvie Joussaume IS-ENES2, sylvie.joussaume@lsce.ipsl.fr Sébastien Denvil IS-ENES2, sebastien.denvil@ipsl.jussieu.fr Francesca Guglielmo IS-ENES2, francesca.guglielmo@lsce.ipsl.fr
Main Shepherd name, affiliation and email	Giuseppe La Rocca, EGI Foundation, giuseppe.larocca@egi.eu
Secondary Shepherd name, affiliation and email	
General part	
Achievements so far (> 200 words)	The refined ERFI scientific use case addresses the forcing of land surface/ecosystem models (instead of the originally planned investigation of radiative forcing). IS-ENES climate models data archived on the ENES Data Infrastructure (DKRZ node of Earth System Grid Federation) were transferred to a OneData EGI node. The VM

	was also configured to fetch data-sets from the original data repository into the OneData volume. ICOS accessed and browsed through the 1-2 TB IS-ENES data and extracted data from historical and rcp45 experiments to build input files for a land surface model (LPJ-GUESS). LPJ-GUESS simulated water and carbon fluxes were compared to in-situ measurements from ecosystem flux towers (Fluxnet 2015 dataset). Simulations were carried out using monthly time resolution. This pilot tested interoperability between institution hosting data (IS-ENES) and institutions responsible for simulations (ICOS, in this case University of Helsinki) and operate in a virtual platform provided by EGI. The EGI Open Data Platform allows integration of various data repositories in distributed infrastructure providing a technical solution where the user can access data in repository and directly create a subset of it in cloud without a need to download source files locally.
Problems encountered	**General aspects:** After an initial delay, and a complete redefinition of the work plan, the actual activity of this pilot was picked up in Aug. 2018. Due to this initial delay the original work plan was iteratively refined. According to the new work plan, the ERFI scientific use case now addresses the transfer of data to produce forcing fields of land surface/ecosystem models. Climate models data (historical and predicted future climate data by IS-ENES) archived on the ENES Data Infrastructure have been transferred to the ICOS Carbon Portal for forcing of land surface/ecosystem models for modelling carbon cycle and greenhouse gas exchange (ICOS). **Technical problems:** CSC-grid certificates can be applied through DigiCert SSO service. However, the user interface contains a bug - certificate generation will not succeed if you write to a textbox asking the purpose of the certificate - leaving it empty will produce a certificate successfully. Personal certificate generated using DigiCert SSO could not be used to access EGI VMOps dashboard. At the end EGI SSO account was created and used to create a VMs topology with the EGI VMOps dashboard. VM created 1.8.2018 was un-deployed 1.11.2018. According to the CESNET-MetaCloud provider's expiration policy, a warning message was sent three weeks before termination to the owner ("appdb-support@iasa.gr") without informing the user of the VM. However, it never reached the user of the VM. **Onedata problems:** An initial challenge was to make multi-petabyte climate model data archives hosted from EGSF hosted by DKRZ accessible to ICOS RI. Due

	to some security policies, it was not possible to configure the IS-ENES data repositories as oneproviders for the EGI DataHub. The connection to the DataHub was tested first time in the summer 2017. The provided solution crashed after loading a few hundred of files in October 2017. After this Cyfronet only succeeded to finalize the configuration of the block space in DataHub in January 2018. But even during the spring 2018, repeatedly, after uploading the files the repository got lost due to technical problems. At the end, DataHub was finally set up and operational by October 2018. But still problems were detected, related to reading and writing rights of the shared disk, which delayed the uploading of files during the autumn 2018. File processing in the DataHub was very slow in the beginning, but got better when technical details related to mounting of the block space was rephrased. The connection between the block space and the oneclient installed in the VM broke down several times, since the client has no direct access to the storage where IS-ENES datasets are available. The connection is limited by the network speed between the VM and the Oneprovider. To have better performance we either had to deploy the Oneprovider closer to the VM where the analysis is performed, or move the VM closer to the Onedata hosting the datasets. Unfortunately, both options were not feasible in the due time. The closest cloud provider was down for several weeks. As a consequence, the creation of monthly averages from NetCDF files was time consuming. **Documentation:** LPJ-GUESS documentation could support the user better. Information and requirements related to different model setups are scattered. Successful runs require personal assistance from persons who are familiar with LPJ-GUESS model. **Modelling**: In Earth System Model output data variable definitions are not homogenous, for example amount of cloud cover is model dependent. Variable names in LPJ-GUESS output data and flux tower data are not identical. Therefore, comparison of LPJ-GUESS simulations to Fluxnet 2015 data set requires manual work before LPJ-GUESS outputs can be compared to measured fulxes. For example, total evapotranspiration from LPJ-GUESS must be calculated as a sum of soil evaporation, transpiration and ET from interception, while Fluxnet 2015 dataset reports total evapotranspiration. Although gap-filled flux tower data was used to compare LPJ-GUESS simulations to measured fluxes , variable naming and variable availability cause problems for the analysis. Although data set should use universal variable names, all station specific data sets does not contain all expected variables. In practice, the only solution is to investigate flux tower files site by site and define a subset of variables that is suitable for all stations.

Data management and handling of sensitive data, with reference to plan.	Not applicable.
List of outreach activities	IS-ENES: presentation of the prototype at ENES scoping meetings and in the frame of the ENES Data Task Force meetings, discussions involving the Earth's Climate System Modelling community. Oral and poster presentation at the EOSC Stakeholder Forum (Brussels, November 2017).
Specific Feedback on:	
Technical challenges and issues encountered	Many. General approach: Two options were considered in the beginning: 1) Use existing Onedata installations (e.g. hosted by EGI) and build a generic data export / data synchronisation pipeline. 2) Set up a Onedata server at DKRZ (or IPSL) with access to (parts of) our archive and integrate this in a Onedata federation used in the demonstrator. Option 2) was discarded for problems in quantifying the resources available. Metadata: One of the approaches initially discussed was to use EUDAT B2Find for the metadata part, yet the situation of ENES data exposure via B2Find was at a different level of granularity and for long term archived data - thus a complete new IS-ENES - to B2Find metadata harvesting pipeline would have to be deployed - including a translator component.

	Implementation: Initial problems with the Onedata candidate releases to implement a transparent access and sharing of datasets. Low network performance to access IS-ENES datasets hosted in Onedata. Lack of stability of the connection, up to complete crashes of virtual machines and data loss.
Proposed measures /suggestions to mitigate technical issues	**Sharing of datasets** From the technical point of view, after careful consideration of real and potential technical issues (see above), the solution proposed and agreed upon was to transfer IS-ENES climate model data hosted on the DKRZ node of the Earth system Grid Federation (ESGF) on a OneData EGI node (Virtual Machine) via http or gridftp (exploiting Synda, a command line tool to search and download files from the ESGF archive). **Hosting of datasets** After IS-ENES estimated necessary disk space to host data as needed by ICOS (daily frequency meteorological forcing), CYFRONET configured the disk space in the EGI DataHub service to allocate 2TB in the Ceph installation. The VM running in one of the EGI providers was also configured with additional software and libraries to fetch data-sets from the original data repository into the Onedata volume, and access to this datasets using the oneclient. Initials bugs in the oneprovider client have been fixed in release rc.11. The access to the EGI FedCloud infrastructure and the configuration of the block space in the EGI DataHub was enabled by WP5. **Performance** To improve the overall performance it has been decided to move the computation close to the Onedata provider where the IS-ENES datasets are physically stored. Unfortunately the only available cloud provider in Poland scheduled an upgrading of the IaaS infrastructure and was not available. For this reason researchers involved in this pilot have been forced to reduce the number of parallel runs.
Political challenges encountered	Nothing to report

Proposed measures /suggestions to mitigate political issues	Nothing to report
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	Moving and adapting services from custom or property platforms to other infrastructure always requires additional effort from both sides, users and system admins. Members involved in this Science Demonstrator have different backgrounds, and not all of them have the necessary skills to manage the interaction with the services provided by EOSC to facilitate the implementation of the pilot.
Proposed measures /suggestions to mitigate cultural issues	To mitigate the cultural challenges reported above, users have been invited to access the EGI FedCloud infrastructure using the EGI VMOps dashboard. The dashboard allows to deploy VM topologies with an user-friendly interface. With the wizard, the user with few clicks can define the technical details of the VM to be deployed in one of the cloud providers of the federation. Moreover, to facilitate the knowledge transfer, in one of the WP5.4/WP6.3 monthly meetings, invited Cyfronet to report about how the Onedata software stack can be used by research communities to implement a transparent access to datasets across different data providers.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	The interaction with the cloud and service providers was liaised by WP5.

Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The configuration of the VM with oneclient to access the block space exported by the Onedata provider was straightforward. The documentation is very detailed and organized to target different profiles: end-users and system admins. During the project, the solution to implement "transparent data access", based on Onedata has significantly improved its functionalities and performance, but the limited network performance delayed the analysis of the datasets and the completion of the work plan.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	The Onedata solution does not support any metadata transfer besides the names of the files. Therefore interoperability depends completely on the user and there is no built-in support for automated workflows. Technical problems prevented that actual results have been obtained and testing of interoperability issues.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	An average scientific users involved in this pilot project (responsible for model runs associated to pilot) had a little of experience of cloud technologies VMs. For such users a basic cloud training event would be very useful, although the technical support was available when needed. Additional, some specific crash-courses and/or webinars on how to manage Docker container, compile and build required libraries (e.g.: CDO and NCO) on cloud computing environment, and focused training on porting scientific applications on cloud-based infrastructure will be also very helpful.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	The policy areas of most relevance for this demonstrator and the science communities that should be addressed to enable maximization of interoperability of services and seamless flow of data are: Policies for infrastructures and services. In this framework, work to go beyond community specific standard could be performed at different levels favoring in the first place communication among RIs, and mapping requirements considering the different levels of advancement, the history and the target community of each RI.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.	This document is a good start, but it avoids clear choices and priorities and puts a lot of responsibilities on the data providers' shoulders.

eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	Performing solutions that work and reactive assistance that shields the user from the complexity of the organisation
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	Improve the network connectivity between the service providers
Other (please specify)	

EOSCpilot: Science Demonstrator Final Report	
Date / Type	2018-02-01
Science Demonstrator Title	http://eoscpilot.eu/science-demos/high-energy-physics (DPHEP)
Representative name, affiliation and email from proposing organisation(s)	Jamie SHIERS, CERN, Jamie.Shiers@cern.ch
Main Shepherd name, affiliation and email	John Kennedy, Max-Planck Computing Centre, jkennedy@rzg.mpg.de
Secondary Shepherd name, affiliation and email	Matthew Viljoen, EGI, matthew.viljoen@egi.eu
General part	
Achievements (> 200 words)	The achievements can only be considered in the context of the goals of this SD. ***OVERVIEW***: Funding agencies today require (FAIR) Data Management Plans, explaining how data acquired or produced will be preserved for re-use, sharing and verification of results. The preservation of data from CERN's Large Hadron Collider poses significant challenges: not least in terms of scale. The purpose of this demonstrator is to show how existing, fully generic services can be combined to meet these needs in a manner that is discipline agnostic, i.e. can be used by others without modification. ***OBJECTIVE***: The high energy physics science demonstrator wants to deploy services that tackle the following functions: 1. Trusted / certified digital repositories where data is referenced by a Persistent Identifier (PID); 2. Scalable "digital library" services where documentation is referenced by a Digital Object Identifer (DOI); 3. A versioning file system to capture and preserve the associated software and needed environment; 4. A virtualised environment that allows the above to run in Cloud, Grid and many other environments.

	TECHNICAL FOCUS: The goal is to use non-discipline specific services combined in a simple and transparent manner (e.g. through PIDs) to build a system capable of storing and preserving Open Data at a scale of 100TB or more. The objective and the simple goal outlined above was not achieved, nor were the "stretch targets" addressed in the SD presentation at the kick-off meeting in Amsterdam in January 2017 addressed. Some limited success was made with the individual services but it was not possible to integrate them into anything approaching a usable service.
Problems encountered	The EOSC Pilot integrates services from 3 well established e-infrastructures: OpenAIRE (e.g. Zenodo), EGI Engage (e.g. CVMFS – already offered by EGI InSPIRE SA3 (led by CERN)) and a Trustworthy Digital Repository (TDR). Equivalent services are used in production by the CERN Open Data Portal, that at the time of writing supports over 1PB of "Open Data" (available via anonymous access over the Internet worldwide). Whilst it was possible to upload a documentation file into the EUDAT B2SHARE test instance (chosen over Zenodo due to the affiliation of the shepherds) and whilst software from the LHC experiments is stored in the RAL CVMFS instance, there were significant delays in finding a site that could act as a TDR for this pilot. There were numerous misunderstandings regarding the scope, duration and scale of the demonstrator; no bulk upload of existing "Open Data" was achieved, anonymous access was not addressed, nor were the 3 services successfully integrated.
Data management and handling of sensitive data, with reference to plan.	"Open Data" is by definition not "sensitive data".
List of outreach activities	e-IRG workshop in Malta in June: https://indico.cern.ch/event/643419/ PASIG workshop in Oxford in September: https://indico.cern.ch/event/664663/

	DPHEP workshop at CERN in March: https://indico.cern.ch/event/588219/
Specific Feedback on:	
Technical challenges and issues encountered	Already described.
Proposed measures /suggestions to mitigate technical issues	The CERN Open Data portal is a production proof that the services needed by this SD can work together. Possibly a meeting with the CERN experts could help solve the problems seen in the EOSC Pilot?
Political challenges encountered	?
Proposed measures /suggestions to mitigate political issues	
Cultural challenges encountered	N/A.

Proposed measures /suggestions to mitigate cultural issues	
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	It was not clear that the objectives from the kickoff event, nor from DPHEP presentations, were sufficiently shared within the project and so many things had to be repeated (and repeated). It became clear that one cannot expect a generic service provider to have the same knowledge of a specific domain as a "domain repository". There is simply a lot of "everyday knowledge" that services providers at CERN and other WLCG sites are aware of (terms such as AOD, ROOT format etc.) This was in part addressed through the series of WLCG Collaboration Workshops held during the WLCG Service Challenge days. It is not possible, realistic or even desirable to repeat such events with e.g. cloud service providers. As a consequence, there needs to be a combination of generic support complemented by domain-specific support. (But this was knew back in EGI InSPIRE SA3, if not before).
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The shepherds were enthusiastic and eagerly sought to find solutions.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	The 3 services mentioned above do not interoperate. But they need to.

Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions?	I am not involved in WP7 so I don't know what they are doing.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	I have not seen the policy document of WP3 (but participated in an interview with Dale Robertson – see below).
Government issues: (WP2): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP2	Does this mean "governance" issues? It should be clear from this report that the "sciences" who are (potential) users of the EOSC services need to have some voice, both in providing feedback on what is provided as well as what is needed / priorities.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	See e-IRG and PASIG presentations. It MUST be possible to integrate services from the different EOSC Pilot service providers, e.g. EGI, EUDAT, OpenAIRE in a simple and straightforward manner. This is described in the "TECHNICAL FOCUS" above and on the EOSC Pilot website since Q1 2017. Can there really be any debate (or confusion) on this point?
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	See e-IRG presentation about KPIs and other metrics.

