D4.3: Consolidated Science Demonstrator progress report

Author(s)	Hermann Lederer (MPG) Steven Newhouse (EMBL-EBI)
Status	Draft/Review/Approval/**Final**
Version	V1.0
Date	20/12/2017

Dissemination Level

[X] PU: Public

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. To achieve this goal, a selection process for Science Demonstrators through Open Calls has been developed, followed by the execution of two Open Calls. Following five initial pre-selected Science Demonstrators, ten additional Science Demonstrators have been selected through two rounds of scientific and technical review processes. An Engagement Model for Science Demonstrators has been developed and applied. Science Demonstrators have been introduced into their roles, and continuous shepherding of Science Demonstrators during their 12 months project execution has been carried out. Shepherding activities and project progress has been monitored in regular meetings. Feedback was collected in reports which are available for all work packages.

[ ] 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)	Hermann Lederer (MPG) Steven Newhouse (EMBL-EBI)
Contributor(s)	John Kennedy (MPG)
Due date	01/01/2018
Actual submission date	20/12/2017
Reviewed by	Natalia Manola (ATHENA), Brian Matthews (STFC)
Approved by	Juan Bicarregui (STFC)
Start date of Project	01/01/2017
Duration	24 months

Versioning and contribution history

Version	Date	Authors	Notes
0.1	19/10/2017	Hermann Lederer (MPG)	Structure
0.2	03/11/2017	Hermann Lederer (MPG)	First draft
0.3	06/11/2017	John Kennedy (MPG)	corrections
0.4	07/11/2017	Hermann Lederer (MPG)	Annex B added, C partially
0.5	13/11/2017	Steven Newhouse (EMBL-EBI)	Minor corrections
0.6	15/12/2017	Hermann Lederer (MPG)	Inclusion of Science Demonstrator descriptions in annex, addressing review of Brian Matthews
0.7	18/12/2017	Hermann Lederer (MPG)	Enhanced conclusion section, addressing review of Natalia Manola
1.0	20/12/2017	Hermann Lederer (MPG)	Polished and formatted

TABLE OF CONTENT

1. EXECUTIVE SUMMARY

The Science Demonstrators play an essential role as early adopters of EOSC, and 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 the proposed governance structures provide the guidelines needed by researchers.

To achieve this goal, structures and processes have been developed that serve as guidelines to select and execute Science Demonstrator projects within EOSCpilot. As a very early task, a selection process for Science Demonstrators through Open Calls has been developed. According to this process, two Open Calls for new Science Demonstrators have been carried out in April and September 2017 which attracted 30 and 26 proposals, respectively. Through scientific and technical review processes followed by a prioritisation step to achieve a balanced representation of science areas and communities, 10 new Science Demonstrators have been selected (five per Open Call). To be able to benefit from Science Demonstrators as envisaged as objectives, an Engagement Model for Science Demonstrators has been developed as guideline for the structuring of the Science Demonstrator activities and their interactions with the different EOSCpilot work packages. New Science Demonstrators have been introduced to their tasks, duties and the guidance provided, and shepherding activities of Science Demonstrators have been started and continuously carried out during their 12 months project execution. In monthly meetings the Science Demonstrator and shepherd activities and respective project progress has been monitored. Reports from Science Demonstrators have been collected after 4 and 8 months, respectively. Final reports are expected after 12 months. These will be available in the final report on Science Demonstrators, together with the respective conclusions.

2. INTRODUCTION

Science Demonstrators are early adopters of EOSC, selected from across a range of science areas. These early adopters help 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 requirements are used to drive and prioritize the integration of the candidate EOSC services to meet the functional and non-functional needs of researchers and to ensure that the proposed technical governance structures will put in place sufficient control mechanisms to satisfy the researcher's needs.

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 in this context.

The functional and non-functional requirements gathered from the Science Demonstrators will be used to drive and prioritize the integration of the EOSC services to meet the needs of researchers across Europe.

The purpose of this document is to provide an interim report on the progress of WP4 Science Demonstrators towards the goal of fostering the implementation of a European Open Science Cloud capable catering for the needs of Open Science in Europe.

The deliverable is organized in five main sections and three annexes. Section 2 describes the progress on the execution of Open Calls for new Science Demonstrators, Section 3 gives details about the newly selected Science Demonstrators, Section 4 describes the shepherding and monitoring of the activities of Science Demonstrators, and Section 5 contains conclusions.

Annex A contains a glossary, while Annex B contains the 8 months progress reports of the first five Science Demonstrators which started in Jan 2017, and Annex C contains the 4 months progress reports of the second five Science Demonstrators which started in July 2017. Annex D contains the background information and the essentials of all planned 15 Science Demonstrator activities.

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. OPEN CALLS FOR NEW SCIENCE DEMONSTRATORS

Science Demonstrators need to satisfy a number of the following criteria, according to the DoA:

- have a strong and well defined scientific challenge addressed by the use of e-infrastructure (e.g. an explicit data analysis challenge or supporting the operation of a Virtual Research Environment, or the dissemination and sharing of data and other research outputs) that will show the scientific excellence and societal impact that could be achieved by EOSC;
- provide data integration, management, interoperability and analysis challenges that will drive the specification and development of services within the EOSC ecosystem that also support open science;
- be representative of a broader scenario that, when established in EOSC, will have impact across Europe and beyond;
- be supported by mature research infrastructures and/or research organizations at a European/National level that will be the long-term consumers of the EOSC;
- commit to publishing or consuming third-party research artefacts (e.g. publications, datasets, tools, workflows) as part of the Open Science model, with the application of FAIR principles, and also as part of the EC Open Research Data Pilot;
- be mature and has demonstrated to be working at scale on existing computational/data/connectivity and other infrastructures (e.g. private, national, European or public clouds/grids/HTC/HPC/network resources) that would become part of the EOSC.

Five Science Demonstrators had been pre-selected through a call prior to the start of EOSCpilot. For the selection of 10 more Science Demonstrators, a procedure for carrying out Open Calls has been developed at project start (see EOSCpilot deliverable D 4.1 for the details). According to this procedure, two Open Calls have been carried out: the first in April, the second in Aug/Sep 2017.

The first Open Call resulted in 30, the second call in 26 proposals. All proposals were subject to technical and scientific review processes with scientific reviewers from the Scientific Advisory Board and technical reviewers from a pool of experts (shepherds, see chapter 4). Many more highly ranked proposals were received than could be supported. A prioritization step ensured a balanced result for represented science areas and science communities. The proposed selections have in both cases been approved by the Executive Committee and the General Assembly.

Both Call executions and selection processes have been finished in time to enable starts of Science Demonstrators on July 1 and December 1. The second Open Call was executed a little earlier than originally planned to allow for project on Dec 1, 2017, rather than on Jan 1, 2018, with an end already in Nov 2018 rather than at the EOSCpilot end on Dec 31, 2018.

4. SELECTED SCIENCE DEMONSTRATORS

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 & 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**): The **photon-neutron** community to 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

Selected Science Demonstrators receive central funding of 12 PMs for their 12 months long engagement with EOSCpilot and their various tasks and missions to help improving the service portfolio and processes essential for the EOSC thus contributing to the advancement of Open Science in Europe.

Details of the engagement model of Science Demonstrators have been developed in the Deliverable D4.1 due April 1 2017.

The engagement model described in D4.1 has been applied to the accepted Science Demonstrators. Kickoff meetings have been organized to familiarize contacts of Science Demonstrators and their respecting shepherds with their tasks and duties and expected contributions. The kickoff meetings for Science Demonstrators took place in January (for the first five pre-selected Science Demonstrators), in July (for the second five Science Demonstrators selected from the first Open Call) and in December (for the third five Science Demonstrators selected from the second Open Call).

The activities of Science Demonstrators have been closely accompanied by shepherding (see next chapter).

The activities of the Science Demonstrators are progressing well. One Science Demonstrator from the pre-selected ones (ERFI) was starting with a delay due to complications in its organization and started only in July together with the second five Science Demonstrators.

5. SHEPHERDING AND MONITORING ACTIVITIES OF SCIENCE DEMONSTRATORS

According to the engagement model proposed in deliverable D4.1, each Science Demonstrator will:

- 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.

All the activities required from a Science Demonstrator to engage with EOSCpilot are co-funded through 12 PMs for 1 year from WP4 Task T4.2 (with the exception of DPHEP/WLCG which made contributions in kind).

The support provided by shepherds, according to Task T4.3, is essential. 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 supports the main shepherd in its 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 interact with their Science Demonstrators independently for carrying out the work plans and help with the identification of issues and solutions.
Shepherds guide and moderate the interactions between Science Demonstrators and WP5 and WP6 and take part in respective meetings organized by WP5 and WP6.

Activities have been monitored through Task T4.1 in monthly video conferences. Feedback to other WPs when appropriate is provided either through the respective shepherds, or through reports. Experiences made by each Science Demonstrator are collected in reports at month 4, 8 and 12. In the context of the engagement process with other WPs, WP5 and WP6 have been invited to devise questionnaires for the Science Demonstrators to better understand their needs in terms of Infrastructure and Interoperability challenges, with answers being collected together with the reports.

6. CONCLUSIONS

WP4 Science Demonstrators in its first year has progressed along the lines described in the WP4 work plan according to the DoA with its three subtasks.

The two Open Calls for Science Demonstrators that have been carried out including the setup, execution and general management of the review and selection processes, resulted in the acceptance of ten new Science Demonstrators from different science areas so that all major science fields targeted by EOSCpilot have been covered.

The engagement model devised for and applied to the Science Demonstrators has ensured the necessary information flux and communication between Science Demonstrators and the project as a prerequisite to obtain the necessary feedback and recommendations with respect to improvements on various aspects of EOSC.

The concept of shepherding - what means the assignment of a main and a co-shepherd with domain and respective IT expertise as main contact for Science Demonstrators with regular meetings – has so far proven to be very useful for guidance and project execution.

First final reports of the first set of Science Demonstrators will be due and available only by the end of Jan 2018. Their evaluation and conclusions as well as those from the other Science Demonstrators will be included in deliverable D4.4 as the essential outcome of WP4.

Preliminary conclusions can, however be drawn by inspecting the 8 months reports of the first Science Demonstrators, by including the general Science Demonstrator feedback and recommendations given during the EOSCpilot organized First Stakeholder Event on Nov 28, 2017 in Brussels.

Accordingly TEXTCROWD is proceeding as expected with no technical or organizational/procedural issues identified.
DPHEP/WLCG is proceeding as expected with minor technical and organizational/procedural issues identified and being addressed.
Photon&Neutron is proceeding as expected with minor technical issues identified. A prolongation beyond the 12 months duration (on own resources) has been suggested to be able to deliver more precise recommendations.
Pan-Cancer is proceeding with minor technical and organizational/procedural issues and suggested a project prolongation by 6 months (on own resources) to be able to provide the envisaged results.
ERFI has started with 6 months delay due to time consuming internal coordination processes in a largely distributed community and will run until mid of 2018, as will the second set of Science Demonstrators.

Preliminary general recommendations were collected from Science Demonstrators during the First Stakeholder event on Nov 28, 2017 in Brussels.

Accordingly, the following areas need to be addressed:

- The Identity provisioning in a federated EOSC model,
- the need for motivation stimulators for scientific communities to adopt interoperable standards,
- the need for support of high bandwidth data connections between data centers,
    - the need for the integration of HTC compute resources with large data resources,

- a way for the usage of EOSC with different underlying data models and strategies to achieve interoperability without collapsing data models,
- the need for cross-discipline data interoperability,
- the tradeoff between performance and resources in a cloud computing model,
- the need for stimulators to make better progress towards the implementation of F.A.I.R. principles,
- the discrepancy between long-term, multi-decade commitments in ESFRI and other projects and short life-times of EC funded projects.

According to a consensus opinion of Science Demonstrators the benefits through Open Science and Data as a significant multiplier for knowledge creation through additional other scientists further working on these data by far outweigh potential misinterpretations or misuse of open data.

Initial concrete recommendations from already include:

- Progress towards better interoperability shall not be made by the standardization of the well established community specific tools, procedures and formats, but by work on interfaces to achieve compatibility through conversions.
- The way how and by whom hardware resources for general usage shall be provided has to be clarified, since the established community specific infrastructures cannot be expected to be opened for general free use on behalf of their respective budgets.
- Issues with copyright laws should be addressed since they block access to data for open use.
- Strategies and concepts for long-term (multi decade) data preservation need to be developed.

.

ANNEX A. GLOSSARY

The definitions below shall be considered for the purpose of this deliverable.

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	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.
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.

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 will drive the further development of EOSC.
Science Demonstrator Representative	Contact person from a specific Science Demonstrator. He/She 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 an approved 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 going to be used.

ANNEX B. 8 MONTH REPORTS OF FIRST SET OF SCIENCE DEMONSTRATORS

This annex contains the 8 months reports of the first Science Demonstrators that started on Jan 1, 2017: TEXTCROWD, PanCancer, PhotonNeutron, DPHEP/WLCG (ERFI started delayed)

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-09-08 / Second Report
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	
Current status	Software development for Italian archaeological NLP tools continues without particular problems. Current effort is devoted to the refining and improvement of the NER system and the strategy definition for CIDOC CRM information encoding and for deployment of the prototype within the D4Science cloud. In particular, the current phase of software development aims to produce a working prototype of the pilot NER system for the first phase of evaluation. Later efforts will be focused on evaluating and improving each component within the NER system's pipeline. Strategies for encoding and exporting identified relevant entities in CIDOC CRM (RDF) format are also under refinement for final implementation within the next period, when a working demo will be made available for the cloud. A beta version of the demonstrator is expected to be available in late October/early November.

Progress made	A technical meeting was held on 10th and 11th July 2017 in Prato to refine terminological tools, identify relevant entities to be extracted from documents and integrate additional linguistic frameworks to be used for POS and NER operations on Italian archaeological reports that will form representative input data for the EOSC Demonstrator. The meeting was also useful to improve the archaeological vocabularies, derived from Italian ICCD standard thesauri, aimed at improving performances of NLP process on Italian documents. Expert annotation by Italian archaeologists of a sample of archaeological reports, to assist both formative and summative evaluation of NLP software tools produced for the demonstrator, continues. Requirements have been refined and tailored to meet the specific requirements of the archaeological field. An initial first evaluation of the pilot NLP system, outlined in the first periodic report, was conducted to discuss the results and look at example annotations and issues. The candidate set of Italian archaeological entities for the NER demonstrator work was reassessed as part of the evaluation exercise and slightly modified. Building on the evaluation of the pilot NLP system, the GATE pipeline has been refined and further developed. The main points were as follows: - importing updated vocabularies and some pre-processing of their content - replacing the Italian OpenNLP with FP7 project OpenNER components via web service calls from GATE, with resulting improvement in NER discovery - checking OpenNER outcomes - refining stemming/lemmatization component - developing POS rules for filtering on nouns when annotating - specialised timespan and period component with pattern based rules. A virtual meeting was held with the D4Science team of CNR, in charge of providing VRE cloud facilities for the final deployment of TEXTCROWD, to develop a strategy for the migration of the tool in the cloud planned for next period. Discussion on technical aspects of this activity, aimed at clarifying and solving potential issues for final deployment, is ongoing within the technical teams. Implementation of the necessary components into the VRE has already completed or is in progress.
Problems encountered	There are currently no available corpora of manually annotated Italian archaeology reports: the Demonstrator project will create one. Potential semantic overlap between particular Italian archaeological concepts (candidate entities) has been discussed. A strategy has been developed to study their use in practice within typical Italian archaeological reports and a prioritization mechanism has been identified.

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.
Outreach activities	Some meetings have been organized with other research groups potentially interested in TEXTCROWD. They concern 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 proposed "Visual Media" pilot. 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 Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	TEXTCROWD is currently oriented to the community of archaeologists and to the management of archaeological textual documents: the tool is currently being evaluated by a team of Italian archaeologists and tested on archaeological information coming from MiBACT and FASTI Online, which is also providing the textual documents used during the current development phase. Feedback from these and other forthcoming archaeological communities will be continuously provided throughout and after the development process.
Adequacy of technical solutions	Evaluation of the NLP tools to be conducted at a later stage in the process.
Missing functionality or services	No
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)

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-04-28/ Second Report
Science Demonstrator Title	Pan-Cancer Analysis in the EOSC (Acronym: PanCancer)
Representative name, affiliation and email from proposing organisation(s)	Sergei Iakhnin European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. iakhnin@embl.de
Main Shepherd name, affiliation and email	Dario Vianello EMBL-EBI, Hinxton, UK dario@ebi.ac.uk
Secondary Shepherd name, affiliation and email	Gergely Sipos EGI.eu, Amsterdam, Netherlands gergely.sipos@egi.eu
General part	
Current status	The Butler scientific workflow system is deployed on the EMBL-EBI's Embassy Cloud (OpenStack-based) using 1000 virtual cores, 4TB of RAM, and 1 PB of Isilon storage. Additional resources have been made available by ComputeCanada and Cyfronet for project use. Work is underway to install Butler at these sites and engage testing.
Progress made	- Improved Butler deployment procedure to require fewer steps - Authored detailed Reference and Installation guides to aid users. - Tested deployments onto Amazon AWS and Microsoft Azure commercial clouds. - Developed project proposal for ComputeCanada resources. - Developed project proposal for Amazon AWS resources. - Implemented operational metric alarms and notifications capabilities in Butler. -

Problems encountered	- Deployments of OneData at EMBL/EBI have encountered repeated failures. - Additional resources (non-EBI) have only been made available to the project in the second week of September, 2017.
Data management and handling of sensitive data, with reference to plan.	- Existing processes (set up for PCAWG) ensure secure data handling on Embassy Cloud. - Similar processes will need to be adapted to other infrastructure providers.
Outreach activities	N/A
Specific Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	No specific feedback
Adequacy of technical solutions	- EMBL-EBI Embassy Cloud coupled with the Butler framework is an adequate solution for processing thousands of cancer genomes. -
Missing functionality or services	- OneData installation for securely moving genomic data between repositories remains a challenge.
Other (please specify)	

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-09-07 / Second Report
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	
Current status	Photon and Neutron facilities have a long history of collaborative efforts in various fields. The PaNdata consortium of European P&N user facilities is continuously developing tools and services for data management and data analysis and standardization of formats and policies. Providing "Data Analysis as a Service" is one particularly important activity aiming to provide portable and scalable solutions with a high degree of user-friendliness. To demonstrate the capabilities of cloud based services and EOSC in particular the P&N demonstrator is focussing on a particular application framework used for serial (femto-second) x-ray crystallography. The framework (crystfel) is particularly suited since it is well documented with vast amount of data retrievable from a mature open access repository (cxidb).

Progress made	**Data access & sharing** Currently, the data relevant for the demonstrator are available from the CXIDB repository and can be harvested with simple curl commands. That works sufficiently well, but is not very convenient for the computational tasks. We aim for two different approaches: one is to establish a OneData provider to access data in the usual posix-style. This is still work in progress. The other alternative is a HDF5server-based solution. The application framework utilizes different subsets of data in different stages of the analysis and both solutions allow data processing with little overhead. We expect the services to be available by end 2017. A closely related project named EUCALL and therein the SIMEX work package has established a simulation pipeline. The pipeline has been dockerized and can conveniently be deployed on HPC clouds. The simulation pipeline allows simulating entire experiments like serial femto-second x-ray crystallography from accelerator (synchrotron or FEL) up to the detector. The simulation data will serve as input to CrystFEL (and other applications) thereby allowing to validate both implementations of the analysis framework as well as the feasibility (signal-to-noise) of the experiment. For that purpose EUCALL has established a zenodo community for upload of simulation data. A small convenience python-wrapper is currently under development which should allow easy and automatic upload and harvesting of such data into the analysis framework. **Analysis framework** The application has meanwhile been improved (not part of the demonstrator), which makes it more "cloud-compliant". In addition GPU-support has been added. The containers (docker/singularity) have partially been updated accordingly. Test of the GPU capabilities in a cloud environment still need to be tested. **Cloud infrastructure**

	The local openstack infrastructure is continuously being improved and web-services for simple container deployment in HTCondor/Slurm hybrid infrastructure being developed. That's not strictly part of the demonstrator, but the demonstrator will certainly be used to test these services later on. We started to test the EGI AppDB VMOps dashboard. That looks extremely convenient and promising. We plan to run some real tests later on.
Problems encountered	The setup of an OneData provider takes a bit more time than expected (mostly due to internal issues) in particular in combination with dCache. The HDF5server based services are also lacking behind. We did not make any attempt to make use of any AAI services yet.
Data management and handling of sensitive data, with reference to plan.	All data relevant for the demonstrator are open access. Handling of sensitive data is hence not 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. Deployment at DESY is still on hold. Adding demonstrator data to an ICAT instance makes however limited sense, since all FEL data are actually being created outside the European P&N landscape (LCLS, SACLA).
Outreach activities	Presentation at the RDA/PaNSIG workshop in Barcelona. Presentation at DESY. We see 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, making a cloud deployment of such standard applications in a cloud environment considerably easier.
Specific Feedback on:	

Communication/ information flow between the science community behind a science demonstrator on the processes	Through various projects and P&N related activities. The science communities are currently not being involved or informed much except for the local user community. For the general P&N user community that would require substantial services in production and is way beyond of the demonstrator (or the EOSC pilot).
Adequacy of technical solutions	For this demonstrator it looks good.
Missing functionality or services	Mostly community specific services (ICAT, HDF5server).
Other (please specify)	

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-08-31 / Second Report
Science Demonstrator Title	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, MPG jkennedy@rzg.mpg.de
Secondary Shepherd name, affiliation and email	Matthew Viljoen, EGI matthew.viljoen@egi.eu
General part	
Current status	DPHEP started as a Study Group, led by DESY, to look into the problem of Long-Term Data Preservation for Future Analysis. Following the signature of a Collaboration Agreement by the main HEP institutes and funding agencies worldwide, DPHEP became a collaboration with a "2020 vision". The services offered by CERN for LTDP were described in an iPRES 2016 paper. These services include an "Open Data Portal" through which data, documentation and complete virtual machines can be accessed. The EOSC Pilot SD "DPHEP" focuses on the latter: trying to build a "mock-up" of the Open Data Portal using services offered by the EOSC Pilot and / or its components (e.g. EGI, EUDAT, etc.) These services include: 1. A Trustworthy Digital Repository (TDR) for the (binary) data itself; 2. Digital Library Services, e.g. based on Invenio, such as B2SHARE, for the documentation; 3. A service for capturing and preserving software and the needed environment, e.g CernVM / CVMFS. The goals of the DPHEP Collaboration are outside the scope of this SD: it is purely an attempt to build a "pop-up" equivalent of the Open Data Portal (itself based on Invenio etc.)