Other (please specify)	See also report of interview with Dale Robertson (WP3): *Dear all,* *I would like to follow up on Hermann's message below relating to input on policies in the context of EOSCpilot.* *WP3 of EOSCpilot is responsible for conducting a policy landscape review, and for identifying, prioritising and developing recommendations for policy issues to address to support the EOSC. Input to our work from a range of policy experts would be very valuable and I would therefore be very grateful if any policy experts from the Science Demonstrators who are willing to contribute their expertise could please contact me by 24 October. This represents an opportunity to help shape policy for the EOSC and thereby to facilitate the EOSC's establishment and development.* *The attached 1-page document provides more information on the expertise we are seeking and the tasks on which we would like to receive input.*

EOSCpilot: Science Demonstrator Final Report	
Date / Type	2018/01/08 / Final
Science Demonstrator Title	TEXTCROWD
Representative name, affiliation and email from proposing organisation(s)	Franco Niccolucci PIN franco.niccolucci@gmail.com
Main Shepherd name, affiliation and email	Kathrin Beck MPCDF kathrin.beck@mpcdf.mpg.de
Secondary Shepherd name, affiliation and email	Thomas Zastrow MPCDF thomas.zastrow@mpcdf.mpg.de
General part	
Achievements (> 200 words)	Semantic enrichment of text documents. • Sets of similar texts concerning related topics (e.g.: archaeological excavation) • Small size (datasets of KBs or MBs, not TBs or PBs) • Enrichment: metadata creation to improve indexing, discoverability, accessibility and reusability Recognition and automatic annotation of entities and concepts in texts • Machine learning model for natural language processing (NLP) and Named Entities Recognition (NER) Knowledge extraction in semantic format • CIDOC CRM (ISO 21127:2006) encoding in RDF format • Machine readable and consumable data to enhance integration and interoperability within CH domain Cloud based architecture empowered by a modular and extensible framework • Intuitive, web-based user interface • User and access management • Cloud storage of private and shared files • Results available and reusable within various Virtual

	Research Environments (VRE) Interoperability of extracted knowledge • Semantic information in CIDOC CRM format: full integration and interoperability with other Cultural Heritage semantic data FAIR Principles implementation • Metadata to be stored in various registries for easy findability and accessibility (e.g.: ARIADNE Registry, PARTHENOS Registry) • Results ready to be reused within the same environment or consumed by other services in different scenarios • Standard NLP encoding to make data interoperable with other NLP tools Machine-learning capabilities • TEXTCROWD has been boosted and made capable of browsing big online knowledge repositories, educating itself on demand and used for producing semantic metadata ready to be integrated with information coming from different domains, to establish an advanced machine learning scenario.
Problems encountered	No annotated text corpora as training data for machine learning algorithms available in Italian. The TEXTCROWD team has created one. • Manual annotation of 400 pages of Italian archaeology reports Potential semantic overlap between particular Italian archaeological concepts (candidate entities). • A strategy has been developed to study their use in practice within typical Italian archaeological reports and a prioritization mechanism has been identified. No user friendly Cloud-based environment available • Desktop pipeline developed in the first phase of development • Cloud migration into D4Science infrastructure in the second phase of development

Data management and handling of sensitive data, with reference to plan.	TEXTCROWD tool works with open and publicly available textual documents. No plan for management and handling of sensitive data required.
List of outreach activities	During the development of TEXTCROWD other research groups potentially interested, have been contacted concerning the possible extension of TEXTCROWD services to other domains, after it has been successfully implemented in EOSC as a pilot. Among others: • researchers in conservation and restoration, to extend the pilot to their data, in the framework of the E-RIHS Research Infrastructure on Heritage Science. • the Italian Ministry of Culture (MIBACT) for assistance and scientific support by us in the creation/improvement of a national (cloud-based) archaeological data management system. TEXTCROWD would be instrumental to manage and index text reports to be stored in the system. This is still in the design phase and will take years for complete implementation. • other potential research teams in the EOSC pilot framework, to integrate TEXTCROWD in their pilots, most notably in the "Visual Media" pilot. • contacts established with another Science Demonstrator for possible integration. All the above-mentioned extensions may require additional work, especially for the creation of specialized dictionaries, to be carried out with additional resources to be provided externally. It must be noted that in all the meetings TEXTCROWD has shown to be useful for its main purpose, i.e. of being a demonstrator of the importance of EOSC for scientific research in the heritage domain. Specific outreach activities: • Posters presented at the CLARIN yearly plenary meeting in Budapest ( ), at the EOSCpilot Stakeholders Forum in Brussels (28-29 November 2017) and at D4RI Conference in Brussels (30 November). • Dissemination video created. • Demo presented at the EOSCPilot Stakeholders Forum during poster sessions. • Scientific publication planned for 2018.
Specific Feedback on:	

Technical challenges and issues encountered	Identification of NLP tools, and in particular of POS and NER frameworks, suitable to analyze linguistic entities in Italian and able to efficiently recognize and mark named entities relevant for the chosen domain (archaeology). Definition of vocabularies and terminological tools to speed up entity recognition and machine learning process for Italian archaeological reports. Identification of a cloud environment suitable for hosting TEXTCRTOWD's modular infrastructure.
Proposed measures /suggestions to mitigate technical issues	Improve usability of cloud infrastructures in terms of user interfaces and reusability among components, to simplify and speed up installation and deployment processes. Provide advanced controls to monitor the status of each component and of the infrastructure as a whole.
Political challenges encountered	None so far. It will be on the researcher to obtain all the necessary permissions to use texts – it is unclear if using texts for NLP infringes copyright. Apparently there is a different situation from country to country.
Proposed measures /suggestions to mitigate political issues	Meeting with EU-level politicians and officers to discuss the issues. However, they are part of a much larger discussion on IPR, and are probably beyond our reach.
Cultural challenges encountered	Communication with the current cloud providers has been efficient, and an interface team has been easily set up. This has not always been the case with other clouds, where the science community was left alone to manage the pilot development and the tools were rudimentary (or at least they appeared so because of lack of support). As regards end users, researchers seem enthusiastic of the opportunities opened by this technology. We already have requests to use it and possibly we will satisfy some to test the pilot.
Proposed measures /suggestions to mitigate cultural issues	Unnecessary as regards the chosen cloud. For others, improvement of user support may be necessary.

Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	As mentioned above, the quality of the interaction depended on the 'usability' of the cloud selected, on the attitude of the managers and on the availability of software. In one case everything went smoothly at the satisfaction of both parties, science community and infrastructure providers. In other cases, this did not happen as we were offered just storage or a virtual machine, what helped little in the pilot development and was therefore quickly declined..
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The initial deployment of TEXTCROWD was to be performed on a platform providing just a Linux virtual machine. No user interface or specific services were available to facilitate the installation. A deep knowledge of the Linux operating system was required for a proper configuration. We did not check the availability of the required libraries as the above limitations were enough to decline. The D4Science platform already provided the required services, libraries and APIs for TEXTCROWD to quickly become up and running. It also provided user interfaces for installation and execution of services both for testing and production mode.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	Interoperability among components would work more efficiently on a platform that is able to provide facilities supporting a modular approach, for deploying and running on demand the required components and services. D4Science, for instance, provided an interoperable framework already including the GATE engine and the features to use TEXTCROWD together with the OpenNLP and OpenNER web services, thus simplifying the operations of deployment and execution in a controlled environment. Interoperability on this kind of platform is also guaranteed by the presence of the source documents and the output results on the same environment, paramount for the efficiency and the reusability of the information.

Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions?	Nothing to comment on this regard.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	There is a general lack of communication between e-infrastructures and research infrastructures, which eventually did not affect the development of TEXTCROWD because we could find, as already mentioned, the right environment to develop it. It seems that e-infra pretend to know better what is necessary and useful for research, what is of course not the case. Great support and excellent communication was instead established with the TEXTCROWD shepherds, showing that this is possible in general with the right approach.
Government issues: (WP2): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP2	Comments here go beyond the development of TEXTCROWD. It seems that the government structure envisages a wide participation of all stakeholders as listed in the policy document. While it is clearly important that all stakeholders are represented in the Forum and contribute to the strategy, it is not clear how scientific communities can play a privileged role, as is expected in a Science framework (the "S" of EOSC). There is a risk that they are just a voice in the choir, not the soprano.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	As already mentioned, the use of D4Science was relatively easy and effective, because of the availability of the necessary services/libraries, or the possibility of quickly incorporating new ones in the D4Science framework. This should be the main requirement in the transition phase to the EOSC.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	Creating an efficient interface between the science community and the e-infrastructure is a necessity. Our shepherds played very well this role. In the EOSC steady state there is a need of creating permanent structures like that, addressing the needs of individual research communities to overcome communication issues. It is also necessary to preserve (as done in TEXTCROWD) the valuable assets developed by research communities in the process of porting them in the EOSC framework without upsetting their functionality. This may require more flexibility than the one envisaged so far: it is the EOSC that must adapt to the research needs, not vice versa. The scheme adopted for EOSC resources, which distinguishes them into Compliant, Compatible and External, suggests that this division happens now, thus relinquishing at the "External" status resources that are essential for some community. This may lead to disaffection to EOSC. Such scheme is perhaps suitable only in the future: at

	present how will it be established, and who will do it, how EOSC compliance couples with research usefulness?
Other (please specify)	TEXTCROWD is going to be considered as one of the services available in the forthcoming cloud-based version of the ARIADNE portal (portal.ariadne-infrastructure.eu) We plan to implement TEXTCROWD for more languages in the ARIADNE platform. We also plan to test it on a real (large) set of documents and verify the results.

EOSCpilot: Science Demonstrator Final Report	
Date / Type	2018-07-23 / Final Report
Science Demonstrator Title	Pan-Cancer Analysis in the EOSC (Acronym: PanCancer)
Representative name, affiliation and email from proposing organisation(s)	Sergei Iakhnin EMBL, Heidelberg, Germany. iakhnin@embl.de
Main Shepherd name, affiliation and email	Gianni Dalla Torre EMBL-EBI, Hinxton, UK gianni@ebi.ac.uk
Secondary Shepherd name, affiliation and email	Giuseppe La Rocca EGI.eu, Amsterdam, Netherlands giuseppe.larocca@egi.eu
General part	
Achievements (> 200 words)	The Butler scientific workflow framework has been set up and tested at three globally distributed cloud computing environments that are based on the OpenStack platform, these include: the EMBL-EBI Embassy Cloud in the UK, the Cyfronet cloud in Poland, and the ComputeCanada cloud in Canada. The three providers have included the repeated provisioning of over 2500 compute cores, 11 TB of RAM, and ~1.5 PB of storage. Specifically: The computational infrastructure at Cyfronet for the PanCancer tenant comprises: 700 vCPU cores, 2.6TB of RAM, 4.9TB of volume storage and 32 floating IPs. The computational infrastructure at Embassy Cloud for the PanCancer tenant comprises: 1000 vCPU cores, 4TB of RAM, 4 TB of volume storage and 1 PB of NFS storage. The computational infrastructure at ComputeCanada for the PanCancer tenant comprises: 1000 vCPU cores, 3.9TB of

	RAM, 62 TB of volume storage and 200TB of NFS storage. • >400 high coverage whole genome samples (~60 TB of data) from the ICGC pediatric brain cancer cohort were downloaded to the ComputeCanada cloud and, • ~50 TB of public data from the 1000 Genomes project was loaded onto the Cyfronet environment from the EMBL/EBI data servers utilising Cyfronet's Oneprovider software. Butler was used to run a genomic alignment workflow (based on BWA and developed at The Sanger Institute) on >400 samples at ComputeCanada and >400 samples at EMBL/EBI Embassy cloud, with over 100 TB of data processed to date. Cyfronet Oneprovider scalability tests were run on the Deutsche Telekom Openstack cloud (OTC) using 300 VMs with 4 VCPUs each, with each one processing a file served via the Oneprovider client, from disks located at EBI. Proper operation of the infrastructure was ensured by Butler's detailed monitoring and self-healing capabilities.
Problems encountered	One of the key challenges of this Science Demonstrator has been the lack of resource availability for performing large-scale computation required by the cancer genomics workloads. Thus, much of the effort during the first six months of the project has been focused on securing additional computational resources. While the EMBL/EBI Embassy Cloud resources have been available from the project outset, the Cyfronet and ComputeCanada environments only became available in the fall

	of 2017, thus limiting the amount of time that could be spent working within these environments in the early project phases. A second major challenge faced by the project has been the set of stability issues encountered at the Cyfronet environment over the course of the project. These included challenges with provisioning of the entire allocated compute infrastructure, as well as major stability issues with the allocated storage and data staging mechanism (Oneprovider). Although these issues precluded full-scale testing of Butler at Cyfronet within the project timeframe, substantial progress was made to resolve the storage and data staging issues by Cyfronet and these were successfully tested at scale in concert with Butler on the Deutsche Telekom Openstack cloud (OTC). Close interaction with the support teams from Cyfronet helped us achieve a stable and reproducible configuration with satisfactory performance. The OTC tests did not reach the physical limits of the infrastructure, nor was the Oneprovider client a bottleneck. - A third major challenge faced by the project has been a storage throughput issue faced at the ComputeCanada environment. Although dedicated storage was made available for the Science Demonstrator, it was not able to provide the throughput necessary by the cancer genomics workload at the intended scale. The IT team at ComputeCanada was able to eventually reconfigure the storage so that a successful analysis at scale could be performed.
Data management and handling of sensitive data, with reference to plan.	- Data management has been handled in line with the standards specified by the International Cancer Genomics Consortium's Data Access Compliance Office.
List of outreach activities	-
Specific Feedback on:	

Technical challenges and issues encountered One of the key takeaways from this Science Demonstrator is that cloud-scale cancer genomics has quite large resource requirements which are only set to grow. The resources utilized thus far have been in line with the ICGC PCAWG project, which is now five years old. Projecting five years forward, demonstrates that the resource requirements for this field of research are set to grow by one or two orders of magnitude as population-scale sequencing projects are undertaken in Europe and elsewhere. Thus, EOSC infrastructure providers should be prepared to operate at much larger scales than have been demonstrated within the context of the EOSC Pilot project.		
Proposed measures /suggestions to mitigate technical issues	Shared storage has proven to be a key bottleneck in two out of the three clouds used by this Science Demonstrator. The EMBL/EBI Embassy Cloud encountered similar storage issues several years prior (outside the scope of the SD) until high performance Isilon storage was procured that was able to meet the needs of the use case at the required scale. Although this SD had a list of technical requirements that was given to potential infrastructure providers ahead of time, they were not able to meet these requirements initially. In order to avoid the type of project delays that have been encountered by this Science Demonstrator because of storage and other issues, it is recommended that each infrastructure provider issue (and remain beholden to) a standard set of SLAs related to the services that they provide, which are backed by routinely exercised benchmarks. This SLA framework will allow appropriate matching between project resource requirements and infrastructure provider capabilities.	
Political challenges encountered		
Proposed measures /suggestions to mitigate political issues		

Cultural challenges encountered	The key cultural shift between HPC and cloud-based models of scientific computation is related to the fact that the science community takes on many more IT management related responsibilities in the cloud model than under HPC. This provides much more freedom to the science community but also creates a lot more opportunity for errors/failures, which may then get raised to infrastructure providers.
Proposed measures /suggestions to mitigate cultural issues	Successfully navigating this shift towards cloud computing requires establishing a precise set of responsibilities and assignment of ownership between the infrastructure providers and scientists with special attention paid to delineating the differences between the new model and those (HPC, grid, etc.) used in the past. This is especially important in areas such as network and information security (when handling sensitive data) to ensure that responsibilities by all parties are understood and data handling protocols are followed.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	Most communication with infrastructure providers occurs via email where two forms are predominant – either a more informal email conversation between any number of parties related to project concerns, or a more formal message to an email alias backed by a ticket queue for a technical issue. These tend to work quite well for most scenarios, although improvements can be made. Specifically, interactions with the ticket queue can be rather opaque, without a clear understanding of the ticket triage and handling SOP and SLAs by the scientific group pursuing the project. The situation would be vastly improved if a standardized support queue front end was made available to clients where they could review all of their tickets along with assigned priorities and resolution history along with a clear and precise definition of the ticket handling Standard Operating Procedures and the SLAs associated with each step and ticket severity.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The general infrastructure-as-a-service capabilities of all providers were adequate although the stability of resource provisioning could be further improved. The availability of storage can be further improved by providing standard shared NFS-type storage and object storage capabilities. Availability of other PaaS type services like databases and

	queues would be beneficial, as currently projects need to set up and manage their own databases and queues. A catalogue of datasets would be beneficial as the inability to move cloud-scale datasets will largely dictate environment choice by scientific groups in the future.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	Uniform identity and access management across environments remains an unsolved issue.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions?	As already discussed in the Cultural Issues section, the biggest skills gap is related to the additional IT management requirements placed upon the scientific community by operating in the cloud. This includes basic Linux administration, networking, security, and operational management within a distributed computing environment. Putting together educational materials using a platform such as edX would be beneficial in closing some of the gaps. Having a responsibilities matrix (RACI chart) for all project stakeholders accompanied by a set of checklists could also help identify the skill gaps one a project-by-project basis.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	Genomic data handling in the cloud remains an area of fragmentation in Europe. Specific policy to provide uniform data handling guidelines across Europe would be instrumental to move forward.
Government issues: (WP2): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP2	

Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	- Standardized IAM service across providers and services Standardized storage services – block storage, shared-NFS, object storage.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	- Established and publicized SLAs for o infrastructure provisioning, o VM MTTF, o network throughput/latency, storage throughput/latency.
Other (please specify)	

EOSCpilot: Science Demonstrator Pre-Final Report	
Date	**2018-Jun-27**
Science Demonstrator Title	Photon & Neutron Science
Representative name, affiliation and email from proposing organisation(s)	Volker Guelzow Volker.Guelzow@desy.de
Main Shepherd name, affiliation and email	Sune Rastad Bahn Sune.RastadBahn@esss.se
Secondary Shepherd name, affiliation and email	Frank Schluenzen Frank.Schluenzen@desy.de
General part	
Achievements so far (> 200 words)	**Current Status:** The Photon and Neutron SD deployed and tested software used by a large community in structural biology at Free Electron Lasers and Synchrotrons on a local OpenStack cloud platform and on local HPC clusters at DESY, Hamburg. The computationally challenging workflow was examined to identify and establish community specific cloud services and gain insight into technical, organizational, legal issues and interoperability requirements. In the meantime several applications from the Photon and Neutron field have been containerized, deployed and partially profiled. We additionally identified some features which would be helpful in further deployment like tools and service facilitating serverless partitioning of data analysis pipelines. Ultimately, we would like to offer more services to PaN communities leveraging cloud technologies to better abstract the complexities behind deployment and orchestration. Therefore, at the end of the project, we started to scale up the on-site OpenStack infrastructure significantly due to successful proof of concept, which has greatly raised interest and visibility in the user community. This development on the hardware side is complemented by integrating new OpenStack modules into our cloud instance. **Progress made:** We tested two software frameworks that are typically applied for online and offline data analysis and that contain several tools and utilities. OnDA (online data analysis) is a modular and scalable utility designed for fast online feedback during serial X-

	ray diffraction and scattering experiments, based on open source libraries and protocols. The Crystfel framework is used for the technique of Serial Femtosecond Crystallography (SFX) and comprises programs for data processing, simulation and visualization. It is a part of a complex, non-redistributable software stack, which is free to use by academia and non-profit organizations. The containerization of those applications as part of XFEL and CFEL data analysis services with interfaces to data sources, additional software and a variety of other cloud services allowed identifying concrete interoperability requirements which are of importance for a successful integration of the analyses platforms and similar systems into the EOSC. For example, both ESS and DESY were running tests on the local OpenStack infrastructure and on the HPC cluster, in particular in view of automatic provisioning of distributed systems and deployment of applications. The tests have shown that the automation stack applied (Foreman, Puppet, cloud-init, cloud-images) is useful to spawn virtual machines on both types of infrastructures, reducing complexity and ensuring interoperability across platforms. However, this solution extends the range of external software and systems that are integrated into the provisioning process and therefore introduces interoperability challenges when solutions are migrated to other cloud platforms at partner institutes, where other automation platforms can occur. The P&N SD demonstrates an appealing solution by running containers as functions in the OpenWhisk Serverless Platform, underpinned by local clusters on the OpenStack Cloud. This achieves full communication through RESTful interfaces and completely decouples work on and with scientific codes and workflows from platform provisioning and maintenance. Benefits are efficient resource usage and greatly enhanced user friendliness as well as better integration of high availability strategies and automated scaling. To provide data exchange between VMs and external resources outside the OpenStack environment, we used AFS, CVMFS, conventional (NetApp) SMB/NFS-shares and the local/federated Owncloud instance. Configurations were easily integrated into the VM configuration and the deployment successfully automated as described above. Templating common data exchange configurations could help to foster interoperability in the EOSC, We will continue to work on this SD and further examine the integration of middleware solutions for mass storage systems like dCache and iRODS, which is a central interoperability aspect as they represent highly distributed systems and solutions to access data sources from different

	cloud providers. It appears the most suitable way to grant access to large datasets on a petabyte-scale. In terms of network access, inter-network trust and speed and dynamically managed overlay networks for cloud VMs and container deployment in multi-cloud environments, we only have the DESY VPN to give access to resources and are interested in consuming services like the GÉANT Multi-Domain Virtual Private Network (MD-VPN). For container swarm communication this has to provide TCP and UDP connections. So far, we have only deployed container swarms on the local HPC cluster and cloud, but we are interested in an interoperability solution that demonstrates how such swarms can be extended to operate in a multi-domain environment and how they can be migrated between clouds. Interoperable swarm and VM provisioning is an essential backend that enables Jupyter notebook users on Jupyter Hub servers in the cloud to run jobs on the underlying infrastructure.
Problems encountered	
	A substantial fraction of Photon and Neutron applications are free for academic use, but imply restrictions on use like explicit agreement to terms of use, in the worst case prohibiting redistribution of software. This could be either solved by docker/singularity registries restricting user consenting to terms of use, or extended sets of custom attributes served by a attribute authority. The practical implementation and scaling of swarms introduces a strong requirement for interoperable registries, and repositories providing trusted services to distribute container solutions. They should apply access and authorization management, user-attribute and role management, thereby keeping track of acknowledged licenses. To guarantee interoperability of such systems, they should be derived from some EOSC standards, AAI, licenses and certificates. Some problems relate to the complexity of utilizing heterogeneous platforms for generic data analysis services. Jupyterhub based integration could overcome at least some of these issues. Best practices how to integrate resources or generic Jupyter based services would be helpful. The SD is/was in contact with EGI and Eur.XFEL on this aspect. Cooperative use of provisioning infrastructures needs a careful balance between security concerns and usability of tools like git, puppet or foreman. The local services are currently secured but not usable in a truly cooperative or federated manner, which make reuse of configuration templates subject to administrative overhead.

	Local networking issues put some hurdles on federation of resources. The current DMZ concept only allows hosting an entire OpenStack infrastructure inside or outside the DMZ, which makes a balanced use of the infrastructure (e.g. opening idle resources to the outside world) and in particular of petabyte-scale storage operated in the DESY-network at least difficult. Mapping and fixing of user-IDs is a critical issue obstructing fully inter-operable integration and sanity of such storage systems. For this discussion, input from the EOSC if there will be EOSC proxy servers; which network communication will be used within EOSC; outcome of the AARC2 project etc. will be valuable. For interoperable trust between networks and cloud providers, also comparable metrics to control SLAs and QoS are needed.
Data management and handling of sensitive data, with reference to plan.	All data relevant for this demonstrator are published and open access available via the Coherent Xray Imaging Database (cxidb.org). Handling of sensitive data is hence not yet an issue. Promoting an open access policy for the P&N facilities is a slowly progressing activity. Latest adopters of the PaNdata derived policy were HZB and European XFEL. The standard data catalogue for P&N facilities ICAT is continuously being developed (almost exclusively) by colleagues at STFC and is making great progress. ESRF is currently establishing ICAT in production. Frequently, data access to Photon/Neutron datasets is subject to client data management plans, variety of policies and usually restricted until publication. This needs to be addressed in the future for sustainable cloud service in the scientific domain. However, to facilitate data publication and reproducibility of data analysis procedure we are currently working on two smaller developments. For data publication of data processing outcomes, data analysis parameters used and additional meta-data, we are implementing a small python data deposition toolbox harvesting most of the information automatically and rendering the information into json/xml streams for semi-automatic deposition and publication. The other development aims to wrap the workflow for this SD by Jupyter kernels. This will provide a simple way to provide the SD as a webservice – e.g. on the EGI Jupyterhub based platform – and also facilitates to publish and reproduce the workflow. In this context we are currently experimenting with serverless architectures (e.g. openwhisk) to partition the SD and scale out individual processing steps on different platforms.
List of outreach activities	Presentation at the RDA/PaNSIG workshop in Barcelona. Presentation at the Photon Science and European XFEL user meeting (in 01/2018) in Hamburg and Schenefeld. Users report a strong increase of container-based application deployment

	during the last couple of month, and several standard application frameworks like DAWN or BornAgain have been assembled and tested, considerably facilitating cloud deployment of such standard applications. In this context, we continue monthly meetings between ESS and DESY. Meetings with EMBL HH, EMBL HD and Eur. XFEL to promote joint efforts related to EOSC. Participation in and contribution to the EOSCpilot stakeholder event. Participation in the EUDAT conference in Porto (01/2018). Agreed to hold a meeting with DKRZ to discuss cloud strategies and best practices. Will also present the SD at a RDA collocated event for the Photon and Neutron Science communities (03/2018).
Specific Feedback on:	
Technical challenges and issues encountered	**1. Licensing** Substantial part of the applications in the Photon/Neutron science domain is "free to use for academic purposes" but subject to restrictive licensing conditions. To provide services based on such applications requires knowing that the consumer has agreed to licensing terms and is indeed an academic user. This can of course be controlled on a per-service basis. It would however be more convenient, scalable to provide such attributes in a central/federated way integrated into the EOSC ecosystem **2. Security** Cloud resources outside the EOSC core like our local OpenStack instances are presumably often behind firewalls in "demilitarized zones" (DMZ). Consuming such resources in the EOSC context might require access to specific EOSC entry points opening only very specific ports or routes to well-defined IPs. **3. Network** Access to cloud instances across domains and in particular privileged access would greatly benefit from inter-network trust and dynamically managed overlay networks for cloud VMs and container deployment in multi-cloud environments. **4. Graphical Access** Many of the Photon/Neutron applications require visual inspection (with GL support) of intermediate analysis steps. This can be solved on a per-instance basis, and is unproblematic for Jupyter based applications. However, a standardized solution or relay services would be beneficial for users.

	5. Technical Readiness Level The technical readiness level of evolving OpenStack modular environments is very heterogeneous and in particular embedding container orchestration in OpenStack introduces strong challenges. Solutions to such challenges should become readily available at least within the EOSC communities.
Proposed measures /suggestions to mitigate technical issues	**1.** Attribute server **2.** EOSC proxy server and/or recipes for establishing proxyable services. **3.** Support for GÉANT Multi-Domain Virtual Private Network **4.** Best practice/standard/relay services for GL accelerated graphical access to cloud VM instances. **5.** Integration of devtools and knowledge tools to accelerate communication and publish issues, findings and solutions.
Political challenges encountered	**1. Licensing** As mentioned, licensing is one of the major challenges in providing SD specific services. For example the CCP4 package is an integral part of the CrystFEL framework used In this SD. The license agreement for the CCP4 package states "*the Licensee may not distribute any CCP4 Application or any Derived Work based on any CCP4 Application to any third party, or share their use with any third party*". Obeying strictly to the terms makes it very difficult to publish container containing the package, or to provide services within EOSC other than for the local user community. For CCP4 the licensor is actually STFC. **2. Function as a Service (Faas)** FaaS seems an appealing way to partition problems and deploy generic/atomic cloud functions in a highly scalable and elastic way. Though such FaaS could be implemented in a domain-specific way, common applications would benefit from EOSC wide availability of specific functions (e.g. AI based methods). This implies secure and reliable namespace administration.
Proposed measures /suggestions to mitigate political issues	**1. Licensing** a) Urge EOSC members licensing "free-for-academic-use" software for an EOSC-friendly license agreement b) Aim for agreements with software providers (of not entirely open software) to allow redistribution within the EOSC cloud. c) Provide means to clearly separate non-academic from academic service consumption. **2. FaaS** Provide secure namespaces (namespace federation) for FaaS.

Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	Acceptance of open data policies is still not very high in the SD user communities. Implementing and adopting open data policies is hence a slow process. Any additional incentives are helpful. Migration of services from well-established platforms (e.g. HPC, Desktop) to cloud based solutions introduces cultural challenges (demand for change management) on both sides, developers and administrators as well as users. This applies in particular to User Defined Software Stacks (UDSS) via container-shipped deployment.
Proposed measures /suggestions to mitigate cultural issues	Offer benefits for EOSC service consumers adopting open data/science licenses for work derived from EOSC services. (if not already in place). Change Management coupled with skill dissemination and training. Security proof of concept, e.g. mapping user IDs consistently to different cloud and HPC platforms including container environments.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	Photon/Neutron facilities are infrastructure and service providers and at the same time also service consumers. Being an integral part of the science community, facilitates communication, for example by transporting any developments to entire user community through a number of events like user meetings, photon science community meetings; by participating in joint projects and collaborations, schools and hands-on training etc.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	So far we almost exclusively have been using the local services. All SD specificities are easy to transfer to a generic EOSC infrastructure except for FaaS based service deployment. Successfully tested EGI Fedcloud (AppDB, AppDB Dashboard) and EGIs kindly provided their kubernetes deployment recipe for Jupyter for comparison. Reviewed and ensured the Open Access OAI-PMH protocol for metadata harvesting and validated the local service using OpenAire. In the envisaged cooperation between PiCo2 and the Photon-Neutron SD we will assess usage of Geant network technologies in use by PiCo2.

Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	Software stacks used in the SD has been containerized both for singularity and docker and run without issue on any platform supporting one of the two container environments. Concerning data management and access we are investigating interoperability between dCache, iRods and onedata. Apart from issues in the combination of dCache and OneData (which are being / have been resolved by the developers) no major obstacles have been identified yet. Networking issues like opening ports, granting access, placing services inside/outside DMZ still have to be addressed. Identity and access management currently requires DESY account provisioning, which needs to be addressed locally.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	**1. Service enablers** Workshops addressing migration from classical compute infrastructure to cloud environments could greatly leverage the rather steep learning curve. This applies to cloud admin technical training in an evolving highly modular landscape, but to also our user group providing incentives for cloud migration. **2. End users** Most of the Photon/Neutron facility users are of course not interested in any of the cloud technicalities. Common statement is still: all I need is a fast desktop. Integration of EOSC services from e-logbooks to HPC GPU engines in a desktop environment should be the ultimate goal. However, Juypter notebooks are a great tool to overcome many of the obstacles. Educating end users, in particular application developing end users, how to embed cloud services in Jupyter notebooks, would be very valuable.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	Services integration into EOSC and mirroring EOSC services should be addressed.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf)? Please comment if necessary.	The responsibility matrices are too fine-grained, and more confusing than helpful. We would highly welcome low access barriers for EOSC compatible services. Decision making workflow in the stakeholder forum needs clarification.

Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	Performant and secure docker registry is a central requirement. Networking and interoperability would be supported by EOSC trusted Proxies, IP-ranges for networking.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	Organisation of skills dissemination, admin technical training, documentation down to the issues, bug tracking, code base provision.
Other (please specify)	An additional point receiving a lot of attention is the graphical output of containerized, parallel software as well as input via such GUIs (mouse, keyboard, instruments). We need solutions for single users and also for multiple users at the same time, esp. for training and education but also to provide cloud services for distributed teams. We have demonstrated VNC-servers and X11 forwarding to work; the former bearing speed and performance challenges and the latter introducing potentially severe security risks. We see a demand for EOSC provided solutions that connect graphical frontends to cloud VMs, enabling interactive usage models, giving graphical input channels to running processes and graphical output channels to deliver e.g. intermediate results. This can facilitate the migration of workflows, but also users and developers from HPC Clusters to Cloud Services.