Progress made	Since the last report, sites offering the requested services for data have been identified (CINES + CINECA). To date, no data has been uploaded, pending a number of clarifications from both sides, including the necessary recipes, as well as clarification of the scope and duration of the "pop-up". (See general part).
Problems encountered	It is unlikely that the initial target of 100TB of data in a TDR will be reached. This goal, intended to show that the EOSC Pilot services could reach the scale of the existing CERN Open Data portal, will therefore be softened to a "proof of concept" – i.e. just a few files of each category.
Data management and handling of sensitive data, with reference to plan.	
Outreach activities	Presentation at the e-IRG workshop in Malta in June. Scheduled presentation at PASIG Oxford (13 Sep).
Specific Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	The DPHEP community has been kept informed through existing DPHEP channels.
Adequacy of technical solutions	Experience has shown that not all TDRs are equal: the specific needs of a given community (in terms of data types, volume, conversion issues etc.) would need to be taken into account for any LONG TERM production solution (which is outside the scope of this SD). To put it another way, the success of this SD as a "pop-up" would not necessarily mean that long-term (i.e. 25 – 30+ year) solutions could automatically be built using these technologies / services / sites. Whilst outside the scope of this SD, these are concerns that would nevertheless need to be addressed as the EOSC moves forward towards a medium or long term solution offering production services.
Missing functionality or services	

Other (please specify)	

ANNEX C. 4 MONTHS REPORTS OF SECOND SET OF SCIENCE DEMONSTRATORS

This annex contains the 4 months reports of the second five Science Demonstrators (plus ERFI) that started on July 1, 2017: EPOS/VERCE, LOFAR, CryoEM, PROMINENCE, ERFI, Life Science Data Sets

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-11-06 / First Report
Science Demonstrator Title	Virtual Earthquake and Computational Earth Science e-science environment in Europe (EPOS/VERCE)
Representative name, affiliation and email from proposing organisation(s)	Andreas Rietbrock, ULIV, andreasrietbrock@gmail.com Alessandro Spinuso, KNMI, alessandro.spinuso@knmi.nl Andre Gemuend, Fraunhofer SCAI, andre.gemuend@scai.fraunhofer.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	

Current status	This first reporting period was focused on setting up the Science Demonstrator and fulfilling the technical prerequisites for the scientific work to be carried out. This mainly involved the preparation of support of FedCloud Cloud infrastructure as backend resources for Workflows of the Portal and extensions of the Science Gateway frontend and backend services to allow the implementation. After the upgrade the current provenance management system, domain metadata and a coverage of the lineage will be discussed to improve usability. We will upgrade the framework used for defining the stream-based processing workflows used inside the jobs (dispel4py) and migrate the workflows where needed. The GUI has already been improved for the Use Case, further improvements are in progress. The scientific part for the misfit calculation is now operational and we are working on the upscaling and defining the best scalable use case. So far, after a simulation for a given earthquake, the corresponding raw recorded data can be easily downloaded from European archives using the Download workflow. Then, both data and synthetic traces can be prepared for comparison by applying fully customized seismological processings through the Processing workflow. Finally, the Misfit workflow offers different possible procedures to compare recorded and simulated seismograms that give quantitative estimates of misfit criteria. Further processing functions and misfit procedures can be easily implemented into the platform based on the chosen use case.

Progress made	The Demonstrator can resort to an existing EGI VO called verce.eu. We have tested the readiness of the cloud providers of the EGI Federation. So far, three cloud providers (IN2P3-IRES, SCAI and HG-09-Okeanos-Cloud) have been configured to support the verce.eu VO and support the research activities of this scientific demonstrator. There is no need for additional resources at the moment. To run the scientific workflow in the EGI Federated Cloud infrastructure, the DCI_BRIDGE VM image [1] has been updated in order to get rid of the LVM support (problematic because unsupported on some FedCloud sites), and add the additional libraries needed by the scientific community. A new VM image is now available for testing. In the last months we have also worked to extend the AAI mechanism used by the WS-PGRADE/gUSE portal in order to: 1.) enable federated access with the OIDC [2] module developed for Liferay by the INFN Catania, and 2.) add support to Per-User Sub-Proxy (PUSP) [3]. The portal is now configured to make REST calls, and download, from the eToken server proxies certificates generated from a robot registered in the verce.eu VO. A portlet has been created to download the proxy and include it in the gUSE certificate store for use in workflows. The integration of the OIDC module and enable federated authentication mechanism is still working in progress. Download, preprocessing and Misfit workflows have been fixed for bugs and validated by domain scientists for their correct behaviour [5]. The frontend has been improved in multiple aspects for misfit calculation, e.g. on the Download Tab (for setting up the acquisition of measurement data from data centers), new query string parameters are automatically defined to limit the search and download of raw measurement data to the corresponding networks and stations of the simulation run. Also, following in the Processing Data Setup Tab, only raw-data download runs which correspond to the selected simulation will be shown. The S-ProvFlow system has been improved in many aspects: Better Rest API methods, Provenance Repository performances,

	Frontend usability and modular “Dockerisation” of each component. Current development branch at [4]: [1] https://appdb.egi.eu/store/vappliance/fedcloud.slave.dci.bridge c.f. gUSE Cloud documentation https://sourceforge.net/projects/guse/files/3.6.8/Documentation/Cloud_Specific_Administration_Guide.pdf/download [2] https://github.com/csgf/OpenIdConnectLiferay/tree/EGICheckIn [3] https://wiki.egi.eu/wiki/Usage_of_the_per_user_sub_proxy_in_EGI [4] https://github.com/aspinuso/s-provenance [5] https://github.com/KNMI/VERCE/tree/master/verce-hpc-pe/src/test
Problems encountered	As described in the progress section, there have been some technical problems met during the period which have been mostly solved. They included (c.f above): - VM image required by gUSE Cloud integration needed to be updated (dci-bridge included in image, Obspy & GridFTP etc. included, LVM issue) - FedCloud integration requires the support of the site for public floating IPs, and a Network policy which allows port access on the port where the DCI-Bridge listens on in the available VM image - Regarding support of Per-User Sub-Proxies, not all software supports PUSP out of the box. E.g. iRODS GridFTP DSI does not support PUSP (all PUSPs will be recognized as the same user, i.e. using the robot certificate).

Data management and handling of sensitive data, with reference to plan.	Input data is mostly publicly available open data, sometimes asking for attribution / acknowledgement. Example of policies from input data used: https://www.orfeus-eu.org/data/eida/acknowledgements/ https://ds.iris.edu/ds/nodes/dmc/terms-of-service/acceptable-use-policy/ Own data: FAIR principles are applied for the management of the simulation and misfit results. Intermediate data-stores for results management and validation will be accessible adopting web standards and ensuring access control mechanisms (e.g. Http interfaces, WebDav with session authentication. Possibly supporting OAuth) enabling browser-based consumption. Provenance granularity and its contextual metadata (user or domain specific) should be customizable. Eventually, it should be managed in stages support different phases of the research-data lifecycle: live (within the science gateway), published and curated (relying on external institutional services and e-infrastructures).
Outreach activities	Presentation and live hands-on training at the Open EPOS Seismology & ORFEUS Annual Workshop in Lisbon, 25-27 October 2017. http://orfeus-epos.ipma.pt/ https://www.orfeus-eu.org/other/workshops/orfeus-epos-seismology/
Specific Feedback on:	

Communication/ information flow between the science community behind a science demonstrator on the processes	The seismological research community results to be particularly interested in a tool that eases the approach of both expert and non-expert users to the fundamental tasks of forward and inverse waveform modelling. The hands-on training demonstrated that the community finds very useful a graphical interface to quickly set up the jobs, explore their outputs and reuse their inputs for multiple experiments. The workflows so far available allow users to perform the main steps towards waveform misfit calculation exploiting worldwide established computing codes. Implementation of further workflows are in the plans as those for seismic source inversion following the interests expressed by the field community. Live demonstrations as well as online distributed tutorials and how-to manuals are fundamental to get the researchers involved and to promote the numerous functionalities of the platform.
Adequacy of technical solutions	Unclear which technical solutions are meant here.
Missing functionality or services	iRODS GridFTP DSI PUSP support. Liferay EGI Checkin Integration (Registration / confirmation of account, RCAuth.eu online CA for X509 certificate). Interesting could be an offered job queuing framework with hosted servers (e.g. comparable to Amazon SQS) and framework methods to automate pulling of jobs, scaling when jobs are left, etc. Additionally a SIMD like system to work on input sets, e.g. the same job on N input files (or alternatively a shared filesystem between VMs like Amazon EFS to implement one more easily than with object storage).
Other (please specify)	

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-11-08 / first Report
Science Demonstrator Title	LOFAR EOSC Pilot
Representative name, affiliation and email from proposing organisation(s)	ASTRON: Hanno Holties, holties@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 NLeSC: Niels Drost, N.Drost@esciencecenter.nl Ronald van Haren, r.vanharen@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 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

General part	
Current status	- Mostly on track. - Started with container deployment of processing workflows using CWL. - Investigating access to further appropriate Computational infrastructure. SURFsara willing to support with access existing systems and early access to their new HTC environment (first half 2018). Would be good to include infrastructure hosted by the other LOFAR partner institutes (FZJ & PSNC). - Investigating appropriate FAIR services/tools for LOFAR
Progress made	- Team formed - Workplan defined - Kickoff meeting: 30/10/2017 - Access to SURFsara HPC cloud for development and testing - First software ported to CWL, executable from a Singularity container on the HPC Cloud.
Problems encountered	- Administrative issues with subcontracting taking quite some time to resolve (get confirmation/consent from project organization). - Access to shared WP4 OwnCloud repository only working from 2017-11-07
Data management and handling of sensitive data, with reference to plan.	For now manually copying (small) datasets to targeted systems. No sensitive data issues. Connection to Archive storage in next phase.
Outreach activities	- Attended Pisa Meeting, presented LOFAR case, R. van der Meer - Poster at Adass XXVII conference, October 22nd – 26th 2017, Santiago de Chile, “What can Open Clouds do for Astronomy”, F. Pasian et al. - Topical Discussion: EOSC/RDA at the Aeneas All-hands Meeting, October 18th – 20th, 2017, Granada
Specific Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	- The community is currently represented by single use case. Will engage other groups and propose to form a community feedback group.

Adequacy of technical solutions	Currently used HPC Cloud is good for quick development and testing but not suitable for scaling up to realistically sized data processing. See also Current Status.
Missing functionality or services	Not yet encountered
Other (please specify)	-

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-11-05/ First Report
Science Demonstrator Title	CryoEM Analysis in the EOSC (Acronym: CryoEM)
Representative name, affiliation and email from proposing organisation(s)	Carlos Oscar Sorzano Sánchez Natl. Center of Biotechnology (CSIC) coss@cnb.csic.es
Main Shepherd name, affiliation and email	Gergely Sipos EGI.eu, Amsterdam, Netherlands gergely.sipos@egi.eu
Secondary Shepherd name, affiliation and email	
General part	
Current status	CNB team is working to make the pilot complaint with the FAIR principles and guaranteed the reproducibility of scientific analysis. To do so, CNB are further enhancing the Scipion application in order to link together raw data, metadata and tools used for the analysis and produce, as a final result, a scientific workflow. The workflow used to describe image processing steps is exported in JSON format and include also a list of movies and binary used during the analysis. The resulting workflow (in JSON format) together with the raw data can be used to reproduce the analysis or produce new data. This can also be done in cloud infrastructure (e.g. EGI). As second step, this JSON file is post-processed with a widget to produce a graphical representation of the scientific workflow.

Progress made	- We have defined the JSON file. - The JSON file can be exported from Scipion, with an option to export source data as a way to increase the reproducibility of the workflow. - It can also be imported from Scipion, along with the optional source data. - We have developed a workflow viewer in Javascript that can be easily integrated in any web page.
Problems encountered	- No specific problem
Data management and handling of sensitive data, with reference to plan.	Currently we have contacted the main raw database in the field to be able to submit the workflow as well as the raw data.
Outreach activities	N/A
Specific Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	No specific feedback
Adequacy of technical solutions	No specific feedback
Missing functionality or services	We need to verify that the solution provided covers most experimental situations.
Other (please specify)	

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-11-24 / first report
Science Demonstrator Title	PROMINENCE

Representative name, affiliation and email from proposing organisation(s)	Shaun de Witt CCFE (shaun.de-witt@ukaea.uk) Pär Strand Chalmers University, par.strand@chalmers.se David Coster MPIPP, dpc@ipp-garching.mpg.de
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	
Current status	Openstack deployed locally. Tested with various configurations. Some issues around mounting of lustre file systems on virtual machines (does not automount, but can be mounted manually). Internally tested some workflows and have prepared some docker images.
Progress made	2017-07: Contract setup. Negotiations with openstack commercial installers 2017-08: Agreement with Bright Openstack for installation. Attempted to get certificate (failed) 2017-09 – EOSC-Pilot Service Provider Workshop attendance. Installation and initial testing of local openstack instance. Attempted to get certificate (pending) 2017-10 Testing running od EFIT equilibrium fitting program on local cloud (success). Two additional test applications created as docker images (not yet tested)

Problems encountered	Primarily internal technical issues especially wrt external access which have yet to be fully resolved. Having problems obtaining certificate (and expect the same when joining VO) since fusion does not have a VO or a local signing authority. Currently trying to use STFC but the request got lost during the approval process.
Data management and handling of sensitive data, with reference to plan.	N/A.
Outreach activities	Presented EOSC-Pilot at IEEE-NSS in Atlanta.
Specific Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	Generally excellent with the shepherd and others being far more active and engaged than I was expecting.
Adequacy of technical solutions	Currently unclear – most of work expected in last 2/3 of project. Much of the initial 3 months was planned for internal work needed to set up for the demonstrator. The bulk of the work will take place Dec/Jan onwards
Missing functionality or services	The use of VOs for accessing the fedcloud will be a significant impediment to take up. Fusion used to have a VO but many external users considered the registration process too onerous and would not likely take up a service requiring this
Other (please specify)	

EOSCpilot: Science Demonstrator Report	
Date / Type	2017-11-06 / first report
Science Demonstrator Title	ERFI

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 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	N/A
General part	
Current status	Science case: discussion ongoing on datasets to be transferred matching scientific use case. The scientific use case is refined iteratively and addresses the forcing of land surface/ecosystem models (instead of the originally planned investigation of radiative forcing) Technical implementation: blocking issues with adoption of OneData services. New release promised but not yet available for exploitation. Climate models data archived on the ENES Data Infrastructure to be transferred to ICOS for forcing of land surface/ecosystem models (deviation from original plan of investigating radiative forcing).

Progress made	-number of files: few thousands. In time with the new EGI DataHub upgrade, explored two different options (with EGI and CYFRONET Team): 1.) Installing OneData provider plugin in the ERFI data repository (ENES data node installed at DKRZ) in order to configure it as data provider and identifying possible, related issues; 2.) Identifying other actors in the EGI Federation (DataHub) interested in hosting datasets and in being used as data providers. discussed: Exploitation of access to climate data/metadata via EUDAT endpoint and related services and tools. Yet the current lack of existing B2Stage (and B2Safe) endpoints with access to IS-ENES data makes this option unfeasible for this demonstrator. Thus the plan is to follow option2 in the future: transfer climate model data on a OneData EGI node (VM) via http or gridftp (exploiting the Synda tool). Contacts with DataHub are being established, yet progress is slow.
	Option 1. Is currently not an option because of blocking issues with the current OneData software (see ICOS report). Option 2 currently is on hold because of communication issues with the CYFRONET team and the rollout of the new OneData release. Due to these problems with OneData, the possibility of an alternative architecture (for option 1 above) proposed by EGI is

Problems encountered	General issues related to current OneData releases and with pace of OneData service offerings upgrades. ICOS: problem with the OneData service (latest version released in Oct 2017) containing a showstopping bug leading it to crash upon loading a few hundred files. IS-ENES: - no clear way to establish collaboration with OneData technical team and/or EGI/OneData service offering (see also technical options above).
Data management and handling of sensitive data, with reference to plan.	No sensitive data involved.
Outreach activities	IS-ENES: presentation of the prototype at ENES scoping meetings and discussions involving the Earth's Climate System Modelling community.
Specific Feedback on:	
Communication/ information flow between the science community behind a science demonstrator on the processes	No clear information flows between science community and EOSC service/infrastructure providers. The "shepard" can help to find / ping the right persons – but what if those are too busy to get back .. etc. On the other hand also science community discussion to distill real use cases with the potential to exploit new service offerings is quite slow.
Adequacy of technical solutions	Currently unclear – best would be to provide test-VM to the science demonstrator with needed infrastructure components installed (e.g. OneData in our case). After successful tests these infrastructure components can be installed at a RI center. Yet the first step is already a blocker.

Missing functionality or services	Unclear – as there is no clear way to evaluate existing functionality.
Other (please specify)	

EOSCpilot:	**Science Demonstrator Report**
Date / Type	Nov 7, 2017 / First Report
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 (CRG) – jordi.rambla@crg.eu Cedric Notredame. Centre for Genomic Regulation (CRG) - cedric.notredame@crg.eu
Main Shepherd name, affiliation and email	Erik van den Bergh, EBI evdbergh@ebi.ac.uk
Secondary Shepherd name, affiliation and email	Matthew Viljoen, EGI matthew.viljoen@egi.eu
General part	

Current status	- Software from workflow from GONL has been made ready, environment is set up - Mock example data is used, computational resources to be added - Docker container image was created with the tools used in the GONL workflow - Preparations are made to move containerized tools to Nextflow - Older tool versions are not available anymore e.g. GATK 1.0. Right now the closest versions are used (1.1), if output can be successfully reproduced it is considered OK - Basic pipeline containerised implementation based on Nextflow able to run in a portable manner
Progress made	Since beginning, see status
Problems encountered	- Ability of resource providers to provide security levels is unclear – the SD suggested a portfolio of security possibilities for each resource provider would be helpful - Size of computational resources needed is not yet clear -> exact requirements will be determined and fed back to shepherd - Old versions of tools not available -> Solved by using nearest available version and testing output is the same - Metadata repository for remastered data that will be used is not chosen -> Not planned to solve yet
Data management and handling of sensitive data, with reference to plan.	All data is handled internally in the EGA at the moment so not an issue currently.
Outreach activities	- Presentation on WP5 meeting in Pisa - Poster + talk at EOSCpilot stakeholder meeting
Specific Feedback on:	

Communication/ information flow between the science community behind a science demonstrator on the processes	Communication with GONL was good and helpful
Adequacy of technical solutions	No issues encountered with technical solutions so far
Missing functionality or services	Catalog of security requirements across resource providers
Other (please specify)	

ANNEX D. OVERVIEW DESCRIPTIONS OF 15 SCIENCE DEMONSTRATORS

Sequence of description: ERFI, DPHEP/WLCG, TEXTCROWD, Pan-Cancer, Photon&Neutron, PROMINENCE, EPOS/VERCE, Life Science Datasets, CroEM, LOFAR, Frictioless Data Exchange, Bioimaging, VisIVO, Hydrology, VisualMedia

Science Demonstrator Title	**ENVRI Radiative Forcing Integration**
Contact Name(s) Organisation(s)	Werner Kutsch, ICOS ERIC Alex Vermeulen, ICOS ERIC Ari, Asmi, ENVRIplus Paolo Laj, ACTRIS Stefan Kindermann, IS-ENES2 (DKRZ) Sylvie Joussaume, IS-ENES2 (IPSL)
Demonstrator Description	The ERFI Demonstrator shall further develop interoperability bet Observational and Climate Modeling Environmental Res Infrastructures. The scientific focus will be on dynamics of greenh gases, aerosols and clouds and their role in radiative forcing technical focus will be on: • improvement of data integration services based on metadata ontologies, • model-data integration by use of HPC, • innovative services to compile and compare model output fro different sources, especially on semi-automatic spatiotemporal scale conversion. The ERFI demonstrator shall also develop some best practice guid for global integrations e.g. in the framework of GEO and the Integ Global GHG Information System (IG^3IS) by WMO. In practice, the demonstrator will estimate the capacity requireme connect the (relatively) heterogeneous in-situ data systems from and ACTRIS to the IS-ENES2 climate data systems. The aim is to pr two-way automated interactivity, i.e. climate model data users ca relevant (climatological or specific-time) observations, and the in-situ users can access to relevant climate model data sets. This demons will also involve processing on spatiotemporal scale corrections.
Science Area	Environment, Greenhouse Gases, Climate Change

Science Demonstrator Title	**WLCG Open Science Demonstrator**
Contact Name(s) Organisation(s)	Bob.Jones@cern.ch; Helge.Meinhard@cern.ch; Jamie.Shiers@cern
Demonstrator Description	*Demonstrate scientific excellence and societal impact* CERN is a practitioner of Open Science continuing on from the role played in terms of Open Access to publications. The LHC experimer have data policies that call for their data to be preserved, for public releases of subsets of their data after a relatively short embargo per and for the ability to reproduce (at least key) analyses. (Data must b preserved before it can be re-used or shared). Experience from these Open Data releases has shown benefits not in closely related disciplines (e.g. theoretical physics) but also in computer science in general.
Science Area	High Energy Physics (Physical Sciences and Engineering according ESFRI classification)

Science Demonstrator Title	**Collaborative semantic enrichment of text-based datasets (TEXTCROWD)**
Contact Name(s) and Organisation(s)	Franco Niccolucci - PIN-University of Florence, Italy, coordinator of ARIADNE and PARTHENOS
Demonstrator Description	Process openly available texts availing of cloud-based tools for sen enrichment, linking and annotation. Set up a personal virtual enviror eventually shared with others, to address specific research questions. The concept is to produce and accumulate in a (semi-) automated collections of semantically enriched texts based on domain ontologies thesauri. Such texts have at present global descriptive metadata b semantic structure. Researchers will generate and revise the enrichm group of texts availing of the (semi-) automatic tools provided, because are directly interested in them; these texts will then be available for searching by others. Thus the total number of enriched and searchable will increase at every use, through a sort of crowdsourcing. Human supervision is required in the process e.g. for disambiguation ar referencing, but may be reduced using “intelligent” tools and well-des thesauri and gazetteers.
Science Area	Language studies, humanities, heritage