EOSCpilot: Science Demonstrator Final Report	
Date / Type	2018-09-27
Science Demonstrator Title	PROMINENCE (Fusion Science Demonstrator)
Representative name, affiliation and email from proposing organisation(s)	Shaun de Witt Culham Centre for Fusion Energy, shaun.de-witt@ukaea.uk
Main Shepherd name, affiliation and email	John Kennedy MPCDF, j.kennedy@rzg.mpg.de
Secondary Shepherd name, affiliation and email	Frank Schluenzen DESY, frank.schluenzen@desy.de
General part	
Achievements (> 200 words)	**Preliminary investigations** We initially looked at existing technologies to see if anything would meet (or help to meet) our requirements. • Used Infrastructure Manager to deploy static SLURM clusters with OpenMPI and Singularity installed, and tested running MPI applications on EGI FedCloud sites. • Used EC3 (https://servproject.i3m.upv.es/ec3/) to deploy elastic SLURM clusters on clouds and run MPI jobs. • Used EC3 combined with OpenVPN to deploy elastic SLURM clusters spanning multiple clouds, enabling us to run MPI jobs on one preferred cloud site but bursting onto others when free resources were no longer available on the first cloud. • Successfully deployed the INDIGO-DataCloud PaaS Orchestrator, including all its dependencies, to investigate whether the Orchestrator could be used for provisioning resources and determining what clouds to use. • Successfully deployed INDIGO-DataCloud FutureGateway to investigate whether it could be used as the frontend for users to interact with. **Development of the PROMINENCE platform**

	A platform was developed which enables users to easily submit containerized jobs and monitor their progress. Infrastructure is automatically provisioned on demand across a range of clouds. • A Python Flask-based RESTful API was developed as the front end. A proof-of-concept integration with the INDIGO-DataCloud Identity and Access Management service (IAM) was successfully tested. • A Python CLI tool was written to make job submission as simple as possible for users. • HTCondor is used for scheduling and executing jobs, and transferring data to/from jobs as necessary. • Hooks were developed to allow HTCondor to automatically provision cloud resources and destroy them as necessary. • Infrastructure Manager is used for provisioning resources across a variety of clouds. EGI FedCloud, OpenStack, Google Cloud Platform and Azure were all successfully tested. • A wrapper for Infrastructure Manager was developed in order to select which clouds to provision resources on. In addition to automatically handling failures and backing off from using clouds which fail, a unique feature is that it provides 'hierarchical cloud bursting'. For example, a user's local private cloud can be used preferentially, followed by a national resource cloud, followed by EGI FedCloud sites, followed by public cloud(s). • Transient storage for jobs is provided by attached disks. For multi-node MPI jobs NFS is used to provide storage across all the required nodes. BeeGFS On Demand was also tested as an alternative to NFS. • Output files are automatically copied to Ceph-based object storage using the Swift API, and users are provided with temporary URLs that can be used to access their data. • Monitoring is provided by Telegraf (metrics collection), InfluxDB (storing time series data) and Grafana (visualization). **Example use case - Tokamak tritium production** A very successful early use case of the PROMINENCE platform was optimizing the performance of tritium production in a nuclear fusion reactor. Ensuring a self-sustaining supply of tritium for use in future fusion reactors is one of the key challenges for the fusion community. Breeder blankets, which occupy the interior of a nuclear fusion reactor, are designed to produce tritium when bombarded with neutrons production by the fusion reactions. Innovative software developed by CCFE enabled the geometries

	and material specifications of blankets to be adjusted to optimize tritium production. All the required Monte Carlo simulations were run using PROMINENCE making use of EGI FedCloud resources. Over a period of around a month up to 300 cores were used across CESNET-MetaCloud, IN2P3-IRES, RECAS-BARI and CESGA. **Other example use cases** A few other example applications from the fusion community were also successfully tested on PROMINENCE: • Geant4: an MPI code which simulates the passage of particles through matter, • ASCOT: an MPI code which simulates charged particles in a Tokamak, • Raysect: a scientific ray-tracing code developed by staff at CCFE. Some performance tests were carried out using a JET CAD model.
Problems encountered	Very limited resources on EGI FedCloud were made available to the Science Demonstrators via the fedcloud.egi.eu VO, which only has opportunistic access to resources. This restricted our testing to scales well below what we would expect to use in a production setting and prevented us from being able to test the scalability of our platform. The challenge of being able make the most of the opportunistic fedcloud.egi.eu VO also directly resulted in us having to develop a new system from scratch rather than making more use of existing services.
Data management and handling of sensitive data, with reference to plan.	N/A
List of outreach activities	A talk was presented at the European HTCondor Workshop 2018: https://indico.cern.ch/event/733513/contributions/3118645/ The PROMINENCE platform was discussed with a number of different groups within CCFE resulting in some interest in the

	service. We will continue to promote this service outside of EOSC both to the fusion community and to wider audiences.
Specific Feedback on:	
Technical challenges and issues encountered	We investigated use of the Orchestrator (and its dependencies including IAM, Cloud Provider Ranker, SLA Manager, CMDB, etc) from INDIGO-DataCloud as the means of deploying infrastructure across clouds. We found that the existing documentation was inadequate and required assistance from INFN in order to successfully deploy it. One significant issue with Orchestrator was that EGI FedCloud sites only supported the OCCI API while support for OCCI had been deprecated in Orchestrator. As it also does not have the ability to handle deployment failures automatically it was found to clearly not meet our requirements. We also investigated use of INDIGO-DataCloud FutureGateway and found the deployment documentation lacking. A number of issues were encountered with Infrastructure Manager during the course of this work, resulting in 8 GitHub issues being submitted. Fortunately the developer responded very quickly and resolved all the problems. Different credentials were used to deploy infrastructure on EGI FedCloud (X.509 certificate) and access storage (access key and secret key). Ideally a single credential would provide access to both compute resources and storage, but this would possibly depend on the existence of interoperable AAI infrastructures.
Proposed measures /suggestions to mitigate technical issues	The INDIGO-DataCloud documentation, in particular relating to deployment and integration of the complete stack, should be significantly improved. AAI interoperability would help to solve the problem of having to use different credentials for accessing compute and storage resources.

Political challenges encountered	Access to fusion data is generally restricted to users with access to specific compute clusters, therefore accessing this data from clouds is not possible as the data is neither open nor even accessible off-site with authentication. This limited us to testing simulation applications or software with well-defined input data for which we could get permission to copy to a cloud. Much fusion software is not open and can only be accessed with permission of the developers. Furthermore, some software is even more restricted due to legal aspects of code usage, completely preventing it from being run on clouds. This restricts the number of fusion applications which can make use of the technology and also prevented us from being able to use Docker Hub for storing container images.
Proposed measures /suggestions to mitigate political issues	The partners in this project (UKAEA, Chalmers University and MPIPP) have put together a bid to develop a platform supporting open data for the fusion community, with the support of EUROfusion. UKAEA (and the Advanced Computing Group) are advocating moving to open source solutions where they exist and are supporting users by testing and validation of results between closed and open source solutions.
Cultural challenges encountered	A lot of fusion software is usually built and run on a very small specific group of clusters at centres carrying out fusion research, and is not really designed for users to be able to build the software themselves. Due to the extensive use of Linux modules and dependence on existing software installed on these clusters it is quite difficult to install the software in a generic environment, e.g. in a container. Additionally, in some cases there are a large number of legacy codes written in IDL which due to licensing issues cannot easily be run in a cloud environment.
Proposed measures /suggestions to mitigate cultural issues	Interest in the use of containers is slowly starting to develop within the fusion community, but it is not yet at all common. The use of clouds seems to be almost non-existent. Training and dissemination in the benefits of containers and usefulness of cloud resources would be helpful.

Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	Very little interaction between us and EGI FedCloud infrastructure providers was necessary. At one point we experienced an incident when all of our VMs running at CESNET-MetaCloud were lost. This was thoroughly investigated by both the site and EGI FedCloud experts, however our VMs could not be recovered. While for worker nodes this type of incident is not an issue, it did impact the frontend of PROMINENCE (a long-running service) which was running on CESNET-MetaCloud. STFC quickly gave us access to S3/Swift Ceph-based object storage when requested. As data management was not a planned activity, we did not integrate with any of the EOSC storage services, but are considering usage of B2DROP for a future evolution of the service.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	Infrastructure Manager was straightforward to deploy and the documentation is quite detailed and useful. The developers responded very quickly to numerous bug reports, as mentioned above. Similarly IAM was straightforward to deploy and is well documented. As mentioned previously, the existing documentation for INDIGO-DataCloud Orchestrator (plus its dependencies) and FutureGateway is inadequate. There are no services in the catalog which have the ability to deploy infrastructure automatically across multiple FedCloud sites. Making opportunistic use of resources is currently very limited within EGI Fedcloud since the general access model assumes a direct connection between a science community and a provider, brokered by EGI.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	Interoperability between the different AAI infrastructures in different organisations is an important issue that needs to be solved in order to simplify access to both compute and storage resources. Increased and standard interfaces between EOSC core components is currently limited, meaning each user needs to perform their own tailored integration.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions?	Within the fusion community there is still a significant skills gap regarding both the use of containers and general cloud computing, as it seems that most of the community only has experience with traditional HPC environments. To some extent this can be overcome by internal seminars, but experience has shown that only a small subset of potential users come to these

	(typically younger researchers); reaching out to more experienced staff is still problematic as they show no interest in the technology. Our efforts to provide a simple, batch like, command may help in this.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	Based on deliverable 3.3 (https://eoscpilot.eu/sites/default/files/eoscpilot_d3.3_final-withannexes-forweb.pdf), we believe the following areas should be prioritised: Item 5 "Develop an Evaluation and Ranking of Openness Maturity of EOSC services, infrastructures and other resources" since this ranking could encourage researchers to make their research artefacts more public Item 7 "Reduce regulatory complexity for researchers" would encourage adoption by reducing paperwork for researchers, but must adhere to overriding privacy or sensitivity issues Item 16 "Develop, support and promote an EOSC Skills and Capability Framework as a common reference point", but extend this beyond RDM to include training on EOSC services Item 26 "Adopt the recommendation of the OSPP Working Group on Rewards and embed Open Science in the evaluation of researchers at all stages of their career"
Government issues: (WP2): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP2	Engagement with stakeholders, particularly research communities, should be prioritised. This should include not only inviting high level people to stakeholders events, but also gaining grass roots support by attending community events with demonstrations, presentations from other researchers (not RI representatives). Attacking adoption from both directions will encourage more rapid adoption of services.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	Run single-node batch jobs and multi-node MPI codes from the fusion community in containers using a single API. Deploy and configure infrastructure on-demand on both private and public clouds. Decide where to deploy clusters based on pre-defined policies, such as preferring local resources and bursting out to external clouds, both academic and commercial, when local resources are busy.

	Automatically copy output data generated by jobs to scratch space which is accessible to users for a specified time period. Provide a portal which has the ability to provide authentication and allows users to define jobs and check the status of their jobs. A container registry which allows users to upload and store images, make use of them in batch jobs but restrict access to specified groups of users.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services	The system should be easy to use.
Other (please specify)	

EOSCpilot: Science Demonstrator Final Report	
Date	**2018-Nov-08**
Science Demonstrator Title	EPOS/VERCE: Virtual Earthquake and Computational Earth Science e-science environment in Europe (acronym: EPOS/VERCE)
Representative name, affiliation and email from proposing organisation(s)	Andreas Rietbrock - University of Liverpool, a.rietbrock@liverpool.ac.uk Alessandro Spinuso, KNMI, spinuso@knmi.nl André Gemünd, Fraunhofer SCAI, andre.gemuend@scai.fraunhof er.de
Main Shepherd name, affiliation and email	Giuseppe La Rocca, EGI Foundation, giuseppe.larocca@egi.eu
Secondary Shepherd name, affiliation and email	Michael Schuh, DESY, michael.schuh@desy.de
General part	
Achievements so far (> 200 words)	The Forward Modeling tool hosted by the gateway has been extended with an additional simulation code (SPECFEM3D_GLOBE). The portlet code was upgraded to Liferay 6.2 to be compliant with gUSE 3.7.5. The integration of 3D virtual globe with a list of other 2D projections. The processing workflows enabling the Misfit analysis (data

	download preprocessing and misfit) have been refactored to support their execution on the FedCloud, in addition to the current HPC resources. Moreover, the lineage and provenance services and repository have been upgraded to a later version of the S-ProvFlow API and storage system. It has been deployed in its containerized version within SCAI resources. This also included the upgrade of the provenance exploration tools, in a version specific to the VERCE portal. About the AAI: On a development copy of the Gateway and its services, an evaluation of the authentication methods has been carried out. The portal has been extended to allow the retrieval of Per-User Sub-Proxy certificates from the eToken proxy certificate service for Workflows, additionally to its community-specific IdP. In a second evaluation, login via OpenID Connect through the EGI Check-In service has been successfully validated, though in the latter case, the retrieval of the required X.509 certificate via the RCauth service could not be finished successful during the runtime of the demonstrator.

Problems encountered	The science gateway software (gUSE) had to be upgraded and required several specific fixes and extensions to support the FedCloud OCCI interface. This included the support of delegation mechanism to authorise the transfer of data from the cloud to the managed iRODS instance. A new release of the DCI-Bridge Virtual Appliance (VA) used by the science gateway to be interoperable with the EOSC cloud infrastructure has been developed. The VA is now registered in the EGI AppDB Cloud marketplace.
Data management and handling of sensitive data, with reference to plan.	Simulation results are sent to iRODS. Their metadata and lineage are extracted during processing and stored within the S-ProvFlow system. They can be exported on demand in PROV compliant formats (XML and Turtle). Experiments can be accessed to be reused, combined and validated by interrogating the S-ProvFlow system through several user interfaces integrated in the Forward Modeling tool.
List of outreach activities	Training delivered in the context of the EPOS project during the ORFEUS/EPOS workshop in October 2017. A hands-on session on the usage of the VERCE portal has been organised showing and testing its main features. The second version of the portal User Guide has been released in April 2018 (https://portal.verce.eu/UserM anual-portlet/html/index.jsp).

Specific Feedback on:	
Technical challenges and issues encountered	The EPOS/Seismological component tried to benefit from previous investments on the science-gateway technology and front-end implementations, by integrating FedCloud resources in the current framework. This activity resulted to be feasible, yet complicated because of the large need of support, which was kindly provided often with in-kind contributions. Also, the demonstrator faced challenges due to perceived gaps in the technical service offering.
Proposed measures /suggestions to mitigate technical issues	The infrastructures should be more specific and technical about their service offering. While the current service catalogues are already an improvement to the prior absence of documentation, they do not provide much more than high-level abstracts. For some services, there is simply too much choice offered in components (without offering them hosted), without any clear recommendations or evaluations. The infrastructures should focus on stable platform services to a concise set of services and openly communicate this.
Political challenges encountered	There are risks brought by the technical choices taken by communities when these require sustained cooperation and support to face technological advances.

Proposed measures /suggestions to mitigate political issues	In the future, such a problem should be solved within the EOSC by keeping track of third parties EOSC compliant tools, promoting them to the communities with enough description about usage and aim.
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	
Proposed measures /suggestions to mitigate cultural issues	
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	Not easy to keep the focus of the communication between the scientific communities interested in doing their research and advancing their career, engineers interested in improving a better and sustainable service, and constantly evolving e-infrastructures. The risks brought by rapid obsolescence is sometimes perceived as too high and the participation of the scientists in the POC and pilots should be acknowledged more. To facilitate the migration of part of the Misfit workflow on the EOSC cloud infrastructures, cloud providers have been invited to allocate some resources for supporting the pilot project. The interaction with the cloud providers was liaised by WP5.

Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The pilot used cloud resources provided by three providers of the EGI Federation (IN2P3-IRES, SCAI and HG-09-Okeanos-Cloud). There is no need for additional resources at the moment. The access to these resources have been enabled by WP5.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	It is not clear what is the near destiny of OCCI and the support of the gUSE technical framework from the EOSC? The reproducibility problem should start to be addressed structurally, scaling from ad-hoc solutions to generic services. Computational tools offered by EOSC should be aware of the existence of these service and use them offering possibilities of adaptation to the use-cases (granularity, privacy rules and visibility), and contextualisation to the user's domain.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	Young researchers in any discipline requiring computational and data-management services should have access to ad-hoc trainings focused on the accomplishment of their research practices through EOSC. For instance, the Jupyter training given at DI4R 2018 was a good starting point to showcase the potential of such service and gain technical feedback. However, next edition should address young researchers.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	

Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	EOSC services and catalogues of tools should consider ongoing efforts by Research Infrastructures to implement and operate services. EPOS perspective is that there will be services that can be exported to EOSC, but other that cannot for different reasons (privacy, security, performance). I guess this is in line with what is envisaged in figure 8 of the document
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	For single users, EOSC should facilitated access to virtual resources globally and with the authorisation and delegation mechanisms put automatically in place. Data-Stores should be also available with configurable level of support for stewardship of the saved data.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	Possibility for expert research-developers to develop serverless workflows, where available resources are allocated and scaled dynamically, removing the burden of configuring specific providers. Such mechanisms should be provided or brokered by the EOSC, who should also advertise to the stakeholders those tools that are compliant with such mechanisms.
Other (please specify)	

EOSCpilot: Science Demonstrator Pre-Final Report

Date	**2018-November-05**
Science Demonstrator Title	**Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability.**
Representative name, affiliation and email from proposing organisation(s)	Jordi Rambla, Centre for Genomic Regulation, jordi.rambla@crg.eu Cedric Notredame, Centre for Genomic Regulation, cedric.notredame@crg.eu
Main Shepherd name, affiliation and email	Gianni Dalla Torre, European Bioinformatics Institute, gianni@ebi.ac.uk
Secondary Shepherd name, affiliation and email	Matthew Viljoen EGI.eu, Amsterdam, Netherlands matthew.viljoen@egi.eu
General part	

Achievements so far (> 200 words)	Our Science Demonstrator is aimed to explore the feasibility of data reproducibility and data re-mastering in genomics. Both reproducibility and re-mastering require pipeline portability and the availability of metadata describing the tools and resources used within a given pipeline. Leveraging a particular dataset stored at EGA (https://ega-archive.org/) and through direct communication with the original data providers, we were able to implement a pipeline aimed to reproduce the EGA deposited data. In parallel, a similar pipeline was assembled using recently implemented tools. Both reproducibility and re-mastering pipelines were implemented using Nextflow (https://www.nextflow.io/), a language developed at CRG enabling scalable and reproducible scientific workflows. Containerized versions of the pipeline were executed at Barcelona Supercomputing Center (https://www.bsc.es/) demonstrating the portability of the pipelines implemented in our SD. Finally, pipelines were deposited in Dockstore (https://dockstore.org/), an open platform for sharing containerized scientific workflows and tools. In order to facilitate their re-use, deposited pipeline containers were described through the meta-data available at Dockstore and additional fields that were considered relevant basing on our experience. In order to assess our ability in reproducing a genomic dataset, data computed through the pipelines implemented in our SD were compared to the EGA deposited ones. Some unexpected limitations forced us to test reproducibility in a fraction of the original data. Unluckily, the usage of partial data introduces a systematic bias, lowering our reproducibility capacity. After evaluating qualitatively and quantitatively the

	generated differences, we demonstrated that disparity between deposited and computed data is low. Furthermore, we proved that produced differences should have a limited effect on the downstream genomic analysis, on which genetic variants and genotypes are obtained. Overall, 97.38% of the variants are called from both data deposited at EGA and data obtained through the reproducibility pipeline. Among the shared variants, 99.66% of the called genotypes obtained from the two datasets perfectly agree, while on average 0.33% of genotypes are discordant on the 13 considered samples. The re-mastering pipeline is aimed to demonstrate the feasibility of data refreshing leveraging recently implemented tools and resources. We proved that data can be easily refreshed and the recent algorithm improvements lower tremendously the required computational resources. Re-mastering pipeline was executed twice using the originally using hg19 version of the human reference genome and the recently released hg38 version. This double execution allowed us to test separately to which extent the tools and the reference genome affect the resulting data. As expected, a gradient of increasing difference is observed when passing from the reproducibility pipeline to the re-mastering pipelines. The highest degree of difference is observed for the re-mastering pipeline using the human reference genome hg38. However, we were not able to assess if the usage of new tools and reference genome generates more precise results, which requires extensive analysis and it is out of the scope of our SD. To conclude, our work demonstrates that is technically possible to reproduce at a high degree genomics data and that current available tools, language and repositories are providing useful means to

	foster reproducibility and data refreshing in science. The resulting workflow application has been deposited in the Dockstore catalog and accessible at this link https://dockstore.org/workflows/github.com/CRG-CNAG/EOSC-Pilot.

Problems encountered	Originally, the SD was conceived to replicate the whole GoNL dataset, but unluckily various factors limited the possibility to perfectly replicate it. Replicability of genomic data requires a detailed understanding of the originally used pipelines and the availability of the original tools and resources. Due to the fast progress of the field, for genomic data obtained just at the beginning of the Next Generation Sequencing (NGS) era, it is much more difficult to collect all the required information and resources. Indeed, the absence of consolidated standards and platforms on which to store the originally implemented pipelines, slowered the understanding of the different steps composing the pipelines. In addition, the originally used version of GATK was not available anymore, while some ancillary files originally provided by the 1000 Genomes Project Consortium, were updated and included also genetic information obtained recently. Overall, the above exposed restrictions limit the possibility to perfectly replicate the original data. Even if the selected dataset is suited for our SD purposes, its size limited to replicate it in its entirety. GoNL dataset is composed by different file types that allow to monitor replicability at each step. However, the usage of outdated tools requires extensive computational resources and prolonged execution time. To overcome these technical issues, we decided to limit our analysis only to a fraction of the whole GoNL dataset.

Data management and handling of sensitive data, with reference to plan.	GoNL consortium was contacted and permission for the usage of data stored at EGA was requested. All data stored at EGA are encrypted in a secure manner. To carry out our analysis, we decrypted the raw fastq files related to the 13 considered samples and we considered only reads mapping to chromosome 21. Filtered raw data were then passed to BSC by using encrypted SSL transfer channel, for pipeline execution in a third-party infrastructure to test pipeline portability. Once pipelines were executed, processed data were returned to EGA to assess and quantify the generated differences compared to the original data. Both raw and processed data were deleted from BSC. Finally, raw data, BSC processed data and difference analysis sensitive data were encrypted at EGA.
List of outreach activities	Purposes and analysis our SD were exposed conferences and in intra-institute meetings. A poster was prepared for the B-Debate conference (Barcelona, 4th-5th October 2018) to explain the impact of our SD in fostering open science policies. Additionally, the SD was presented during the EGA annual Science Advisory Board (SAB) meeting (Barcelona, 18th-19th June 2018) and evaluated by a panel of experts. Finally, the SD was also presented in the intra-institute course "Nextflow: Reproducible in silico Genomics" (Barcelona, 14th-15th September 2017).
Specific Feedback on:	

Technical challenges and issues encountered	In order to perfectly replicate the original data, it was required the usage of the Genome Analysis Toolkit (GATK) version 1.0. Unluckily, due to outdated requirements, we were not able to install the originally used version. As explained previously, the size of the selected datasets posed unexpected computational challenges.
Proposed measures /suggestions to mitigate technical issues	In substitution of the original GATK version, we installed the closer available version (GATK v1.2). To mitigate the issue related to GoNL dataset size, we carried out the analysis only in a fraction of the total GoNL dataset. To this end, we considered only 4 out of the 250 families were considered. The selected families represent the ones corresponding to the 25^{th}, 50^{th}, 75^{th} and 100^{th} percentiles of the distribution of input file sizes. Furthermore, we limited our analysis considering only genomic data mapping to chromosome 21.
Political challenges encountered	No political challenges were encountered
Proposed measures /suggestions to mitigate political issues	
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	Since all parties and collaborators are currently working in genomics and have familiarity with the performed analysis, no particular cultural challenges were encountered.

Proposed measures /suggestions to mitigate cultural issues	Promote the adoption of best practices for scientific workflow reproducibility such as: 1) describe the computational environment with a package manager such Conda/Bioconda whenever possible. 2) Capture the computational environment using a container technology eg. Docker or Singularity. The container should be depositing in a community collection like BioContainers (https://biocontainers.pro/). 3) Use workflow system like Nextflow to enable portable deployments across different cloud and clusters with minimal changes 4) Deposit the data analysis code into a public source code management platform 5) Disseminate the resulting application through a open catalog like Dockstore which implement FAIR principles and community best practices.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	GoNL consortium was open to understand the SD requirements and when possible quickly provided the requested information in detail. However, some expected difficulties have arisen during pipeline implementation process. Indeed, our whole SD is aimed to overcome the inevitable issues generated by no-portable pipelines and by the absence of pipeline metadata, descriptors and standards in genomics. Within the SD, BSC provided the computational resources and the infrastructure to execute the pipelines implemented at CRG. The communication between the SD partners was fluent and emerging technical issues were immediately communicated and when possible addressed.

Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	Given that our SD is based on the processing of sensitive data owned by third-party data providers, we could not use yet any service implemented within EOSC for data processing. Moreover, given that EGA is itself a repository, we did not store any produced data in the repositories available at EOSC.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	To foster interoperability, we adopted Nextflow to implement our pipelines. Nextflow is an emerging language for genomic pipelines and it is among the ones accepted by Dockstore, a repository endorsed by GA4GH for sharing Docker-based tools. GA4GH is a global effort involving major stakeholders for the definition of standards in biology and the implementation of tools service the whole biological community. Moreover, pipelines were deposited using a docker image, which contains nearly all the required tools and resources for pipeline execution. Unluckily, GATK required version (v1.2) for the reproducibility pipeline is protected by copyright and cannot be distributed within our image. To overcome this issue, we documented this problem and we adopted a meta-command showing the options to use. Finally, to promote interoperability, a FAIR representation of the objects and their relationships is required. FAIR-compliant semantic repositories address the above issue and we think that this kind of repository is missing in the EOSC ecosystem.

Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	The experience provided by this SD proves the need to foster the adoption of standards to describe and document genomic pipelines. Even if the field is moving toward open science policies, pipeline portability and documentation are still perceived by providers as time consuming tasks that are not fully part of their scientific obligations. Continuous training is required to form a new generation of scientist more engaged in adopting open science policies and behaviors.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	No specific feedback
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	No specific feedback

Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	From our work in this SD, we conclude that standards and specific metadata to describe pipelines, and long term repositories for such artifacts, are a priority to foster reproducibility in genomics. Based on our experience, we suggest to implement an universal method valid across several scientific fields on which it should be possible for each step of a pipeline to declare: the used tool, its version, the input, the options and the output. Finally, the presence of links to software containers to the used tools would provide a determinant contribution to data reproducibility. These containers should be available as long as the raw data is, in order to keep the whole experiment reproducible.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	The stability of the above mentioned functional requirements is necessary for long term open science policies.
Other (please specify)	

EOSCpilot: Science Demonstrator Pre-Final Report	
Date	**2018-Jul-19**
Science Demonstrator Title	Cryo Electron Microscopy
Representative name, affiliation and email from proposing organisation(s)	Carlos Oscar Sorzano Natl. Center of Biotechnology (CSIC) coss@cnb.csic.es
Main Shepherd name, affiliation and email	Gergely Sipos EGI gergely.sipos@egi.eu
Secondary Shepherd name, affiliation and email	Giuseppe La Rocca EGI giuseppe.larocca@egi.eu
General part	
Achievements so far (> 200 words)	The CNB team has advanced in bringing the FAIR principles to Electron Microscopy (EM) Single Particle Analysis. We have concentrated in the first stages of the image processing pipeline as carried out in Scipion (http://scipion.i2pc.es). These steps are normally carried out in large EM facilities and we have developed tools to be able to export the workflow as well as the data in a way that the early stages can be fully automated. The workflow is written as a Json file that can be read from Scipion and also visualized in a web browser through a javascript widget. We have contacted the EMDB and EMPIAR coordinator (the largest public databases for EM) so that structural biologists can submit the data as well as their workflow to these databases, and any other researcher can reuse the data and reproduce the same results as the original researcher up to the estimation of the microscope aberrations. We have written a tool for automatic submission of these files. We have also evaluated the possibility of using Common Workflow Language as a standardized reporting tool for other programmers. Although, the EMDB coordinator likes this idea, we all thought that making such a step requires more maturation and that it should be addressed by a future extension of the activities developed in this EOSC pilot. At the moment, we are defining use cases in which the EOSC hardware facilities can be used by the EM facilities and/or structural biologists to perform their analyses.

Problems encountered	No specific problem
Data management and handling of sensitive data, with reference to plan.	We have contacted the EMDB and EMPIAR databases. They have agreed to accept the workflow descriptions sent by Scipion as well as incorporating the web viewer to allow its visualization. We have devised a mechanism for automatic submission of the data as well as the workflow description to the public databases.
List of outreach activities	Contacting the EMDB and EMPIAR databases. Contacting the developers of Common Workflow Language.
Specific Feedback on:	
Technical challenges and issues encountered	There is no large repository of acquired data so that at acquisition time, the data cannot be transferred along with its FAIR descriptor to any repository.
Proposed measures /suggestions to mitigate technical issues	It would be very useful to implement a mechanism of data storage so that the metadata and workflows associated to the acquisition are deposited in some common place (currently there is no hardware for this), and eventually recovered for its submission to the final structural biology databases.
Political challenges encountered	EMPIAR is only accepting data associated to final results. New acquisitions have no place for depositing its metadata and associated workflow. Eventually, this pilot will be regularly used at facilities. A mechanism for user identification (either the facility, or the

	final user) should be provided if EOSC is to provide hardware for the image processing.
Proposed measures /suggestions to mitigate political issues	Definition of a policy for acquisition metadata and workflows, and another one for user authentication (technically this is solved) for using the EOSC machines.
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	Structural biologists are not used to machines in the cloud, EGI, etc. Many of them do not know which options they have available.
Proposed measures /suggestions to mitigate cultural issues	Training and dissemination in the use of cloud machines.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	The Electron Microscopy Data Base has been very open to incorporate our visualizer into their database. This integration would facilitate external users to verify the image processing steps performed to attain the three-dimensional structure deposit.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	What is missing is a full description of the whole image processing pipeline.

Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	We have explored Common Workflow Language as a description of the processes performed. Although promising, at the moment no further step has been taken, since it has to be decided community wise.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	There is a need of users understanding why it is important to share the data and the processing giving raise to the final structure.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	Investment in hardware for a public repository of newly acquired metadata and processing workflows.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	No specific feedback
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	A well defined language to report workflows.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/	Stability over time of this language.

wiki/Non-functional_requirement)	
Other (please specify)	

EOSCpilot: Science Demonstrator Final Report	
Date	**2018-Nov-15**
Science Demonstrator Title	LOFAR
Representative name, affiliation and email from proposing organisation(s)	ASTRON: Hanno Holties, holties@astron.nl Emanuela Orru, orru@astron.nl Tammo Jan Dijkema, dijkema@astron.nl Yan Grange, grange@astron.nl Rob van der Meer, meer@astron.nl Raymond Oonk, oonk@astron.nl (switched to SURFsara during the project) NLeSC: Hanno Spreeuw, h.spreeuw@esciencecenter.nl Niels Drost, n.drost@esciencecenter.nl Ronald van Haren, r.vanharen@esciencecenter.nl Arnold Kuzniar, a.kuzniar@esciencecenter.nl Stefan Verhoeven, s.verhoeven@esciencecenter.nl Ben de Vries, b.devries@esciencecenter.nl Felipe Zapata, f.zapata@esciencecenter.nl Valentina Maccatrozzo, v.maccatrozzo@esciencecenter.nl Elena Ranguelova, e.ranguelova@esciencecenter.nl Laurens Bogaardt, l.bogaardt@esciencecenter.nl Rob van Nieuwpoort, r.vannieuwpoort@esciencecenter.nl SURFsara: Robert Griffioen, robert.griffioen@surfsara.nl Axel Berg ,axel.berg@surfsara.nl Natalie Danezi, natalie.danezi@surfsara.nl Coen Schrijvers, coen.schrijvers@surfsara.nl Raymond Oonk, Raymond.oonk@surfsara.nl (switched from ASTRON during the project) CWL Project: Michael Crusoe, mrc@commonwl.org Pythonic.nl:

	Gijs Molenaar, gijs @ pythonic.nl INAF: Fabio Pasian, pasian@oats.inaf.it
Main Shepherd name, affiliation and email	Thomas Zastrow, MPCDF, thomas.zastrow@mpcdf.mpg.de
Secondary Shepherd name, affiliation and email	John Kennedy, MPCDF, John.kennedy@mpcdf.mpg.de Nicolas Fabas, MPCDF, nicolas.fabas@mpcdf.mpg.de
General part	
Achievements so far (> 200 words)	LOFAR is a European scale distributed radio astronomical instrument that since the start of Science operations in 2012 has generated over 30 PB of data. This data is distributed over three Long Term Archive (LTA) locations hosted by SURFsara (NL), FZJ (DE), and PSNC (PL). A typical imaging dataset has a size of 30 TB and requires further processing before scientific results can be generated. The further processing is typically conducted or requested by end-users and is ideally done on a compute cluster close to the storage location. Given the inhomogeneity of compute clusters and the complexity of the processing applications in terms of (legacy) code and dependencies, it is essential to be able to deploy and execute the applications in a portable manner. As a Science Demonstrator in the EOSCpilot project, the deployment of LOFAR processing workflows within the EOSC infrastructure using containers and a standardized workflow definition language has been investigated and Proof of Concept functionality has been implemented for a web service that runs a processing workflow from data stored in the LOFAR LTA and metadata from the LOFAR LTA has been mapped to, and stored in, a web-based Linked Data platform to gain experience with exposing LOFAR data by applying general purpose FAIR data principles and tools. We have demonstrated that the Common Workflow Language standards are sufficient to enable the portable description of radio astronomy data analysis pipeline via three real world examples. The first pipeline is Prefactor which is used for performing the first calibration of LOFAR data. The second pipeline is Presto. Presto is a framework to search for pulsars in radio frequency datasets. The final pipeline is called Spiel. Spiel is a radio telescope simulator, which generates 'dirty' images from an input image. For the project, we started with making Debian packages for all required software. For our pipelines, we evaluated three

	containerization platforms, Docker, singularity and uDocker (user-space Docker). Singularity is supported by various HPC providers. To tie our software together, we used Common Workflow Language (CWL) which is a standard for describing data analysis tools and workflows. There are various (open source) workflow engines that can execute processing workflows and have support for the CWL syntax. This makes CWL satisfy our flexibility requirement and makes a vendor lock-in much less likely. The engine chosen for the first deployments is the open source Python-based Toil, a complete but still lean implementation that supports parallelization and scheduling. We have run the pipelines on a variety of platforms, A MacBook, a Linux Server, the SURFsara HPC cloud environment , Cartesius, the Dutch national supercomputer, and finally on the HTDP system as part of a beta testing effort which included the staging and retrieval of a 15 TB dataset from the LOFAR Archive. In this case we retrieved data hosted by FZJ site -- which is one of the three LOFAR archive sites. Benchmarks were performed on small datasets to evaluate the relative performance on the different systems. An experimental webportal based has been created that allows users to select a dataset from the LOFAR LTA Catalog and initiate a default processing pipeline . Finally, an experimental Docker-deployable FAIR/Linked Data platform (lofar-ld) has been created based on open source Virtuoso and FAIR Data Point services, demonstrating the translation of LTA-associated metadata into semantic (RDF-based) form using controlled vocabularies and ontologies. For FAIR data access, we provisionally conclude that the International Virtual Observatory Alliance provides the most important functionality and standards for data sharing within the astronomical community. General purpose FAIR data services may become useful tools to increase visibility and re-use of (radio-) astronomical data in non-astronomical domains but a practical application is yet to be demonstrated.
Problems encountered	During the development, we found some bugs in the CWL implementations and ambiguities in the CWL standard. For example, nested directory structures, typical for radio astronomy datasets (measurement sets) were not properly handled. This has been fixed. Another important addition to the standard introduced upon our request is the optional in-place

	writing. When datasets grow in the order of terabytes you do not want to copy the results around but modify them in place. Lastly, we have found a bug where intermediate array products were not properly sorted, resulting in alignment problems if you use multiple arrays. This issue has been resolved. The response from the CWL community, the CWL reference engine, and the Toil CWL engine has been excellent. Toil depends on a scheduler for running jobs on a cluster which needs to be deployed if none is at hand. We decided to try to set up a Mesos cluster to handle scheduling on a mini cluster deployed on the SURFsara HPC cloud but were not successful and did not have enough time left to fully investigate this issue so this setup could not be benchmarked. A typical real-world LOFAR dataset is multiple terabytes in size. Being allocated sufficient workspace connected to the processing system can be a challenge. The CWL runners we tested are not aware of data locality. However, the CWL standard does not make a unified storage assumption so it is possible to switch to more advanced workflow orchestrators in the future without having to change our pipelines. During HTDP beta-testing, a not-yet solved shortcoming of Toil was exposed: Re-downloading of Docker images (from which Singularity images are built) on each run. Also, Singularity support is limited in that Singularity images are only supported through a bas-Docker image import. A CWL runner with complete native support for Singularity images would be preferred. Also, Toil, natively, requires the user to specify the --mem option for slurm jobs. This is done for tracking purposes, however for our software this option is dangerous as our software does not respect memory very well. Container security: Docker is known to be insecure in a multi-user environment. Singularity seems to be more fitted for this purpose. Although Singularity does contain an executable with a 'setuid bit' set, which might expose a security risk, many HPC sites have adopted this technology. Rigid dependencies between Cuda drivers and Nvidia kernel modules: If a container containing a Cuda version not matching the hosts Nvidia kernel module version is used, GPU acceleration just does not work. This breaks the host-platform independence. For Docker there is a workaround in the form of a Docker extension, but if a different container technology

	there is no solution (yet). Although support for Singularity is growing, there is not yet standard support for it on all HPC clusters in the EOSC infrastructure. Since maintaining local installations of (legacy) software with complex dependencies is not sustainable, this limits use of computational platforms that do not support containerized (preferably Singularity based) deployments. As a consequence, data may be required to be copied to alternative locations, which at the level of tens to hundreds of terabytes becomes a serious issue.
Data management and handling of sensitive data, with reference to plan.	As in plan (for pilot using public data, currently manually copied to target systems), no sensitive data.
List of outreach activities	- EOSC Stakeholder Forum, Brussels 28 - 29/11 2017 - Poster presentation at DI4R, Brussels 30/11 - 1/12 2017 - Blog post by Gijs Molenaar: https://medium.com/@gijzelaerr/portable-radio-astronomy-data-processing-pipelines-4e6ba8b00ca3 - ASTRON Daily Images (1/12/17, 16/1/18): http://www.astron.nl/dailyimage/ - …
Specific Feedback on:	
Technical challenges and issues encountered	We decided not to evaluate the CWL + Singularity deployment on the SURFsara grid cluster for lack of support for CWL on that system although this type of cluster is one of the most fitting systems for processing large LOFAR datasets. Retrieving data from a tape backend and large data volumes at the level of tens to hundreds of terabytes is a major challenge when the data are not local.
Proposed measures /suggestions to mitigate technical issues	Support for CWL and, in particular container based, preferably Singularity, deployments on all EOSC clusters (including grid/HPC). A wide(r) deployment of high capacity network connections between computing centers; on demand point to point connections.
Political challenges encountered	Sustainability of services in a collaborative environment: Who takes responsibility for defining, building and supporting community-oriented services. (Infrastructure provider,

	Expertise center, Research support organization). Each needs acknowledgement and visibility.
Proposed measures /suggestions to mitigate political issues	Direct collaborations involving all stakeholders, transparent propagation of acknowledgements (supported e.g. by standardized policies).
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	Getting a shared view on objectives and path towards it between scientists, engineers, infrastructure providers, and support organizations.
Proposed measures /suggestions to mitigate cultural issues	Early and continued involvement of all stakeholders.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	Interaction with SURFsara has been excellent as has been the support by EOSCpilot Shepherds as well.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	There is already a large number of services and tools available within the EOSC, often with overlapping functionality – it is a challenge to find and evaluate appropriate service components effectively. For each community to survey the whole landscape and evaluate all options is not practical.

	An annotated service/tool catalog including specific strengths and weaknesses, enabling communities/users to provide 'reviews' and tips would be a useful help. The surveying could be facilitated through a structured set of demo's and/or workshops. Broad support for VM/containers would be extremely useful to deal with application deployment complexity (for LOFAR, but likely for many communities). General purpose support for the CWL standards would be useful as well. What we are currently missing in our setup are appropriate workflow management and monitoring tools. We aim to look e.g. at Airflow in a future follow-up. It could be considered to provide a central EOSC linked data platform as a low-threshold entry point for communities that are new to the technology. Finally, data locality is an issue for Radio Astronomy in particular. We are limited by (network) connectivity between source and destination storage on the one hand and appropriate computational infrastructure on the other hand. This will become a serious issue when expanding our demonstrator to include multiple locations (e.g. FZJ and PSNC for LOFAR).
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	Addressed: - Portable and flexible pipeline/workflow definition & deployment using containers. - Data access and storage (FAIR data repository) - User Inerface for definition & starting pipelines. As was the intention of the demonstrator, the above issues where addressed at the level of Proof of Concept/technology experiment level. As such, they were successful and will be used to direct future activities aimed at setting up pilot and ultimately production services in all of the above areas. Not covered within EOSCpilot (status 13/11/2018): - Make workflow useable over more systems. - High bandwidth connectivity. - Federated access Of these, the latter (federated access) is going to be addressed in an extension for the LOFAR demonstrator. Workflow usability will continue to be worked on in other projects. The bandwidth issue does not currently seem to be considered an issue to be addressed at EOSC level.

Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	Scientific service enablers need to learn the new language for describing workflows. The tutorials are good, but a workshop for CWL programming would be beneficial. Users need to gain skills in developing pipelines in an abstract way. There needs to be more awareness that thinking about pipelines in an abstract way makes it easier to deploy on large scale. End users know little about containers, VM's & CWL. Training needed. Service enablers need to identify the most appropriate or generic solution(s). it is not always clear which infrastructure/service will be the best fit or is most easily adapted for a specific purpose. Information and training needed.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	In general, for the LOFAR community, it could be stated that data sensitivity is not much of an issue and our interest is in: - Facilitating open and accessible data - Ensuring IP through appropriate acknowledgments - Guidelines and support for ensuring protection of personal data in accordance with law while maintaining openness and traceability of scientific provenance.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	Yes, we agree and applaud the change of community stakeholder role from 'Advisory' to 'Steering'. Concerning Radio Astronomy, it will be useful to consider allining the governance model for the European SKA Science Data Center, as being developed e.g. in the H2020 AENEAS project, up with that for EOSC.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	- Support for containerized software deployment - IO/throughput performance - CWL Workflow execution and monitoring (for End Users and Service Enabling organizations) - Move away from user owned X509 certificates

Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	- Scalability (IO & data volume) - Usability of (grid) infrastructure - User and Service Enabler support by infrastructure providers
Other (please specify)	-

EOSCpilot: Science Demonstrator Pre-Final Report	
Date	**Final report**
Science Demonstrator Title	Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories (FDE)
Representative name, affiliation and email from proposing organisation(s)	Petr Knoth, Senior Research Fellow, KMi, The Open University, UK, petr.knoth@open.ac.uk
Main Shepherd name, affiliation and email	Thomas Zastrow, Max Planck Computing and Data Facility (MPCDF), thomas.zastrow@mpcdf.mpg.de
Secondary Shepherd name, affiliation and email	Erik van den Bergh, European Bioinformatics Institute, EBI/EMBL, evdbergh@ebi.ac.uk
General part	

Achievements so far
(> 200 words)
	○ Collected baseline data for a comparison of the ResourceSync protocol with OAI-PMH
	○ Conducted planned experiments addressing various configurations.
	  ○ OAI-PMH speed assessment across a variety of repository platforms and over 1,000 repository systems.
	  ○ Evaluation of three different ResourceSync setups (standard, batch and materialised dump) across several datasets of scientific publications
	  ○ Recall evaluation of OAI-PMH full text harvesting and their benchmark to ResourceSync harvesting.
	○ New synchronisation approach proposed to the ResourceSync community. The new approach has also been developed and tested. And the source code is released on https://github.com/oacore/rs-aggregator
	○ Research paper describing the results and the methodology of the experiments described above 90% ready for submission to JCDL 2019
	○ Conducted interview and questionnaire with the TEXTCROWD science demonstrator. A report has been written on how ResourceSync can improve the discoverability and interoperability of this Science Demonstrator tool.
	○ Conducted interview and questionnaire with High Energy Physics science demonstrator. A report has been written on how ResourceSync can help address and mitigate the Science Demonstrator challenges encountered.
	○ Supported ARC (OpenAIRE) in their efforts of adopting ResourceSync harvesting from our projects ResourceSync endpoint called Publisher Connector (http://publisher-connector.core.ac.uk/resourcesync/)
	○ Established a commercial collaboration with Naver Academic

	(https://academic.naver.com/) who now harvest data from the CORE aggregator (core.ac.uk) to Naver Academic in production using ResourceSync.
Problems encountered	• Time scalability issues of ResourceSync in scenarios requiring the synchronisation of large numbers of small files (typically metadata files). We have responded to this issue by developing a new approach we call ResourceSync Batch (On Demand Dump), described in this blog post: https://blog.core.ac.uk/2018/03/17/increasing-the-speed-of-harvesting-with-on-demand-resource-dumps/). During the course of the project, we have explored possibilities of making ResourceSync Batch part of the ResourceSync protocol specification and would like to continue this work following the end date of the project. • Issues around creating a fair benchmarking environment, considering effect of network latency, geolocation, data size, HTTP connection overhead, repository platform implementation differences, differences in average resource size. These issues have been discussed and tackled in our experimental design by: ○ Drawing comparisons only for synchronisation tasks conducted on exactly the same data ○ Planning to limit the effect of network latency on the results ○ Analysing variance in response time and average resource size of repositories per repository platform (for example, EPrints, DSpace, OJS, etc.). • In order to run the recall comparison of Full texts across different repositories, we communicated with a number of repository managers asking them to supply numbers of full texts available in their systems, to serve as a gold standard. Some replied without any issue but around 40% of them had a less clear idea on what is the number of full text resources available in systems they manage, demonstrating that the institutional repository software is not yet resource oriented and highlighting how the use of ResourceSync could also push the Repository Software producers to improve the management of their data. • Timing constraints on the implementation of the collaboration with the other Science Demonstrators. The structure of the demonstrator didn't allow a thorough implementation or testing for or with data of other SDs but only a feasibility study that will need further investigation in the future.
Data management and handling of sensitive data, with reference to plan.	We currently do not work with sensitive data. The data we experimented with are already managed as part of CORE using established processes.

List of outreach activities	• Blog post on the introduction of ResurceSync Batch[1]. This proposal has also been discussed with the creators of ResourceSync and was sent to the ResourceSync mailing lists. Digital Public Library of America (DPLA) positively responded to this effort, declaring they encountered the same problems and indicating this could be a solution for them. • Participated in the EOSC Pisa meeting. Having several discussions with other EOSC demonstrators and exploring the possible interactions with the harvesting interoperability efforts of WP6. • A submission and presentation at Open Repositories conference: Knoth, P. and Klein, M. (2018) Evaluating the performance of OAI-PMH and ResourceSync, Open Repositories 2018, Bozeman, Montana, USA • A submission and presentation at the DI4R conference: Knoth, P., Cancelliri, M. and Klein, M. (2018) Frictioneless Data Exchange Across Repositories, Digital Infrastructures for Europe, Lisbon, Portugal • Started a conversation with 4Science (https://www.4science.it/en/) about the implementation of a ResourceSync endpoint in the DSpace software. • Presented a poster at Repositories Fringe in Edinburgh about the preliminary results of the experiment. • Meeting organized in collaboration with the Next Generation Repositories working group for repository platforms to adopt ResourceSync. Key platforms, including DSpace, EPrints and Fedora participated and are working on adoption.
Specific Feedback on:	
Technical challenges and issues encountered	Listed in the problems section.
Proposed measures /suggestions to mitigate technical issues	Key technical challenges have been addressed by the demonstrator. One improvement created by the demonstrator requires a change/extension in the protocol specification.
Political challenges encountered	None
Proposed measures /suggestions to mitigate political issues	N/A

[1] https://blog.core.ac.uk/2018/03/17/increasing-the-speed-of-harvesting-with-on-demand-resource-dumps/

Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	The main challenge encountered has been the outreach to repository managers and institutions to move forward with the implementation of ResourceSync. Mainly in the data collection instance, we realized that in some institutions there is a knowledge and skills gap between between technologists who implement repositories and those who manage it. This is an issue when it comes to integration of ResourceSync in different platforms.
Proposed measures /suggestions to mitigate cultural issues	Outreach activities on promoting the use of a new protocol are needed at all levels of the decision chain and must inlcude repository managers, repository platforms and aggregators developers, architecture decision makers.
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	The stage of the demonstrator is to prove is validity and finally become widely adopted. The interaction with the Repository Managers has been positive and there is a wide interest in bringing innovation on 20 years old protocols. The New Generation Repositories working group is including ResourceSync in its recommendations and possibly will impact on the recognition of innovation in this field to promote interoperability.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The Science Demonstrator worked on a basic client implementation created by the protocol creator. By applying the concepts at a wider scale it was immediately clear that the protocol/implementation of the protocol needed some changes to support scalability. The changes have been implemented in a new client and a proposal to extend the protocol has been started.

Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	The experience built by running this Science demonstrator proved that ResourceSync is a valid approach on improving the interoperability at scale. Adoption from big providers is the next thing to pursue. Policy push and recommendations from across the sector would help in its wider adoption.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	There are multiple ways of adopting ResourceSync, but not all will provide scalable solutions. Adopters need better information and guides that will help them to make the right choices in deciding which aspects of the protocol they should support. This depends on the amount of their data, frequency of updates and synchronisation use case.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	Recommend ResourceSync as a default data and metadata exchange protocol for all repositories operating within EOSC.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	No opinion, need to see it in action.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	In general, the protocol can be applied in any digital repository to facilitate the synchronization. The main requirements are about disk space and processing power, in case of big volumes of data this could be challenging and correct trade-off needs to be found. Moreover, most of the data hosted in the EOSC services should have a metadata definition that could be shared to enable an easier distribution and discoverability.

Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	There is a need on creating a distributed/federated infrastructure where all the components should easily communicate and there should be any major difficulty for other components to join in. Using a widely used, well tested/benchmarked and adopted standard instead of creating new one could help in fostering many of EOSC's processes. We would like to see ResourceSync adopted within EOSC for the purposes of helping the integration and interoperation of distributed systems and we hope that the steps (including benchmarking and the produced source code) we have conducted in the SD will be useful for the initiation of such processes.
Other (please specify)	

EOSCpilot: Science Demonstrator Pre-Final Report	
Date	**November 2018**
Science Demonstrator Title	Bioimaging
Representative name, affiliation and email from proposing organisation(s)	Dr Jean-Karim Hériché, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany heriche@embl.de
Main Shepherd name, affiliation and email	Gianni Dalla Torre, EMBL-EBI, Hinxton, UK gianni@ebi.ac.uk
Secondary Shepherd name, affiliation and email	Tony Wildish, EMBL-EBI, Hinxton, UK wildish@ebi.ac.uk
General part	
Achievements so far (> 200 words)	The Bioimaging EOSCpilot science demonstrator has two goals. The first one is the prediction of new cellular functions for human genes using publicly available image data from genome-scale loss of function experiments. The second one is to assess how the EOSC could be used for image analysis tasks. In this respect, the project's computational needs are fairly representative of a large class of image analysis problems requiring feature extraction from images followed by application of machine learning algorithms. In a first stage, the project has extracted and further processed thousands of features from over 2 million images using the Embassy Cloud, reducing the data size from ~5 TB to 45 GB. To parallelize computation in the first stage of the project, a cluster of 192 vCPUs, each with 4 GB RAM,

	was created. The second stage of computation started using a different configuration with 24 vCPUs but with 32 GB of RAM. While computing on one particularly large dataset, the amount of RAM available in the cloud became limiting and the lack of a high-performance file system prevented implementation of alternative solutions. The issue was resolved by moving the second part of the computation to the local data center at EMBL Heidelberg, where we are currently running the final analyses to evaluate the results.
Problems encountered	See below specific feedback.
Data management and handling of sensitive data, with reference to plan.	N/A
List of outreach activities	None so far.
Specific Feedback on:	

Technical challenges and issues encountered	As the main computation phase has finished, we can make several observations concerning the use of the cloud for image analysis: 1- Setting up and configuring a tenancy in the cloud requires a level of IT expertise beyond what can be expected from most computational biologists and is definitely beyond the reach of cell biologists who would only occasionally need to run image analysis tasks. In fact, the project wouldn't even have been able to start and subsequently run without the excellent and regular support of EOSCpilot shepherds. In particular, several stumbling blocks were identified: - Defining the parameters of the tenancy for the project was not easy and took time. For example, how many vCPUs and how much RAM they should have are parameters that depend both on the underlying cloud technology (e.g. what hardware is available, how fast it is...), on the software used (how efficient it is both in terms of memory use and computation speed) and on the data to be processed. - Creating a compute cluster and deploying software across all nodes are tasks that are typically done by the IT team supporting an HPC but which in the cloud are transferred to the user. Having to write and run scripts with obscure parameters is going to be a major obstacle to adoption. 2- Defining the deployment size of a data processing pipeline can be difficult because requirements are not always comparable from one data set to another. A lot of computation time has been spent on various issues including adapting and testing the code to run in the cloud environment. Although cost wasn't taken into account for this project, we surmise that this could become prohibitively expensive depending on the cost model adopted for cloud ac-

	cess. 3- The cloud environment is not as flexible as a local HPC: - Meeting the requirements of the different parts of the computation requires a complete tear-down and recreation of the cluster whereas, in a local HPC, one can just specify the resource requirements for each job. The alternative is over-specification of the cloud tenancy requirements to cover all needs. - The cloud doesn't offer a high performance shared file system which means that I/O-bound operations need to be rethought and substantially reworked compared to a local HPC with such a file system. 4- Many standard image analysis software require image data to be in files on the local file system. Most of this software would need to be rewritten to get their input directly from object storage as typically used in the cloud. The alternative is to add a preprocessing step to save object storage data to a local file which adds overhead to all computation but could be mitigated by availability of a high-performance file system.

Proposed measures /suggestions to mitigate technical issues	1- User-friendly and user accessible platform to set up and provision nodes in a cloud environment targeted at non-specialist users and streamlined and quick interaction with IT support staff. 2- Because defining the deployment size of an established data processing pipeline can be difficult, a generous cost model that doesn't penalize testing of pipelines and overprovisioning of tenancies should be considered. 3- Provisioning of high performance shared file system for I/O bound operations. 4- Standardized environment and tools for deploying software.
Political challenges encountered	N/A
Proposed measures /suggestions to mitigate political issues	
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	N/A
Proposed measures /suggestions to mitigate cultural issues	

Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	Remoteness of IT support team (i.e. shepherds) affected communication. In particular, response time was much slower than with our local IT support team.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	N/A
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	N/A
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	Setting up and configuring a tenancy in the cloud requires a level of IT expertise beyond what can be expected from most computational biologists and is definitely beyond the reach of cell biologists who only occasionally need to run image analysis tasks. - Provide more training to non-expert end users. - Provide user-friendly services targeting non-expert users.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	

Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf)? Please comment if necessary.	The current high-level wording makes it hard to understand the concrete consequences for end-users.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	Object storage access for computation in a cluster of vCPUs with shared high-performance file system.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	Non-expert user-friendly tenancy set-up, including software deployment.
Other (please specify)	

EOSCpilot: VisIVO Science Demonstrator Final Report	
Date	**2018-Nov-7 / Final report (12 Months)**
Science Demonstrator Title	**VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics**
Representative name, affiliation and email from proposing organisation(s)	INAF Astrophysical Observatory of Catania Alessandro Costa, alessandro.costa@oact.inaf.it Fabio Vitello, fabio.vitello@oact.inaf.it Eva Sciacca, eva.sciacca@oact.inaf.it Antonio Calanducci, antonio.calanducci@inaf.it Ugo Becciani, ugo.becciani@oact.inaf.it INAF IAPS Sergio Molinari, molinari@iaps.inaf.it
Main Shepherd name, affiliation and email	Michael Schuh, DESY michael.schuh@desy.de
Secondary Shepherd name, affiliation and email	N/A
General part	
Achievements so far (> 200 words)	The Astrophysical community has set up a new suite of cutting-edge Milky Way surveys that provide a homogeneous coverage of the entire Galactic Plane and that have already started to transform the view of our Galaxy as a global star formation engine (http://vialactea.iaps.inaf.it). New instruments have delivered information of unprecedented depth and spatial detail spanning the electromagnetic spectrum. This Science Demonstrator is aimed at the integration in the EOSCpilot e-infrastructure of the visual analytics environment based on VisIVO (Visualization Interface for the Virtual Observatory) (http://visivo.oact.inaf.it/) and its module VLVA (ViaLactea Visual Analytics). The VLVA application has integrated European Open Science Cloud (EOSC) technologies for the archive services and intensive analysis employing the connection with the ViaLactea Science Gateway (https://vialactea-sg.oact.inaf.it/). The archiving services are being deployed within the EGI Federated Cloud toward the assurance of a FAIR access to the surveys data and related metadata. The science gateway has been integrated with the EGI Check-in (https://www.egi.eu/services/check-in/) service to enable the connection from the federated Identity Providers and with the EGI Federated Cloud (https://www.egi.eu/services/cloud-compute/) to expand

	the computing capabilities making use of a dedicated virtual appliance stored into the EGI Applications Database (https://appdb.egi.eu/store/vappliance/visivo.sd.va).
Problems encountered	During the development we found some bugs in the WS-PGRADE/gUSE gateway framework for the connection with the EGI Federated Cloud and the EGI Check-in. The problems related to the connection with the cloud has been solved by fixing some configuration details and upgrading some software libraries. While the problems related to the connection with the federated log-in services have been solved by modifying/customizing the portal plugin of the gateway (see https://github.com/evasciacca/OpenIdConnectLiferay/tree/EGICheckIn). Furthermore, we developed our own lightweight gateway with an ad hoc RESTful APIs to expose a simple set of functionalities to define pipelines and executing scientific workflows on any Clouds resources, hiding all the details of the underlying infrastructures. As it's not straightforward to use Federated Identity from a desktop application, we added a local user management system to our lightweight service. In turn, we route the requests via gUSE to the Cloud computing resources, making use behind the scene of a Robot certificate stored in a secure eTokenServer. We keep track of any request coming from the VLVA desktop client logging actions and users in our service DB.
Data management and handling of sensitive data, with reference to plan.	As planned only public input data are deployed to the EGI Data Resources to be reached from the VisIVO ViaLactea application.
List of outreach activities	The Science Demonstrator has been presented by Fabio Vitello during the EOSCpilot All Hands Meeting, Pisa, Italy (8-9 March 2018). The Science Demonstrator was also presented by Alessandro Costa during the INAF ICT Meeting, Catania, Italy (10-14 September 2018). A poster entitled "Data Knowledge Visual Analytics Framework for Astrophysics within EOSC" has been accepted to be presented by Eva Sciacca during the conference DI4R, Lisbon, Portugal (9-11 October 2018). A poster entitled "VisIVO Visual Analytics Tool: an EOSC Science Demonstrator for data discovery" will be presented by Ugo Becciani during the ADASS 2018 conference, Maryland, USA, (11-15 November, 2018).

	A poster and a light presentation will be given by Eva Sciacca during the EOSC Stakeholder event, Wien (21 November, 2018) EOSCpilot has been acknowledged in the following publication: ● Vitello, F., et al. "Vialactea Visual Analytics Tool for Star Formation Studies of the Galactic Plane." *Publications of the Astronomical Society of the Pacific* 130.990 (2018): 084503.
Specific Feedback on:	
Technical challenges and issues encountered	There has been a delay on the VM configuration for deploying the data resources needed by the science demonstrator due to CESNET-MetaCloud technical problems.
Proposed measures /suggestions to mitigate technical issues	It would be useful to plan measures for ensuring e-infrastructure reliability.
Political challenges encountered	We were planning to make usage of some cloud storage to save VLVA users' sessions data. It seems that EGI FedCloud Virtual Organizazion doesn't not provide any official object storage service at the moment. We made a request to the EGI Task Force mailing list for 100GB of storage (for 6 months) and one of the partner replied, offering storage for a monthly cost, motivating the request as acknowledged in any dissemination activity related to the EOSCPilot project.
Proposed measures /suggestions to mitigate political issues	EOSC infrastructure providers should be prepared to operate at much larger scales than the ones demonstrated within the context of the EOSCPilot project. The astrophysics community handle massive data and petabytes of data are expected from future space missions and ground based facilities such as the Cherenkov Telescope Array or the Square Kilometer Array.
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	N/A

Proposed measures /suggestions to mitigate cultural issues	
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	The problems and issues encountered have been tackled and solved thanks to the suggestions of shepherds, EGI staff and technical people involved in the development of the technologies and software employed by the Science Demonstrator. We have also profited from the experiences, results and knowledge of other previous science demonstrators (e.g. EPOS-VERCE for the gateway services and the connection with EGI FedCloud and the EGI-CheckIn).
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	Many services and service catalogues are already available. We have so far evaluated mainly the EGI service catalogue but it would be convenient to have an interface to help evaluating each service based on previous experiences, user ratings and tutorial material.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	N/A
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	All the technical and configuration details related to the use of the EOSC e-infrastructure should be completely hidden for the end users, since astrophysicists would like to focus only on the scientific results. The scientific service enablers should be trained for the configuration of the specific EOSC services they would like to employ and should be advertised of new services that would potentially impact their scientific community.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the	N/A

policy document of WP3	
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf)? Please comment if necessary.	N/A
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	- integrate visual analytics services; - increase the access of computing resources for analysis; - provide services for a federated and interoperable virtual environment enabling collaboration and re-use of data and knowledge; - optimization of the archiving of complex data (and meta-data) such as the multi-wavelength surveys under FAIR principles.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	- Usability, enhance usability for end users - Accessibility and Integrability to facilitate the integration of new components coming from the scientific community within the EOSC - Documentation - Failure management, to help also debugging failing computations.
Other (please specify)	

EOSCpilot: Science Demonstrator Pre-Final Report	
Date	**2018-Nov-29**
Science Demonstrator Title	FAIRifying eWaterCycle and SWITCH-ON
Representative name, affiliation and email from proposing organisation(s)	Rolf Hut, TU Delft, r.w.hut@tudelft.nl Nick van de Giesen, TU Delft, n.c.vandegiesen@tudelft.nl Jerom Aerts, TU Delft, j.p.m.aerts@tudelft.nl Niels Drost, NLeSC, n.drost@esciencecenter.nl Berend Weel, NLeSC, b.weel@esciencecenter.nl Gijs van den Oord, NLeSC, g.vandenoord@esciencecenter.nl Inti Pelupessy, NLeSC, i.pelupessy@esciencecenter.nl Maarten van Meersbergen, NLeSC, m.vanmeersbergen@esciencecenter.nl Martine de Vos, NLeSC, m.devos@esciencecenter.nl Ronald van Haren, NLeSC, r.vanharen@esciencecenter.nl Stefan Verhoeven, NLeSC, s.verhoeven@esciencecenter.nl Yifat Dzigan, NLeSC, y.dzigan@esciencecenter.nl Vincent van Hees, NLeSC, v.vanhees@esciencecenter.nl Berit Arheimer, SMHI, berit.arheimer@smhi.se Michael R. Crusoe, VIDE, mrc@darboeigos.eu Axel Berg, SURFsara, axel.berg@surfsara.nl Sergio Andreozzi, EGI, sergio.andreozzi@egi.eu Lukasz Dutka, Cyfronet, lukasz.dutka@cyfronet.pl Anton Frank, LRZ, anton.frank@lrz.de
Main Shepherd name, affiliation and email	Gianni Dalla Torre, EBI, gianni@ebi.ac.uk
Secondary Shepherd name, affiliation and email	Tony Wildish, EBI, wildish@ebi.ac.uk
General part	
Achievements so far (> 200 words)	Central to the science of hydrology is the localised nature of the medium through which water flows. This fact leads to a large amount of hydrological models, specifically made for a certain region (catchment).