Science Demonstrator Title	**Pan-Cancer Analysis in the EOSC** (Acronym: **PanCancer**)
Contact Name(s) Organisation(s)	Jan Korbel, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany & Brazma, EMBL-European Bioinformatics Institute (EMBL-EBI), Hinxton, UK.
Demonstrator Description	Cancer is the leading cause of death in several European countries, results in €120 billion in annual health related-costs in Europe, and its relative impact on society is expected only to grow due to demographic changes. Progress in DNA sequencing technology has revolutionized this field[1-3], and hundreds of thousands of cancer genomes will become sequenced in Europe by the end of this decade resulting in exabytes of data. But there are clearly unprecedented computational demands to be met for utilisation of these data for the benefit of patients[4]. Pan-cancer analysis – collaborative data sharing and analysis across countries and cancer types to maximize the statistical power for health-related discoveries – is a powerful EOSC science demonstrator use-case, due to: (*i*) the relevance of this research for science and society, and (*ii*) close links and synergy with a relevant major international pilot study with European leadership and governance[1] – the Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative, as an international forerunner in meeting challenges related to the sharing, secure storage, and cloud-based processing of sensitive patient genomic data[4,5]. **Aims**: (1) PanCancer will develop interoperable IT frameworks to enable standardized sharing and large-scale processing of cancer genomes with other molecular and clinical data on the EOSC, to enable biological and translational breakthroughs. (2) We will employ these interoperable frameworks to process ~10,000 cancer whole genome sequenced (WGS) tumour-normal pairs from 20 most common cancer types. Such a panel of cancer WGS samples will lead to markedly increased biological discovery power over the state-of-the-art. A main research focus will be on uncovering genome-wide patterns of different types of genetic variation, which requires availability of WGS data, and integrating these with molecular, demographic and clinical data. (3) We will additionally integrate >20,000 cancer exomes, expression and epigenome data to facilitate discoveries of alterations affecting genes. In the future this frameworks will facilitate a shift from the focused use of specific research cohorts to incorporating orders of magnitude larger health data (*e.g.* Genomics England's 100,000 Genomes Project). While such secondary data will come with separate physical storage and ethics procedures compatible with the legal provisions of respective countries, PanCancer will allow to proactively create suitable standards and interoperability.
@Science Area	Life sciences (*i.e.* basic biology & translational biomedicine)

Science D.Title	**Research with Photons & Neutrons**
C. Name(s) and Organisation(s)	Volker Guelzow (DESY), Thomas Schneider (EMBL), Rudolf Dimper (ESRF), Sune Rastad Bahn (ESS), Andreas Schwarz (European XFEL), Jean-François Perrin (ILL)
Demonstrator Description	Photons and Neutrons are widely used for research in many scientific fields. Examples listed under the topic Science Area The production of highly intense photon & neutron beams requires large Research Infrastructures (RI) provided by the partners. Research at these RIs makes use of large area detectors, multi-channel detection, and high repetition of measurements. This leads to large quantities of data and raises the need to perform data analysis in an efficient manner. Thousands of users of the RIs propose, conduct and analyze data from scientific experiments in a wide range of application domains. Access to the RIs is gained through peer review of the scientific proposals. Often these groups are small teams of scientists coming from universities and research organizations. Many of these groups use RI's at various locations in Europe according to the specific characteristics of the instruments/beamlines and often more than one analytical facility for the same experiment. Key issues are data storage, sustained access to the data and an efficient data analysis ecosystem. Some of the facilities have adopted an open data policy. However, re-analysis of open data is currently not in the remit of the RIs. We can foresee two important impacts of a successful EOSC initiative for our user communities: -Enable processing of data coming from different RI's and/or remote processing of data by several groups residing at various universities/research institutes." -Allow to re-process (or process differently) large volumes of (or complex) Open Data produced at the analytical facilities. This demonstrator has the following goals: -Help scientists through the usage of cloud technology to gain access to federated resources for data analysis. -Demonstrate -in cooperation with other work packages- the feasibility of providing services in the context of large analytical facilities. -Collect requirements of the scientific communities for a future cloud solution on technical and socio-economical issues.
@Science Area	Physical sciences, life science, chemistry, drug design (pharmaceutics), material scienc environmental sciences, cultural heritage,

Science Demonstrator title: **PROMINENCE: HPCaaS for Fusion**

Principal Investigator and team members :

Dr Rob Akers

Head of Advanced Computing, Culham Science Centre , Abingdon , Oxfordshire , OX14 3DB, UK

Shaun de Witt

Culham Science Centre , Abingdon , Oxfordshire , OX14 3DB , UK

Dr David Coster

Max-Planck-Institut für Plasmaphysik, Boltzmannstr. 2, 85748 Garching, Germany

Dr Par Strand,

Director of Chalmers e-Science Centre, Chalmers University of Technology, SE-412, 96

Gothenburg, Sweden

John Corne

Bright Computing BV, Kingsfordweg 151, 1043 GR Amsterdam, The Netherlands

Brief Summary of the Science Demonstrator:

Access to HPC facilities are vitally important to the fusion community, not only for plasma modelling but also for advanced engineering and design, materials research, uncertainty quantification and advanced data analytics for engineering operations (e.g. condition monitoring). The requirements for HPC class machines is expected to only increase as the community prepares for the next generation facility, ITER. However, access to HPC class infrastructure is quite restricted and obtaining time on this class of machine for algorithm development, testing and optimisation is already problematic. A few fusion centres have local access to smaller HPC class clusters but larger scale workflows and smaller fusion research centres are required to competitively bid for time on hardware such as PRACE Tier-0 and Tier-1 facilities, often also requiring visiting the centre.

Within this demonstrator we propose to make HPC class machines available as a cloud like service to the fusion community, in a similar way to those available through AWS (https://aws.amazon.com/hpc) and Microsoft Azure

(https://azure.microsoft.com/en-gb/solutions/big-compute). This is currently not a feature offered by any of the current e-infrastructures (EUDAT, EGI, Indigo-DataCloud, PRACE, etc), but has been investigated in other H2020 and FW7 funded projects (cloudSME, HOST). Building and collaborating with these projects and the current infrastructure projects we aim to demonstrate

that by using an industry standard Openstack instance at CCFE we can make HPC class nodes together with fast interconnects accessible to other members of the fusion community to exploit. The pilot infrastructure will be made available to the Eurofusion MST community for joint research (covering four large scale facilities: MAST-U at Culham, WEST at CEA, ASDEX Upgrade at IPP Garching and TCV and EPFL/CRPP Lausanne).

Description of the Scientific Demonstrator:

The MST experiments listed above all allow analysis of their data by the entire Eurofusion community; the infrastructure to achieve this has been in place for well over a decade. However, the system is based on restricted ssh access to sites with data for each experiment held in localised silos, thereby lacking scalability and flexibility required in the run up to ITER for running different codes with very different requirements. Analysis must be run local to the data which often means it is restricted to certain periods and is limited by local resources. By providing HPCaaS with a demonstration of bursting out from a local resource onto a larger pool of resources we will demonstrate that these limitations can be overcome at least as far as computational resources are concerned. In addition, the use of containerisation for HPC class codes means that the end user no longer needs to work with local site admins to ensure the required software and hardware resources are available, but can build up and maintain their own images which can be run on any cloud instance with the possibility of bursting onto a larger virtual infrastructure. Issues associated with the remote access between data and compute are being addressed by the Fusion EOSC-Hub Competency Centre (involving STFC, UKAEA, CEA and FZJ); for this specific work we will ensure that data is copied within the container to demonstrate the principle of running HPCaaS.

Science Area :

Primarily Energy (2.7) and Plasma Physics (1.3), but including Materials Engineering (2.5).

Science Demonstrator title:

EPOS/VERCE: Virtual Earthquake and Computational Earth Science e-science environment in Europe

Principal Investigator and team members: Andreas Rietbrock - University of Liverpool

Emanuele Casarotti - INGV

Horst Schwichtenberg - Fraunhofer SCAI Alessandro Spinuso - KNMI

Brief Summary of the Science Demonstrator:

Several seismological events, not last the 2016 destructive Amatrice earthquake sequence (Italy), that are still causing victims and damages to historical sites, have shown again the urgent need to understand the complexity of the underlying processes of an earthquake rupture/sequence. In the last decade, the availability of open access and high quality seismic observations has increased exponentially (100s of TBs). Open Source and community supported 3D seismic simulation tools in complex 3D media have become available (also supported by advances in HPC) and are now used by an increasing number of researchers. However, combining large amount of high quality data with complex 3D seismic simulations is still extremely challenging and has been accomplished only for a few regions on Earth. Furthermore, civil protection agencies have an increasing need of computing realistic scenarios of earthquake shaking to aid emergency planning and coordination of rescue efforts.

The VERCE project has pioneered a VRE to support researchers using established simulation codes on

high-performance computers in conjunction with multiple sources of observational data. This is accessed and organised via a science gateway that makes it convenient for seismologists to use these resources from any location via the Internet. Their data handling is made flexible and scalable by community-developed libraries, such as ObsPy (http://obspy.org), and data-intensive tools, such as dispel4py (https://github.com/dispel4py/dispel4py). It connects to federated data services of the FDSN (http://fdsn.org) to discover and ingest observational raw and parametric data delivered by worldwide seismological networks. Provenance driven tools (S-ProvFlow, https://github.com/KNMI/s-provenance) enable the rapid exploration of the results and of the relationships between data processes and users, which accelerates understanding and method improvement.

Description of the Scientific Demonstrator :

The EPOS/VERCE VRE currently supports two important use cases in the field of computational seismology research: (i) Earthquake Simulation: Synthetic Seismograms for public Earth models and earthquakes' source solutions via the execution of HPC simulation codes (SPECFEM3D, SPECFEM3D_GLOBE); (ii) Raw data acquisition & Misfit: The simulated data may be compared to real observations stored in distributed third-parties archives (FDSN) to evaluate and refine the Earth model. Building on these achievements published in a peer reviewed paper presented at the 2015 IEEE 11th International Conference on e-Science we propose for the EOSC demonstrator to expand the VRE to be able to compute realistic scenarios of earthquake shaking and visualise and compare the results to recorded strong motion records to address the needs of civil protection agencies for more accurate earthquake scenario estimates.

The exploitation of computational and storage resources will use standard middleware (globus, OCCI Cloud interfaces), besides authentication/authorisation technologies (X.509 VOMS) that are widely supported by the existing e-infrastructures. An intermediate data management layer, based on iRODS, stores experimental results obtained by simulations and misfit calculation. Users setup and control all the phases independently by interactively configuring dedicated workflows that are transparently translated into data-intensive methods and mapped to scalable resources.

HPC centres and institutional clusters across Europe have been used in the past with successful results, coupled with data and metadata management capabilities. Currently the portfolio of resources is expanding with the EGI FedCloud. Open-data and reproducibility is guaranteed by the automated adoption of community vocabularies enriched by Provenance information, in compliance with the W3C PROV data-model. Trainings, on which feedback was collected, have been carried out and the platform is actively used, as can be seen by e.g. a research story published for EGI (https://www.egi.eu/use-cases/research-stories/amatrice-earthquake/).

Science Area :

Researchers from computational Earth Science (FOS 1.5: Earth and related Environmental sciences) with a clear aim to include civil protection agencies.

Science Demonstrator title:

Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies

reproducibility, reusability and interoperability

Principal Investigator and team members : Coordinators:

• Jordi Rambla. Centre for Genomic Regulation (EGA Project Manager. CRG - Centre for Genomic Regulation)

• Cedric Notredame. Centre for Genomic Regulation (Group Leader. Notredame's lab - Bioinformatics and Genomics Programme. CRG)

Team members:

• Josep Lluís Gelpí. Barcelona Supercomputing Center (ELIXIR ES Technical Coordinator. BSC - Barcelona Supercomputing Center)

• Luiz Bonino (CTO FAIR Data. DTL- Dutch Techcentre for Life Sciences -)

• Angel Carreño (EGA Systems Manager - CRG)

• Paolo Di Tommaso (Nextflow author. CRG)

• Romina Royo (Software Engineer. BSC) Brief Summary of the Science Demonstrator:

Services like the European Genome-phenome Archive (EGA, https://ega-archive.org) or big consortia portals offer life science datasets for re-analysis at the requester computing facilities. Usually, these datasets have been processed with reference genomes and analysis pipelines that were current at the time, but become obsolete quite fast. Probably, every new requestor of such datasets would start their analysis by re-mapping with a current reference genome and running current gold standard tools or pipelines like the Pan-Cancer one. It is quite probable that, when possible because of huge resources involved, the requestor would like to reproduce the results from the original project.

Indeed, as most projects do not have a proper data management strategy, data are not made available according to the FAIR principles.

The EGA would like to avoid such wasteful and redundant processing by providing refreshed versions of the original datasets, while increasing the homology and interoperability between data for crossed analyses, in a similar way that the Pan-Cancer project applies a common pipeline to datasets initially processed by each individual cancer project. When offloaded, the process should happen following high privacy preserving procedures like the ones used by EGA.

This demonstrator will show how to minimize such wasting of time and resources by leveraging EOSC to offload the computing burden that such up-to-date process will put on EGA operational resources. It will foster reproducibility by packaging the original pipeline into standardized workflows, and using container technology to make such packages portable across different platforms. The new pipeline will be packaged in homologous ways, fostering reuse and reproducibility. As well, the process would be leveraged to apply FAIRification pipelines. The results will be send back to the EGA, thus every new user will benefit from those re-mastered datasets, while dataset discoverability will increase accordingly.

Description of the Scientific Demonstrator :

• Keeping datasets updated and comparable is a challenge that tampers discoverability, comparison between linked data and reproducibility, e.g. every request for a genomic variant must take into account which reference genome was used to re-map to the matching request.

• This pilot will require the management of sensitive data, to apply solutions for the reproducibility of analysis, and test a way to make data more interoperable using standard ontologies and FAIR principles.

• The components of this pilot have been developed and set up independently and tested in laboratory conditions, both at the CRG and the BSC, but never put to scale in terms of a production service.

• The need for making science reproducible, easily portable, and interoperable is common to all science fields that rely on analysing data obtained by third parties. A broader applicability end-to- end scenario would be hard to find.

• The pilot is supported by the EGA (internationally co-managed by the EBI and CRG in the context of ELIXIR Europe), the CRG Bioinformatics and Genomics program (a national Center of Excelence that participates in international projects like ENCODE) and the BSC, a national supercomputing facility that is part of the HPC European program. All of them data and service producers and consumers.

• The main goals of the pilot are to consume third party datasets stored at the EGA, consume reference datasets hosted at different services, produce updated datasets that will made available through the EGA, make those datasets interoperable in FAIR terms, and demonstrate how to make reproducibility portable via packing and sharing tools and workflows. Therefore, consuming and producing research artefacts is at the core of the pilot.

• Given that all parts of the pilot have been developed and made available independently, the pilot would just integrate them in a format that is suitable for EOSC, we consider that it is possible to have it completed in 12 M with the suggested 12PM effort.

Science Area :

1.2 Computer and information sciences

1.6 Biological sciences

Science Demonstrator title:

Acronym: **CryoEM workflows**

Full title: **Linking distributed data and data analysis resources as workflows in Structural Biology with cryo Electron Microscopy: Interoperability and reuse**

Principal Investigator and team members :

Jose-Maria Carazo (JMC). Director of the Instruct Image Processing Center (I2PC), Spanish National Center for Biotechnology, CSIC-CNB. JMC has long been working in the cryoEM area. As per Google Scholar on 23/04/2017, he has close to 10.000 citations and a h-index of 56. He directs the Instruct Image Processing Center (I2PC) at the Spanish National Center for Biotechnology CNB-CSIC. At the level of cryoEM key resources, JMC Coordinated the Bioimage EU project that was at the inception of current repositories for EM data (the “EMDB”, for Electron Microscopy Data Base, at the European Bioinformatics Institute). At the level of image processing, he has developed multiple methods as well as software, being responsible for the XMIPP and Scipion suites, that have been downloaded close to 1000 times by unique users just in the last year.

Pablo Conesa (PC). Instruct Image Processing Center (I2PC), Spanish National Center for Biotechnology, CSIC-CNB. PC has a biology degree and has long been working in the software development industry. He joined, for 5 years, the European Bioinformatics Institute (EBI-EMBL) working on repositories and data standardization, and recently, joined I2PC as the technical leader of Scipion, a framework for integrating the main EM software packages.

Laura del Caño is a Senior Software Engineer with a Ms in Physics. She has worked in the area of IT consultancy on applications design and development. She has also worked in different European research projects, such as the JCOP Framework, building Control Systems for the LHC detectors at CERN, or EGEE and GRIDCC, related to development of Grid infrastructures for science. She currently works on the Scipion project at CNB-CSIC, where she works as a software developer and is also in charge of deployment of Scipion software on the Cloud.

Brief Summary of the Science Demonstrator:

Structural Biology ("SB") aims at providing a detailed understanding of the 3D structure of macromolecular machines, many times reaching atomic resolution, as a fundamental step in the understanding of biological function. Among the structural biology (SB) techniques at the core the Research Infrastructure for SB, Instruct, microscopy under cryogenic conditions ("cryo-EM") is currently the fastest growing area, having been nominated "Method of the Year (2015)" by Nature.

Typically, cryoEM starts with the acquisition of thousands of "movies", extremely noisy and large, at specialized facilities. Multiple image processing operations are then performed on these images by many different data analysis tools until a quantitative 3D electrostatic map is obtained, conceptually forming an "image processing workflow". Individual processing steps may be performed at the experimental facility, at supercomputer centres or in the scientist's home instutition, so that typically the workflow is geographically distributed.

Public databases exist for the deposition of 3D maps (Electron Microscopy Data Bank (EMDB)), as well as for the deposition of key supporting evidence (EMPIAR). However, the way from the movies to the maps is full of case-dependent methodological image processing choices (the elements of the "image processing workflow"), linking different data sources at different sites and using a variaty of software with multiple versions. There is no standard method for recording these steps, and they are currently not deposited anywhere. Some details obviously can be found in publications, but they are far from being complete and, in any case, their description changes from author to author. Mining these heterogeneous data sets could bring new light to best practices and yet unknown data and analysis bottlenecks, but this currently impossible.

In this Science-Demonstrator we want to address the proper reporting of cryoEM image processing workflows, ensuring provenance at the level of data and analysis tools, linking workflow

information with raw data either at cryoEM facilities, individual laboratories or public repositories, so that reproducibility of scientific results were enhanced, data and analysis workflows could be reused and properly mined, allowing for a deeper level of interoperability among information sources.

Description of the Scientific Demonstrator :

The strong and well-defined scientific challenge of "cryoEM Workflows" is clear: In the most rapidly growing area of Structural Biology (cryoEM), a key piece of information linking distributed data

and data analysis tools is not being properly reported and organized (the image processing workflows). This situation has an immediate negative effect in scientific reproducibility, information sharing and data interoperability.

Structural Biology is an area that has been fast in realizing the need of information sharing. However, technology evolves quickly and current SB databases, specifically the main cryoEM information resource (the European-based EMDB at the EBI (European Bioinformatics Institute)), are lacking proper ways to capture key information. In this Demonstrator we will work in close coordination with INSTRUCT -the RI for Structural Biology-, EMBL-EBI and ELIXIR –the RI for life science information- and WestLife VRE, to develop: (1) Standardized ways to report cryoEM image processing workflows and, (2) Extensions of cryoEM image processing engines reporting workflows, which will be coordinated with EMBL-EBI to assure a very wide accessibility.

Currently, the proposing team is at a stage in which a precise action within the timeline of this

Demonstrator can be made. The following considerations sustain this claim:

(1) The Instruct Image Processing Center already develops Scipion (http://scipion.cnb.csic.es/m/home/), which is the workflow-oriented image processing framework that would be extended.

(2) The main proposer of this project (JMC at the I2PC) is also "Biotool editor for cryoEM image processing" in ELIXIR-Excelerate and partner in its WP5 (Interoperability). He is, therefore, very close to current initial specifications of the starting Common Workflow Language, which would be used as a starting point.

(3) The relationship between I2PC and EMBL-EBI is very well founded. As an example, EMBL-EBI currently uses two of the web analysis services provided by I2PC in the context of WestLife on their EM web-resources pages.

Regarding the capacity of this Demonstrator to grow from its initial focus on cryoEM, it is clear that the need to link data and data analysis in distributed environments is common to virtually all areas of research. Certainly common to most Structural Biology and Biophysical approaches within Instruct.

Science Area :

Structural Biology. This denomination refers to the study of the three-dimensional structure of biological macromolecules in the pursuit of better understanding biological processes and functions at the molecular level.

Frascaty field: 1.6 Biological sciences

Science Demonstrator title:

Open Science Cloud access to LOFAR data

Principal Investigator and team members :

- Hanno Holties, Rob van der Meer ASTRON (NL)
- Rob van Nieuwpoort, Netherlands eScience center (NL)

• Axel Berg, SURFsara (NL)

• René Vermeulen, ILT the International LOFAR Telescope consists of user consortia in 6 EU countries and two data centers besides SURFsara, in DE and PL.

• Michael Crusoe, Common Workflow Language, (LT),

• Fabio Pasian, INAF (IT)

• STFC (UK)

Brief Summary of the Science Demonstrator:

The goal of this Demonstrator is to allow for the science community to locate, access, and extract science from the LOFAR archive without being an expert on data retrieval and data analysis tools. At the moment the LOFAR data archive is operational and mostly used by experts.

The pilot will develop services, based on existing tools such as Xenon, CWL, Docker, Virtuoso, that allow users to initiate processing on data stored in a distributed, large-scale archive. This implies the ability to define and run custom processing tasks on multiple heterogeneous hardware platforms using data in multiple storage locations. The system will be based on workflows standards (CWL) and Containers (Docker, Singularity) to ensure reproducibility, and increase FAIRness of the system as a whole. The resulting system will be available for any scientist with access to the EOSC, and usage of SURFsara and the other ILT data centers will ensure ample storage and compute resources.

As a result, users can easily create new scientific results based on archived data products. It provides users with large-scale compute resources not likely to be available at their home institutions. It produces an overall multiplication of the total science output of the LOFAR archive. It also demonstrates how to arrange and organize comparable data archives and analysis infrastructures in the context of the EOSC, as well as enabling FAIR access for the first time in this domain.

Description of the Scientific Demonstrator :

The demonstrator will have a strong scientific impact thanks to the use of the advanced EOSC e-infrastructure. The invaluable existing LOFAR data will be made readily available to a much larger and broader audience, enabling novel scientific breakthroughs. Important discoveries are regularly made by re-analysing exiting astronomy data. For instance, an intermediate-mass black hole, an important missing link, was recently discovered in the Sagittarius constellation, by re-analysing 25 years of existing observations of pulsar PSR B1820 30A, which was already discovered in 1990.

Data integration and data interoperability allow users to exploit the sensitivity of multiple instruments, and are the driving force behind discoveries like this. The open science enabled by this proposal, in combination with the EOSC ecosystem will be a catalyst to make this happen with LOFAR data as well. This is of key importance, since LOFAR is two orders of magnitude more sensitive in its frequency range compared to previous instruments.