	The fact that these models are made by different individuals, in different programming languages severely hinders re-use and reproducibility. This in turn is leading to a "crisis of reproducibility" in Hydrology. This science demonstrator seeks to create a fully FAIR Hydrological forecasting system, combining local and global models. With this we showcase and hope to demonstrate best practices on how data as well as software can be made FAIR in Hydrology. We started by doing an assessment of the various input data sets required for our project. Even determining the FAIRness of a dataset proved problematic, as critical information needed for this assessment such as guarantees of long term availability of a repository is often not publicised. In the end we decided that creating a local copy of needed input data was the best course of action, though recognize this cannot be considered FAIR yet. Using a combination of the CWL standard for workflows, Cylc, and Docker software containers, we were able to create a fully reproducible (low resolution) version of the eWaterCycle forecast. Output data is stored via OneData, and available for analysis in a notebook environment, as well as visualization in a web application. The forecast is running daily now, without any manual intervention. We plan to deploy the system to one or more supercomputer systems, and increase the resolution of the model. On this version we can do a verification of model output to assess its scientific merit. In addition we plan to integrate the work done in this demonstrator into our larger eWaterCycle II project, with a focus on FAIR Hydrological modelling https://www.ewatercycle.org/
Problems encountered	FAIRness (or lack thereof) of input data blocked progress. Attempts to improve FAIRness of datasets from third parties led to nothing. We explicitly tried to improve FAIRness without involving the original author, as involving authors of these existing datasets, often created by large organizations, is not something possible within the scope of a small projects such as ours, and in general not something individual scientists have the resources for. The biggest problem faced in making existing datasets FAIR was a lack of authoritative metadata. For instance, for one dataset it was unclear if two versions found online are the same, or different variants of the same model. Another problem faced is that improving accessibility of a dataset often requires redistributing a dataset, something not always allowed under the licence. Or, even worse, it is sometimes unclear if this is allowed.

	If input data is not Open Data, this greatly hinders what can be done in terms of processing this data. E.g. no automated checking of data availability and quality, and caching of data problematic. Only very low-level services seem to be part of EOSC at this time (VMs, clusters, etc); but our needs are for higher level services OneData system is a really nice concept, but is missing some key usability features. One example is the token based invite system. Being able to invite users via email as is more customary would save users from copy-pasting long tokens. OneData also needs to be available out-of-the-box on platforms used. Running the High-resolution forecast is still work in progress, as the demonstrator started so late due to administrative issues: the contracts were signed late by the consortium and the SMEs in our demonstrator can only start working when they also get paid, they lack the financial buffer that a university or large institute might have.
Data management and handling of sensitive data, with reference to plan.	The demonstrator uses a number of external datasets. The results of the forecast will be published using Zenodo. No sensitive data is used or produced by this demonstrator.
List of outreach activities	Attendance to the following EOSC related events: ● EOSCPilot all hands meeting, March 2018 ● Second EOSC EOSC Stakeholder Forum, November 2018 As our demonstrator is only now nearing completion, we expect more outreach activities in the near future.
Specific Feedback on:	
Technical challenges and issues encountered	This demonstrator makes use of the OneData system. Specifically, a storage node at CNAF. We've had some stability issues with this experimental node. The reason we used OneData was to have easy access to our data from all our environments. For some, installing software and creating mounts is problematic, so for those we would like to have a client application rather than installing OneData into the system. From the documentation of OneData we gathered OneData solely relies on FUSE mounts. This made it tricky to get going on the infrastructure we use, and we ended up installing OneData as a system service on the cloud infrastructure we are using at SURFsara. Not having up to date packages for common Linux distributions was problematic. As a result of writing this report we have found out that access via HTTP is

	possible via the CDMI protocol. We are planning to try this out in the future, and would like to suggest to mention this in the OneData user documentation. It is currently hidden in the “advanced usages” section, while we consider downloading a file a basic usage of the system. In general, the lack of a coherent set of high level services meant we had to build our entire demonstrator from the VM-level up.
Proposed measures /suggestions to mitigate technical issues	In our opinion, EOSC is in need of a coherent set of high level services aimed at researchers. The current set of services are too disjoint to be useful. Some prime examples of services we would like to see: ● File sharing (Dropbox for large data sets) ● Execution services for standards-based workflows ● Authentication and Authorization services ● Persistent storage (with identifiers like DOIs)
Political challenges encountered	In general, we struggled with getting our input datasets into a FAIR state. As these were not created by us, attempts to make these FAIR led to nothing. Most of the input dataset required for our research are not open data. This makes it impossible to build open science on. Most of the communication from the project (with the exception of the Shepherds) has been procedural, perhaps as a result of our demonstrator starting late.
Proposed measures /suggestions to mitigate political issues	The EOSC should actively engage dataset providers for creating a FAIR version of each dataset. Individual researchers are incapable of doing this themselves, mostly due to lack of time. Creating Open, not only FAIR, datasets should be one of the goals of the EOSC. The EOSC should focus more on creating a community of users and providers, including intermediate roles such as RSEs and Data Stewards.
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	e-infrastructure providers in general do not speak the same language as scientists. In this project we did not encounter this issue as a number of Research Software Engineers (RSEs) were responsible for much of the technical work.

Proposed measures /suggestions to mitigate cultural issues	It is important to recognize the intermediary roles required for the EOSC to be successful, such as Data stewards and Research Software Engineers (RSEs).
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	There was little direct communication between the science community and the infrastructure providers, as this was all done via the RSEs working on the project.
Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The WP5 service catalogue was not ready when we were creating the technical design of our demonstrator.
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	It would have helped greatly if support for services such as OneData was present on a number of computing facilities. As it stands now, the integration of services seems to all be up to the user of the EOSC.
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	None encountered.

Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	No comment.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	The governance seems adequate for the EOSC. We do stress the need to actively engage with and react to real user's needs for the EOSC.
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	Our demonstrator is in need of a set of high level services: ● File sharing (Dropbox for large data sets) ● Execution services for standards-based workflows ● Authentication and Authorization services ● Persistent storage (with identifiers like DOIs)
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	Support for researchers in the form of Data stewards, Research Software Engineers (RSEs), and other roles, is as important as the (compute) services offered.
Other (please specify)	

EOSCpilot: Science Demonstrator Pre-Final Report	
Date	**2018-Nov-7**
Science Demonstrator Title	***Visual Media:*** *a service for sharing and presenting visual media files on the web*
Representative name, affiliation and email from proposing organisation(s)	ISTI-CNR (IT): Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione Roberto Scopigno, r.scopigno@isti.cnr.it PIN (IT): Polo Universitario Città di Prato Franco Niccolucci, franco.niccolucci@gmail.com MIBACT – ICCU (IT): Ministero dei Beni e delle Attivita' Culturali, Istituto Centrale per il Catalogo Unico Sara di Giorgio, sara.digiorgio@beniculturali.it
Main Shepherd name, affiliation and email	Thomas Zastrow, MPCDF zastrow@mpcdf.mpg.de
Secondary Shepherd name, affiliation and email	Erik van den Bergh, EBI evdbergh@ebi.ac.uk
General part	
Achievements so far (> 200 words)	A workplan has been defined at the beginning of the project (M1), in collaboration with the science demonstrator partners, and delivered on Dec. 18th, 2017. We organised a meeting with our partners to discuss the structure and content of the metadata gathered for each visual data file uploaded by the users; we have slightly revised our original metadata organization and this will be implemented in the Visual Media server web forms (the ones enabling data upload). We had one in-person meeting with the D4SCience staff and several further contacts (phone, skype) to plan the activities of integration of the D4Science facilities. The planned work was divided in three phases: adding authentication, personal storage management and scalability (of the pre-processing tasks, if needed).

	We have incorporated the authentication feature of D4Science. Implementation work started in January 2018 and we finalized the implementation of the authentication in the following month. Authentication included support of: D4Science authentication, Google authentication and a password-less login option (based on email). Managing the authentication was not just incorporating the initial login and authentication dialog; conversely, it required several changes to the internal organization of the Visual Media Service and its functionalities/interfaces (e.g. possibility of presenting visually only the data owned by a specific user, notifications about the job status through the D4Science 'social' system). Authentication features have been tested in the following period. After having introduced the authentication, we have implemented a few small services connected with the now available user data. As an example, we have provided a web page for allowing the single user to define his own profile (email notification, institution, etc). We did also several changes to the code of the Visual Media Server. Among those are the following: • the internal 3D browsing visualizer has been updated with the last version of the **Nexus** data conversion and rendering engine; • we changed the **format** adopted to manage **RTI data**, incorporating the new Relight format recently proposed by CNR-ISTI (published at Web3D 2018 Conf. in last June). This new format enables: better data encoding (improved space/quality ratio), enables a standard multi resolution format based on image pyramids (following the Google maps approach), is compatible with deepzoom and zoomify formats, supports IIP and IIIF for interoperability; • we have introduced the automatic production of **thumbnails** and use them in the viz interface and in the browsing interface; • we added an option to support the **online configuration** of the visualization tools. We worked on the data browsing component of the Visual Media Service, to provide an option to access/visualize just

	the proprietary data of the specific logged user. We also included the possibility to load visual media data (data files) from the personal storage area of any specific user (from the D4Science user workspace). Finally, we are published a link to the Visual Media Service on the project web (https://eoscpilot.eu/social-sciences-and-humanities-visualmedia-service-sharing-and-visualizing-visual-media-files-web), to make it accessible to EOSC users. User testing activities started on September (we asked to users to upload new data material and to test the revised and extended service). This activity progressed quite well in Sept-Oct. 2018, with many new datasets uploaded by users and several bugs notifications (which were promptly corrected). The overall evaluation of the users was very good. We published two papers with acknowledgement to EOSC: Federico Ponchio, Massimiliano Corsini, Roberto Scopigno, "A Compact Representation of Relightable Images for the Web", Proc. ACM Web3D (International Symposium on 3D Web Technology), 20-22 June 2018, Poznań, Poland, page 10 - 2018 - **Best Paper Award** at the confeerence. M. Potenziani, M. Callieri, M. Dellepiane, R. Scopigno,"Publishing and Consuming 3D Content on the Web, a Survey", Foundations and Trends in Computer Graphics and Vision, Now Publishers Vol.11(?), pp.95, 2018 (in press)
Problems encountered	No mayor problems to be mentioned. The activities proposed in the initial workplan have been implemented according with the initial time planning. The implementation of the new features was successful.
Data management and handling of sensitive data, with reference to plan.	New datasets have been requested to users in the final test and assessment period (Sept-Oct.). The VM Service includes some disclaimer related to data ownership and right to use. We think that the personal data processing policy of Visual Media Service is compliant with GDPR.
List of outreach activities	We have presented Visual Media Service to potential users in several events/conferences.

	Two scientific papers have been published (see above); the first paper was awarded the "Best Paper Award" at the ACM Web3D conference and invited for submission to a journal (the revised paper submission is under way).
Specific Feedback on:	
Technical challenges and issues encountered	The instruments used (mostly D4Science) were sufficiently easy to incorporate in our existing software and worked fine. The support of the responsible people was very good (both in personal communications and supported code examples / software documentation). Supporting and enhancing the availability of FAIR data is a key issue in many scientific domains. The Visual Media Service tries to propose a solution to some key issues: how to manage the multiplicity of visual data types, the many visualization applications available/needed for each data type, the complexity of publishing those data on the web to allow easier sharing between experts. Instead of working at the data standardization level, we decided to work at the data tools level (offer a single generic tool to allow the community to share and visualize visual data on the web).
Proposed measures /suggestions to mitigate technical issues	none to report
Political challenges encountered	The main political challenge in the Cultural Heritage domain is to convince potential users of the added value of uploading visual media results and to make them available to the community on the web. Visual Media Service has been mainly designed for the Cultural Heritage community. Many people in this community are still very protective with the data they produce or are working with. The goal is therefore to convince them that the added value of sharing with remote colleagues the data is more important than the possible risk of having unique access over those data or to lose control of the data.
Proposed measures /suggestions to mitigate political issues	We have discussed quite extensively in the last few years on the issue of which type of reward could/should be given to scientists which accept to share data. There is nowadays no reward or incentive to increase the quantity (and quality) of the data shared. Visual data are a

	basic and very critical resource in the scientific process, in many disciplines (Cultural Heritage is a good representative). We need some incentive to foster data sharing (for example, considering quantity and quality of data shared as one of the items evaluated in a CV in promotion or hiring processes).
Cultural challenges encountered (including communication challenges between science communities and e-infrastructure providers)	The offer of standard e-infrastructures is often too much oriented to ICT developers (offering basic technologies). If we consider the specific Cultural Heritage domain/community, then most of the potential users are looking for **ready to use solutions/tools** rather than libraries to be used for SW development.
Proposed measures /suggestions to mitigate cultural issues	Increase the number of tools or services built on top of basic infrastructure resources (I think that the EOSC SD follow in some sense this policy).
Interaction between Science community and infrastructure providers: What worked well, what not so well, suggestions how to facilitate and streamline this process	The interaction between the CH community and the infrastructure providers was mediated by the SD proponents (ourselves). We managed internally the contacts with the users, which used a service which was enriched by some functionalities derived by the EOSC infrastructure. The interaction between this SD and the infrastructure providers (as far concerns the resources used in this SD) was fine. More generally, it is still not easy to convince Cultural Heritage experts to check/use/deploy the features supported by standard ICT infrastructure providers. This is due to cultural divides (not easy to orient in web sites usually written for ICT people, or to use resources which require a medium or high-level skill in computing). In this sense, we think it is more useful and easier to adopt if infrastructures for CH people offer medium-level services rather than low-level libraries. The Visual Media Service is an example of such a class of final-users services (not oriented to SW developers, but to final users).

Services and service catalogue (WP5): What worked well, and what are missing functionalities or services	The description available on the project web on the service portfolio could be improved, maybe organizing the services by groups and giving to the final users some easier instruments for searching on the project web interesting services/functionalities (possibly, avoiding to redirect the reader to very long documents or project reports).
Interoperability issues: (WP6): What has been addressed and how well, what remains to be addressed?	We have considered some interoperability issues related to the visual data accessibility trough different tools/architectures/web services. Concerning the easier case (2D images) we are following the approach proposed by IFFF (International Image Interoperability Framework, https://iiif.io/).
Skills issues: (WP7): Where do you see deficits in education and training? What are your suggestions? Please differentiate between end users and scientific service enablers	Our experience is that training is an important and critical activity for any infrastructure. In other projects we have organized thematic summer school (1-week duration) to skill the potential users on the tools or resources developed by the infrastructure. This can also be beneficial in helping the users to endorse approaches based on a data sharing policy. We think that the training offer should be enforced in a possible future incarnation of the EOSC project.
Policy issues: (WP3): What areas should be addressed with priority with respect to this science area; comments on the policy document of WP3	Some effort in encompassing the issues related to visual data types and standards, as well as IPR issues and polices.
Government issues: Do you agree with the governance framework proposed by WP2 (https://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf, Executive Summary, page 6)? Please comment if necessary.	We are mostly in favor, we agree it is really critical and important to adopt a stakeholder-driven approach (see sect. 4.1. of the document referred on the left).
Functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Functional_requirement)	Maybe a more clear and synthetic description of the EOSC project and resources would have been helpful for groups entering the project from scratch, just for the implementation of a SD.

	Given the small time and effort given to each SD, it is quite complex to understand fully the structure of the project, what is the scope of the many WPs, which are the resources, and to orient yourself in the framework of a big project. Support by EOSC pilot shepherds was useful.
Non-functional requirements from your Science Demonstrator to drive and prioritize the integration of the EOSC services (https://en.wikipedia.org/wiki/Non-functional_requirement)	Development of standardized policies around data sharing. Development of standardized polices to visualize and analyse visual data. Stability in time of the resources shared by the infrastructure.
Other (please specify)	Our service will become part of a new EC project, **ARIADNE Plus** (now in negotiation phase). This is a quite large project, with more than 40 partners from the archaeology domain. Thus, a quite intensive use of the Visual Media service is expected in this project. We therefore plan to monitor our service usage of resources in the course of 2019 and, according to possible processing and storage bottlenecks, we will evaluate the possibility to use other EOSC resources (this time considering data and processing replication on different servers).