The LOFAR archive is already developed and demonstrated to be working at scale on existing infrastructures. The International LOFAR Telescope (ILT) consists of user consortia in 6 EU

countries and three data centers (besides SURFsara, there are centers in DE and PL). The partners in this proposal are committed to publishing papers, datasets, tools, and workflows following the Open Science model with the application of FAIR principles, and also as part of the Open Research Data Pilot in H2020.

Although this proposal deals with a concrete use case in radio astronomy, many of the existing tools we will port to the EOSC ecosystem (i.e., Xenon, CWL, Docker, Singularity) are broadly applicable. NLeSC already uses the tools in urban modelling, coupled climate simulations and virtual research environments in chemistry and archaeology.

Science Area :

Science area 1.3, Physical Sciences:

Research communities who benefit from to the science case are LOFAR users, Square Kilometre Array (SKA) user community, Radio astronomy, Astronomy, Astrophysics, and other data-intensive research domains.

Science Demonstrator title

Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories

Principal Investigator and team members

The project will be carried out jointly by teams at two institutions:

Knowledge Media institute, The Open University (OU), UK - http://kmi.open.ac.uk/ - Lead institution

Los Alamos National Laboratory (LANL), USA - http://www.lanl.gov/ - Partner institution

It is understood that Los Alamos National Laboratory, as a non-EU based institution, will not receive funding for this project. LANL is interested in participating in the proposed demonstrator project and engagement with EOSC despite this fact.

Team members - OU

Petr Knoth, Senior Research Fellow, KMi, The Open University, UK - Principle investigator for the

project

Lucas Anastasiou, Developer - GR7, KMi, The Open University, UK Giorgio Basile, Developer - GR6, KMi, The Open University, UK Team members - LANL

Martin Klein, Scientist, Los Alamos National Laboratory, USA, Principle Investigator at LANL Herbert Van de Sompel, Scientist - Team Leader, Los Alamos National Laboratory, USA

Additionally, we discussed our proposal with the following two existing EOSC demonstrators and agreed how we will engage with them during the course of the project (details discussed in the relavnt sections):

TEXTCROWD: Franco Niccolucci, University of Florence

High Energy Physics: Jamie Shiers, CERN

The team proposing this work consists of professionals from the Open University responsible for the national aggregation service CORE and scientists from Los Alamos National Laboratory who are editors of the ResourceSync specification and participated also in the creation of OAI-PMH protocol. It therefore has the required expertise, collaboration network and the necessary hardware resources available to make this

demonstrator a success.

Brief Summary of the Science Demonstrator

A single scientific repository is, if considered by itself, of limited value. Real benefits come from the ability to exchange information effectively and in an interoperable way, enabling the development of a wide range of global cross-repository services. However, exchanging metadata and content across scientific repositories in the EU (but also outside of the EU) is mostly based on a 15-year-old technology, symbolized by the OAI-PMH protocol. This protocol a) is unsuitable when there is a need to exchange large quantities of metadata, b) suffers from inconsistent implementations across providers and c) was only designed for metadata transfer, omitting the much needed support for content exchange.

We propose to pilot a demonstrator service for fast and highly scalable exchange of data across

repositories storing research datasets, manuscripts and scientific software. The data exchange in the demonstrator will be based on the ResourceSync protocol, that was designed and the first scalable implementation of which has been developed and deployed by the project team. This work will enable more efficient and effective information exchange between EOSC data providers and services, which is an essential step towards the realisation of the EOSC vision.

The demonstrator will provide argument for modernising existing legacy communication mechanisms routinely used by thousands of research repositories. The code developed for the demonstrator will lower the barrier to adoption of ResourceSync across data providers irrespective of scientific discipline. To provide evidence for this argument we will:

Quantitatively evaluate ResourceSync against OAI-PMH on the use case of aggregating millions of

resources from scientific repositories to the CORE aggregator and enabling others to keep in sync with relevant parts of this dataset.

Engage with the existing EOSC demonstrator “TEXTCROWD” to investigate and architect how this

technology can be applied to overcome the challenge of effectively sharing data from Social Sciences and Humanities repositories to make more resources available for TEXTCROWD.

Engage with the existing EOSC demonstrator “High Energy Physics” to investigate how this

technology can be used to aid the discoverability of data from CERN’s Large Hadron Collider.

Description of the Scientific Demonstrator

The demonstrator for fast and highly scalable exchange of metadata and content across repositories will be based on the ResourceSync protocol (http://www.openarchives.org/rs/toc). The first implementation of ResourceSync that scales to millions of items has been developed and deployed by the CORE team at the Open University, UK (OU) (slides: https://www.slideshare.net/petrknoth/seamless-access-to-the-worlds-open-access-research-

papers-via-resourcesync) in collaboration with Los Alamos National Laboratory, USA (LANL) and the Data Archiving and Networked Services, NL (DANS).

The demonstrator will showcase how scholarly communication resources, i.e. research datasets, manuscripts and scientific software, can be effectively, regularly and reliably exchanged across systems using the ResourceSync protocol. It will apply ResourceSync on real-world use cases with millions of

resources and a representative set of repository and service platforms reflecting the diversity of the EOSC ecosystem. The data synchronization will be shown:

a) across a cross-disciplinary network of repositories and

b) between repositories and global value-added services, such as those used in research evaluation, aggregation and workflow execution.

We will use the demonstrator to evaluate the efficacy of this solution, benchmarking it against the current state-of-the-art (mainly OAI-PMH) in terms of:

- speed (time)

- complexity (steps required to complete)

- reliability (recall)

- freshness (e.g. average time gap between syncs)

The empirical evaluation will be carried out along a set of dimensions, including:

- metadata only vs. metadata and content exchange

- batch vs. incremental exchange

- sequential vs. paralelised implementation of ResourceSync

The diversity of platforms considered will include repository platforms, such as EPrints, DSpace, Fedora or CKAN, and services, such as GitHub, CrossRef, Dryad, OpenAIRE, CORE or PMC. Additionally, by working with two existing EOSC demonstrators, we will investigate and describe a solution for application of this technology in the domains of High Energy Physics and Humanities and Social Sciences.

Science Area

1.2 Computer and information science, its results will be demonstrated on use cases in 1.3 Physical sciences, 5. Social sciences and 6. Humanities and can be extended to all areas of science.

Science Demonstrator title

Mining a large image repository to extract new biological knowledge about human gene function.

Principal Investigator and team members

Prof Jason Swedlow, University of Dundee, Euro-BioImaging

Dr Alvis Brazma, EMBL-EBI, Euro-BioImaging

Dr Jan Ellenberg, EMBL, Euro-BioImaging

Dr Jean-Karim Hériché, EMBL

Mr Balaji Ramalingam, University of Dundee, Open Microscopy Environment

Mr Josh Moore, University of Dundee, Open Microscopy Environment

Dr Simon Li, University of Dundee, Open Microscopy Environment

Brief Summary of the Science Demonstrator

Image-based genome-scale RNAi and small molecule inhibitor screens generate a wealth of image data that remains unexploited after their original publications. Re-analysing available image data from published RNAi screens would provide new biological knowledge about cellular functions of human genes. In addition, integrative analysis of multiple image data sets would take advantage of complementary information provided by the use of different assays and reporters in different screens. However, the major obstacles to the combined analysis of multiple screen image data sets are the complexity and the size of the data, which preclude easy network transfer and require a high-performance computing infrastructure for processing in a reasonable amount of time.

This project will leverage an existing strategic collaboration between the Euro-BioImaging- and BBSRC-funded Image Data Resource (IDR; http://idr.openmicroscopy.org) and the EMBL-EBI Embassy Cloud. IDR holds >40 systematic imaging datasets, >1 Mio experiments, and imaging data related to >19,600 human gene and >31,000 drugs or small molecule inhibitors. In this pilot project, we will establish the resources required to perform comprehensive machine learning analyses on these datasets, with the ultimate goal of identifying functional connections between genes and/or small molecules that target them based on image-based phenotypes. The pilot will test the validity of this approach and also demonstrate how a large cloud-based collection of published datasets can be re-used for novel discovery. Besides generating testable hypotheses about cellular functions, this study will also produce re-usable infrastructure and analyses for generating value from published image data.

Description of the Scientific Demonstrator

The proposed pilot is based on the Image Data Resource (IDR; http://idr.openmicroscopy.org), an online database of image datasets, with annotations and access for browsing, search and re-analysis. IDR holds 41.5 TB of image data in 35.7 Mio image planes and ~1.03 Mio individual experiments, and includes all associated experimental (e.g., genes, RNAi, chemistry), analytic, and functional annotations. >90% of the functional annotations have links to defined, published controlled vocabularies and ontologies. IDR datasets sample a variety of biomedically-relevant biological processes like cell shape, division and migration, at scales ranging from nanometer-resolution localisation of proteins in cells to millimetre-scale structure of tissues, from studies in S. pombe, S. cerevisiae, A. thaliana, D. melanogaster, M. musculus and H. sapiens.

The pilot project will analyse publicly available image data from genome-scale RNAi screens and small molecule screens to gain insights into cellular functions of human genes and try to identify inhibitors of these functions. The work is organised around the following points:

1- A feature vector will be computed for each image of every data set. This step is already under way

in the production IDR system in the Embassy Cloud.

2- Two strategies to mine the integrated data for new gene functions will be developed:

- A semi-supervised approach in which the data will be queried to find genes involved in specific cellular functions. This will again make use of the Embassy Cloud to parallelize queries to be able to explore a wide range of functions.

- An unsupervised approach based on tensor factorizations to define groups of genes and molecules

involved in the same biological process.

3- We will assess the results using independent data covering previous knowledge such as annotations with Gene Ontology or Cellular Microscopy Phenotype Ontology terms, both of which are already supported in the IDR.

By using image feature extraction followed by application of machine learning methods, this project is representative of many microscopy data analysis pipelines. As such, it will demonstrate how image data can be accessed and reused via the cloud after publication. In addition, the infrastructure, methods and data produced can be scaled and reused.

Science Area

1.6 Biological sciences

Science Demonstrator title

VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics

Principal Investigator and team members

Principal Investigator

Alessandro Costa (INAF - Catania) is a senior INAF Technologist and Researcher. He has got his Master's degree in Telecommunications Engineering in 2000 and he works at INAF since 2001. He works as computer scientist in Scientific Visualization, Authentication & Authorization Infrastructures (AAI), High Performance Computing and Virtual Observatory.

He has been taking part to the CTA Project and in particular to the Data Management activity (2013-present).

The list of most recent European Projects (last five years) where Alessandro Costa took part is: SCI- BUS, ER-flow, VIALACTEA, INDIGO DataCloud, ASTERICS.

He is the leader of the visualization task in the H2020-funded AENEAS project.

He is the INAF coordinator for the H2020-funded AARC2 project.

Team Members

Ugo Becciani (INAF - Catania), holds a permanent position as Astronomer Researcher. He has been PI and CoPI of several research projects on different astrophysical topics, data exploration, visual analytic, 3D visualization, grid computing, virtual observatory and advanced studies in supercomputing and parallel computing.

Eva Sciacca (INAF - Catania), researcher focuses on activities related to: scientific visualization, data analysis, science gateways, workflow management systems and data systems exploiting distributed

infrastructures.

Fabio Vitello (INAF - Catania), post-doctoral researcher in the field of HPC, authentication &

authorization and scientific visualization.

Sergio Molinari (IAPS/INAF) leads the Star Formation Group at IAPS. He has almost 20 years experience in Galactic star formation, with more than 170 refereed publications with H-index 53, and more than 25 invited talks at international conferences. He is PI of the Herschel/Hi-GAL key- project and of the FP7-funded VIALACTEA project.

Brief Summary of the Science Demonstrator

The Astrophysical community has currently set up a new suite of cutting-edge Milky Way surveys that provide a homogenous coverage of the entire Galactic Plane and that have already started to transform the view of our Galaxy as a global star formation engine. New instruments have delivered information of unprecedented depth and spatial detail spanning the electromagnetic spectrum. While a huge progress has been made in the last two decades in understanding the evolution of isolated dense molecular cores toward the onset of gravitational collapse and the formation of single stars and protoplanetary systems, a lot remains still hidden such as the relative importance of

gravity, turbulence, the perturbation from spiral arms and the triggering from explosive events, in shaping the diffuse Interstellar Medium into molecular clouds and driving their fragmentation and evolution into dense filamentary structures and dense clumps that will originate the young clusters that are the new stellar nurseries of the Milky Way Galaxy.

The volume and complexity of current new survey datasets and scientific challenges calls for a

radical re-evaluation of the current science and data analysis techniques.

The proposed approach is the integration in the EOSCpilot e-infrastructure of a visual analytics environment based on VisIVO (Visualization Interface for the Virtual Observatory) (Becciani, 2010; Costa, 2011; Sciacca, 2015), developed by INAF Catania. In the ViaLactea project (FP7) a scientific use case was implemented as a VisIVO module to study the star formation regions on our galaxy (Molinari, 2016). The tool, named ViaLactea Visual Analytics (VLVA), is being continuously used by

the astrophysical community: https://github.com/inaf-oact-VisIVO. VLVA is integrated with INAF IA2

service (Molinaro, 2016) that exposes a wide variety of surveys and catalogues of the Milky Way.

Description of the Scientific Demonstrator

VisIVO framework and the VLVA tool have being used by the astrophysics community to study the star forming regions accessing data available by INAF IA2 service.

VisIVO is an application specifically designed to deal with multidimensional data, it leverages the

Visualization ToolKit exploiting its processing and visualization functionalities including the most advanced visualization algorithms and techniques including vector, tensor, scalar, texture, and advanced volumetric methods. VisIVO has been used to better understand the multiple physics principles behind star formation regions. VisIVO contains many generic visual analysis instruments that make it reusable for complex multivariate tabular data in different fields. As an example, a project recently approved by the Italian Ministry of Research is aimed to customize VisIVO features for geological studies of volcanic fractures and

for environmental geohazards analysis. The framework implements Virtual Observatory (VO) protocols, such as the Table Access Protocol to access VO data and Simple Application Messaging Protocol to interoperate with other VO-compliant tools.

The sustainability of the proposed science demonstrator is guaranteed thanks to the support of INAF Catania Observatory and Institute for Space Astrophysics and Planetology in Rome being the data analysis and the scientific visualization one of the main mission for INAF.

The usage of innovative infrastructures, thanks to the deployment of visualization services in EOSC, will allow an integrated approach combining visualization and data analysis.

The porting activity on EOSC ecosystem will be carried out in 12 months considering that the tool is already deployed at INAF HTC/HPC infrastructures.

Science Area

1.3 Physical sciences: Astronomy (including astrophysics, space science

Science Demonstrator title

Switching on the EOSC for Reproducible Computational Hydrology by FAIR-ifying eWaterCycle and SWITCH-ON.

Principal Investigator and team members

Dr. ir. Rolf Hut - Delft University of Technology Dr Niels Drost - Netherlands eScience Center Dr Berit Arheimer - SMHI

Michael R. Crusoe - Common Workflow Language

Dr Axel Berg - SURFsara

Dr Sergio Andreozzi - EGI Foundation

Dr Lukasz Dutka - CYFRONET

Dr Anton Frank - Leibniz Supercomputing Centre (LRZ), Bavarian Academy of Sciences and

Humanities

Prof dr. ir. Nick van de Giesen - Delft University of Technology

To best make use of the time (12PM) available for this demonstrator, the core development team of our demonstrator will be formed by a small team R.Hut, (TU Delft, 4PM) N.Drost (Netherlands eScience Center, 4PM), with extensive knowledge of both the scientific use cases and technical solutions proposed. The remaining 4PM funding will be divided between the remaining partners, depending on challenges faced, distributed through TU Delft, thus minimizing fragmentation and making sure that the EOSCpilot

consortium does not need to deal with too many partners.

Brief Summary of the Science Demonstrator

Central to the science of hydrology is the localized nature of the medium through which water flows. No two stretches of soil and vegetation are the same, making empirical parameterization a

necessity. This has led to a large number of local hydrological models that work well for given watersheds. Recently, a new challenge has been accepted by the hydrological community to build global models. The eWaterCycle project produced a fully open source operational hydrological model, producing worldwide streamflow and soil moisture forecasts (http://forecast.ewatercycle.org). In the meantime, the European FP7 project SWITCH-ON (http://www.water-switch-on.eu/), has built an ICT environment that allows for collaborative experiments and sharing of protocols, data and research results. This science demonstrator proposal seeks to integrate the top-down approach of eWaterCycle with the bottom-up approach of SWITCH- ON.

The vision is a computational environment that allows for easy integration of local models in a global model. This environment will allow the hydrological community to include local knowledge,

including models and data, in large scale models without extensive knowledge of HPC.

Both models and data will adhere to the Findable, Accessible, Interoperable and Reusable (FAIR) principles. For the data part, we will ensure FAIRness of data by storing all data and metadata using well defined standards (e.g., NetCDF), and using the Onedata (https://onedata.org) platform for storing, accessing and publishing data.

FAIRness of models and other software is often overlooked (Hutton et. al. 2016, Hut et. al. 2017). We will make this a spearpoint of our demonstrator; by using only open source software, coupled with standards such as the Common Workflow Language (http://commonwl.org) for formalizing scientific workflows, and container platforms such as Docker and Singularity, we will ensure results are truly reproducible and reusable.

Description of the Scientific Demonstrator

In this EOSC science demonstrator we will combine lessons learned from SWITCH-ON and eWaterCycle to make the next step towards reproducible, reusable, open, data driven science in Hydrology, by showcasing a modern approach to large scale simulation-based science. We will create FAIR versions of data required for running our hydrological models. We will make these available along with the results of our models for further usage by researchers in both hydrology and related fields. All data will be stored using the EGI DataHub system based on the open source Onedata technology, allowing seamless access to data from different compute infrastructures

We will also FAIRify the software; the two hydrological models themselves, and any software used for preprocessing and analysis of the results. To this end we will make all software available, both in source code form, and as ready-to-run software containers based on Docker and Singularity. As it is vital to have a complete and accurate description of all the steps and data required to calculate the output and analysis of these models, we will make use of the Common Workflow Language standard to formalize the process.

The resulting system will at least contain all data and software needed to run the eWaterCycle global hydrological forecast, coupled with the HYPE local model from the SWITCH-ON project, in this case for Sweden. This will then serve both as a demonstrator, and as a basis for scientists to extend with their data and models.

We will show how our software can function on any available suitable resource, by using both different Supercomputers (the Dutch National Supercomputer Cartesius and SuperMUC at the Leibniz-Rechenzentrum) and Cloud environments (the EGI federated cloud and the Dutch HPC Cloud).

Science Area

1.5 Earth and related Environmental sciences: Hydrology

1.5 Earth and related Environmental sciences: Water Resources

Science Demonstrator title

VisualMedia: a service for sharing and visualizing visual media files on the web

Principal Investigator and team members

Partners participating to the development activities (using funds):

Roberto Scopigno, ISTI-CNR, Visual Computing Lab, Pisa, Italy (principal investigator) Carlo Meghini, ISTI-CNR, Nemis Lab, Pisa, Italy

Franco Niccolucci, PIN, Prato, Italy

Achille Felicetti, PIN, Prato, Italy

Sara di Giorgio, MIBACT – ICCU, Roma, Italy

Partners participating to the use cases and assessment activities:

Christodoulos Chamzas & Konstantinos Stavroglou, Athena Research Center, ILSP-Xanthi Branch, Greece

Ute Dercks, Photothek des Kunsthistorischen Instituts in Florenz - Max-Planck-Institut, Germany/Italy

Adeline Joffre, TGIR Huma-Num, CNRS, France

Brief Summary of the Science Demonstrator

Visual aspects are of paramount importance for Cultural Heritage (CH) research. Images of artifacts, monuments and sites are the most widely used support to analysis and interpretation; 3D models are also playing an important role in CH research and management. Although visual technologies have achieved a high level of sophistication, their use is still fragmented among the many individual

researchers, small research teams and different institutions that separately manage their own digital image library. Often there is also a limit in the ability to choose, install and use the most advanced visual tools even when they are available as Open Source. The availability of a generic public web service demonstrated

with the ARIADNE Visual Service (http://visual.ariadne-infrastructure.eu/) has been welcomed by users, who have however pointed out possible improvements of the current service version. The SD will provide researchers with a system to publish on the web, visualize and analyze images and 3D models in a common workspace, enabling sharing, interoperability and re- use. The system will also improve findability, supported by the metadata enrichment component (to be added in this SD): at present, many heritage images have poor metadata that make them unsuitable for finding and accessing.

The SD will consist of the following modules:

• Web publication and Visualization of high-resolution images, RTI (i.e. dynamically re-lightable images), and 3D models, all encoded in a multi-resolution format

• Thumbnail-based browsing interface

• Customization of presentation through the selection of different browsing systems for 3D models

• Metadata enrichment based on automatic processing of related documentation (reports, captions,

etc) and/or manual input/revision.

It is anticipated that the images uploaded to the system may eventually form a public distributed digital library, accessible through the cloud using the Visual Tools described here. However, this aspect will be analysed but not implemented by the SD.

Description of the Scientific Demonstrator

The SD provides easy visualization and presentation of complex visual media assets. It is an automatic service that allows to upload visual media files in the cloud (by means of a web interface) and to transform them into an efficient web format, making them ready for web-based visualization. All processing is done in an automatic manner. Images may be grouped in collections, uploaded in a single action and presented in a coordinated manner.

Therefore, the service will be instrumental to a wider dissemination and sharing of data, in a context

(CH or Digital Humanities) still characterized by scarce sharing or reuse of visual data resources. The service supports publication on the web and browsing of three types of visual media:

• High-resolution 2D images (input converted in a multi-resolution format and browsed in real time, zooming in and out);

• Reflection Transformation Images (RTI), also known as Polynomial Texture Maps (PTM) images, i.e.

dynamically re-lightable images;

• 3D models (triangulated meshes, point clouds and textured models).

For each media type, the service supports automatic conversion to an efficient multi-resolution representation, offering data compression, progressive transmission and view-dependent rendering. At visualization time, the image browser is based on an interface for easily browsing the list of

images (represented by thumbnails) and selecting the one to be rendered in full resolution. Each

media is presented visually with a standard browser (one for images, another for RTI images and a third one for 3D models). In the case of 3D models users may customize the behaviour and graphical interface of the 3D browser, since alternative templates for 3D content are provided, to address specific sub-types of 3D scenes and to extend the flexibility of the overall system.

As regards the metadata, the service allows a manual input using a simple form and/or the automatic creation based on terms list or automatic processing of related text documents. Integration with TEXTCROWD is planned for this purpose.

Science Area

Cultural Heritage (CH), i.e. museums and artworks, CH restoration, Archaeology:

- 6.1 History and archaeology

- 6.4 Art